Functional Compensation of Mouse Duplicates by their Paralogs Expressed in the Same Tissues

Agusto Luzuriaga-Neira, Krishnamurthy Subramanian[1], and David Alvarez-Ponce *

Biology Department, University of Nevada, Reno, Reno Nevada 89557

[1]Present address: Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854.

*Corresponding author: E-mail: dap@unr.edu.

Accepted: 30 July 2022

Abstract

Analyses in a number of organisms have shown that duplicated genes are less likely to be essential than singletons. This implies that genes can often compensate for the loss of their paralogs. However, it is unclear why the loss of some duplicates can be compensated by their paralogs, whereas the loss of other duplicates cannot. Surprisingly, initial analyses in mice did not detect differences in the essentiality of duplicates and singletons. Only subsequent analyses, using larger gene knockout data sets and controlling for a number of confounding factors, did detect significant differences. Previous studies have not taken into account the tissues in which duplicates are expressed. We hypothesized that in complex organisms, in order for a gene's loss to be compensated by one or more of its paralogs, such paralogs need to be expressed in at least the same set of tissues as the lost gene. To test our hypothesis, we classified mouse duplicates into two categories based on the expression patterns of their paralogs: "compensable duplicates" (those with paralogs expressed in all the tissues in which the gene is expressed) and "noncompensable duplicates" (those whose paralogs are not expressed in all the tissues where the gene is expressed). In agreement with our hypothesis, the essentiality of noncompensable duplicates is similar to that of singletons, whereas compensable duplicates exhibit a substantially lower essentiality. Our results imply that duplicates can often compensate for the loss of their paralogs, but only if they are expressed in the same tissues. Indeed, the compensation ability is more dependent on expression patterns than on protein sequence similarity. The existence of these two kinds of duplicates with different essentialities, which has been overlooked by prior studies, may have hindered the detection of differences between singletons and duplicates.

Key words: duplicates, singletons, essentiality, gene expression.

Significance

A gene's loss can, under certain circumstances, be buffered by its paralogs, which results in duplicates exhibiting a lower essentiality than singletons. However, the circumstances under which this buffering can occur are poorly understood. We hypothesized that the loss of a gene can only be compensated by its paralogs if such paralogs are expressed in the same tissues as the lost gene. In agreement with our hypothesis, we found that mouse "compensable duplicates" (those with paralogs expressed in the same tissues) are less likely to be essential than "noncompensable duplicates" (those without such paralogs), whose essentiality is equivalent to that of singletons.

Introduction

Gene duplication is a crucial evolutionary mechanism, because it is the primary source of new genes (Ohno 1970; Zhang 2003). As duplicated genes tend to perform similar functions, gene duplication is a source of functional redundancy (Gu et al. 2003; Conant and Wagner 2004; Liang and Li 2009). As a result, under certain circumstances, duplicated genes can have a backup role, meaning that they

can compensate for the loss of their paralogs (Gu et al. 2003). However, it is unclear why certain genes can compensate for the loss of their duplicates whereas others cannot.

Functional compensation by duplicates has been inferred in several model organisms by comparing the proportion of essential genes (P_E) between singletons and duplicates. If the functional loss of duplicated genes can often be compensated by their paralogs, the P_E of genes with no duplicates (singletons) should be higher than the P_E of duplicated genes. A higher P_E in singletons has been observed in several organisms, including the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the plant *Arabidopsis thaliana* (Gu et al. 2003; Conant and Wagner 2004; Hannay et al. 2008; Hanada et al. 2009; Wang, Birsoy, et al. 2015). In contrast, early analyses in mouse did not detect significant differences in the P_E of singletons and duplicates (Liang and Li 2007; Liao and Zhang 2007). Only more recent analyses, using larger data sets and correcting for a number of confounding factors, have found differences between mouse singletons and duplicates (Liang and Li 2009; Makino et al. 2009; Chen et al. 2012; Su et al. 2014; Acharya et al. 2015; Kabir et al. 2019).

Several confounding factors have been suggested as potential sources of bias in the estimation of P_E in mouse singletons and duplicates. Makino et al. (2009) found that the mouse knockout data set is enriched in duplicated developmental genes and in ohnologs (duplicates resulting from the two rounds of whole-genome duplication (WGD) though to have affected a common ancestor of vertebrates; Ohno 1970), which results in inflated P_E estimates. However, Liang and Li (2009) showed that the enrichment in developmental genes does not cause a significant bias in the P_E estimates of duplicates because the enrichment equally affects the entire mouse knockout data set. The degree of divergence between duplicates is considered to be another confounding factor. Studies in worms, yeast, and plants have shown a negative correlation between sequence similarity with the closest paralog and essentiality (Gu et al. 2003; Conant and Wagner 2004; Hanada et al. 2009). In contrast, in mouse the correlation is positive or not significant (Liao and Zhang 2007; Su and Gu 2008). Based on this idea, Liao and Zhang (2007) suggested that if researchers tend to avoid targeting genes with close paralogs (because such genes are unlikely to produce detectable phenotypes), the mouse knockout data set might be biased toward duplicates with higher P_E. There is a positive correlation between essentiality and the number of protein–protein interactions (Jeong et al. 2001; Batada et al. 2006; Reguly et al. 2006; Liang and Li 2007). Thus, centrality can also bias P_E estimates because, unlike in *Escherichia coli*, yeast, worm, and fly (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009), mammalian duplicates display more centrality than singletons (Liang and Li 2007; D'Antonio and Ciccarelli 2011; Doherty et al. 2012), which could explain the similarity between the P_E estimates of singletons and duplicates found in mouse analyses. Chen et al. (2012) suggested the age of the gene family (phyletic age) as another factor causing bias in the P_E estimation. They stated that old mouse singletons and duplicated genes are the preferred targets for knockout experiments because they present a higher rate of detectable phenotypes. Kabir et al. (2019) identified the developmental stage of expression as another confounding factor, finding that pairs of paralogs that are both essential are more likely to be at different stages of development than pairs of paralogs that include one or two nonessential genes.

In complex organisms, each gene is expressed in a set of tissues. We hypothesized that the loss of a gene can only be compensated by its paralogs if these paralogs are expressed (at least) in the same tissues. For instance, the loss of a gene that is expressed in the brain cannot be compensated by a paralog that is exclusively expressed in the liver. However, previous analyses have largely ignored this requirement for gene buffering (expression in the same tissues), which may have hindered the detection of differences in the P_E of singletons and duplicates in mouse.

Here, we distinguish between two types of duplicates: those with paralogs that are expressed in the same tissues (which we term "compensable duplicates") and those without such kinds of paralogs (termed "noncompensable duplicates"). We classified a gene in the former category if, for each tissue in which it is expressed, it has at least one paralog that is expressed in that tissue. We found that both kinds of duplicates behave differently: noncompensable duplicates have a P_E that is similar to that of singletons, whereas compensable duplicates exhibit a significantly lower P_E. We show that the differences in P_E between compensable and noncompensable duplicates are independent of a number of potentially confounding factors. These results indicate that functional compensation of a gene's loss by its paralogs is generally only possible when such paralogs are expressed in the same set of tissues as the lost gene. This may explain why previous analyses have detected no differences or only moderate differences between singletons and duplicates analyzed as a whole.

Results

Compensable Duplicates are Less Essential than Noncompensable Duplicates

Earlier analyses have shown both nonsignificant and significant differences in the proportion of essential genes (P_E) between singleton and duplicated mouse genes (Liang and Li 2007; Liao and Zhang 2007; Su and Gu 2008; Makino et al. 2009; Su et al. 2014; Acharya et al. 2015; Kabir et al. 2019).

We re-evaluated this difference using a more recent version of the mouse knockout data set. Of the 22,175 mouse protein-coding genes, phenotype data were available for 8,730 genes (of which 4,215 are essential and 4,515 are nonessential). Of those genes, 1,964 are singletons and 6,766 are duplicates. Singletons exhibit a P_E of 57.08%, whereas duplicates exhibit a P_E of 45.73% (fig. 1 and table 1), and a Fisher's exact test (FET) showed significant differences in the P_E of both groups ($P < 2.2 \times 10^{-16}$). These results agree with more recent studies that found P_E to be lower in duplicates than in singletons in mouse (Su and Gu 2008; Liang and Li 2009; Makino et al. 2009; Kabir et al. 2019) and indicate that duplicates can, under certain circumstances, compensate for the loss of their paralogs.

We hypothesized that the loss of any given duplicate could only be compensated by its paralogs if these paralogs are expressed in the same tissues as the lost gene. This would mean that only a fraction of duplicates would be "compensable" and thus expected to exhibit reduced essentiality. To test this hypothesis, we divided duplicated genes into two categories based on the patterns of expression of the genes and their paralogs: compensable duplicates (those with paralogs expressed in the same tissues) and noncompensable duplicates (those without paralogs expressed in the same tissues). Out of the 6,766 duplicates, we found 5,562 to be compensable and 1,204 to be noncompensable. Compensable duplicates exhibited a P_E of 43.82%, whereas noncompensable duplicates displayed a P_E of 54.57% (fig. 1, table 2), and the FET detected significant differences between both classes of duplicates ($P = 1.24 \times 10^{-11}$). In addition, we detected no differences between singletons and noncompensable duplicates (FET, $P = 0.1740$). Combined, these results support our hypothesis that duplicates can only be compensated by paralogs that are expressed in the same tissues.

Singletons, Compensable Duplicates, and Noncompensable Duplicates are Different in Terms of Age, Gene Function, Type of Duplication Event, and Similarity with the Closest Paralog

Previous studies showed that a number of factors, including gene age, the number of protein–protein interactions, gene function, type of duplication event, and degree of similarity with the closest paralog, influence essentiality (Jeong et al. 2001; Gu et al. 2003; Conant and Wagner 2004; Batada et al. 2006; Prachumwat and Li 2006; Reguly et al. 2006; Liang and Li 2007; Su and Gu 2008; Makino et al. 2009; D'Antonio and Ciccarelli 2011; Chen et al. 2012; Doherty et al. 2012; Zhu et al. 2012; Su et al. 2014; Kabir et al. 2019). Therefore, these are potentially confounding factors: it is possible that these factors differ between singletons and duplicates, and/or between compensable and noncompensable duplicates, and that these differences account for their differences in P_E. To account for this possible effect, we first evaluated whether these factors differ

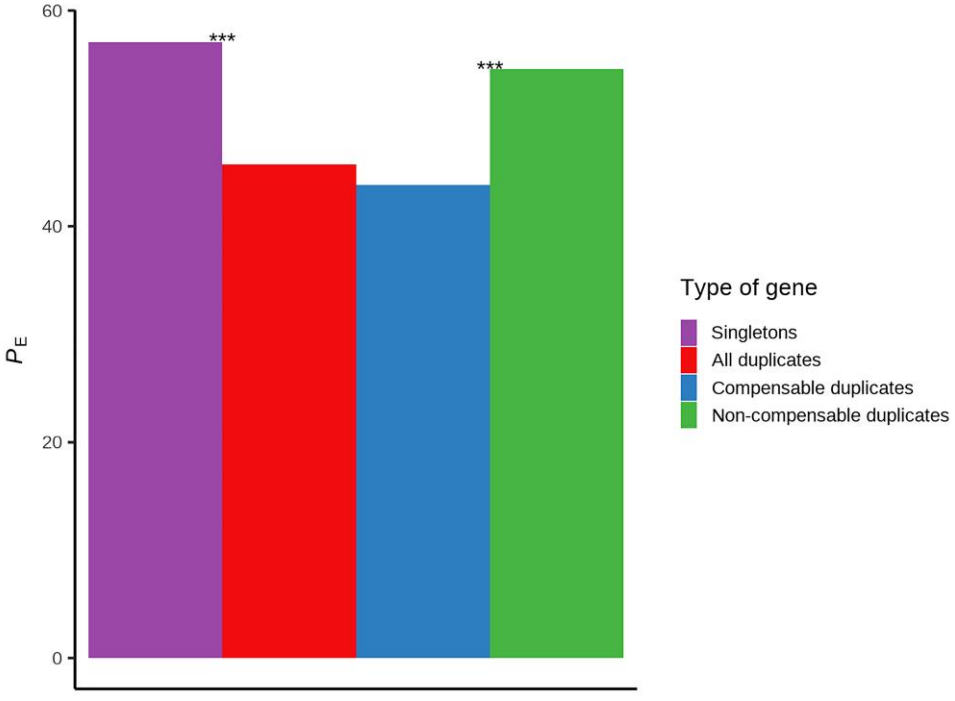

Fig. 1.—Differences in the percentage of essential genes (P_E) between singletons and duplicates, and between compensable and noncompensable duplicates in the mouse genome. Fisher's exact test significance levels: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

Table 1
Comparison Between Singleton and Duplicated Mouse Genes

	Singletons			Duplicates			
	N	***n***	**%**	***N***	***n***	**%**	**FET *P*-value**
Essential	1,964	1,121	57.08	6,766	3,094	45.73	$<2.2\times10^{-16}$***
Developmental genes	1,964	660	33.60	6,766	2,102	31.07	0.0988
	N	**Average**	**Median**	**SD**	***N***	**Average**	**Median**	**SD**	**MWU *P*-value**
Gene age	1,963	964.56	797	239.37	6,724	730.67	454.6	674.77	0.0121*
Interactions	1,383	14.40	6	28.15	5,054	16.99	7	53.16	0.0811

FET, Fisher's exact test; MWU, Mann–Whitney *U* test; *N*, number of genes with available data for each variable; *n*, number of genes in a particular category.
*$P<0.05$.
**$P<0.01$.
***$P<0.001$.

between singletons and duplicates (table 1), and/or between compensable and noncompensable duplicates (table 2). Second, we controlled for these factors, observing in all cases that the differences in P_E are not due to differences in these confounding factors (figs. 2–6 and supplementary table S1, Supplementary Material online).

We observed that, on average, singletons are older than duplicates (average ages: 964.65 and 730.67 Ma, respectively; Mann–Whitney *U* [MWU] test, $P=0.0121$). However, they do not significantly differ in their number of protein–protein interactions or in the percent of developmental genes (table 1). We also observed that compensable duplicates present a higher proportion of developmental and WGD genes (31.86% and 38.60%, respectively) than noncompensable duplicates (27.33% and 33.05%, respectively; FET, $P=0.0019$ and 0.0003, respectively). In addition, noncompensable duplicates are older on average than compensable duplicates (average ages: 768.10 and 722.53 Ma, respectively; MWU test, $P=0.0235$). Moreover, compensable duplicates present a higher average percentage of sequence similarity with their closest paralogs than noncompensable duplicates (47.42% and 42.15%, respectively; MWU test, $P=<2.2\times10^{-11}$). No significant differences were observed in terms of the number of protein–protein interactions (table 2). To test whether these differences account for the differences in P_E between compensable and noncompensable duplicates, we controlled for each of these potentially confounding factors (see the following sections).

Gene Age Does not Explain the Low Essentiality of Compensable Duplicates

Previous analyses have shown that essential genes tend to be older than nonessential ones, indicating that gene age has an impact on essentiality (Chen et al. 2012). This, combined with our observation that noncompensable duplicates are on average older than compensable duplicates (table 2), might potentially account for the high essentiality of noncompensable duplicates. To account for this possible effect, we divided genes into three age groups using age estimates from the ProteinHistorian database (Capra et al. 2012). Group 1 includes genes belonging to young gene families (originated 0–361.2 Ma; $n=2,752$), group 2 includes genes belonging to families of intermediate age (454.6–842 Ma, $n=3,271$), and group 3 includes genes belonging to old families (910–4,200 Ma; $n=2,141$). These thresholds were chosen so that the resulting age groups had a similar number of genes.

In line with previous observations (Chen et al. 2012), we observed that older genes exhibited a higher P_E than younger ones (fig. 2). In addition, within groups 2 and 3, singletons are more essential than duplicates (FET, $P=7.02\times10^{-7}$ for group 2 and $P<2.2\times10^{-6}$ for group 3), and noncompensable duplicates are more essential than compensable duplicates (FET, $P=2.46\times10^{-5}$ for group 2 and $P=3.46\times10^{-8}$ for group 3; fig. 2). No significant differences were detected within group 1 between singletons and duplicates (FET, $P=0.692$) or between compensable and noncompensable duplicates (FET, $P=0.2223$). These results indicate that differences in gene age do not account for the differences in P_E between singletons and duplicates or between compensable versus noncompensable duplicates (at least for genes belonging to old families and families of intermediate age).

Sequence Similarity with the Closest Paralog does Not Explain the Low Essentiality of Compensable Duplicates

Previous analyses showed that duplicates resulting from recent duplication events (and thus exhibiting a high degree of similarity with their closest paralogs) are underrepresented in the mouse knockout data set. A likely explanation is that such duplicates are less likely to produce phenotypes upon knockout, and are thus less likely to be targeted by scientists conducting knockout experiments. This effect has been proposed as a factor reducing the differences in the P_E estimates of singletons and duplicates (Su and Gu 2008).

Analyses in yeast, worms, and plants showed a positive correlation between sequence divergence from the closest

Table 2
Comparison Between Compensable and Noncompensable Duplicated Mouse Genes

	Compensable duplicates			Noncompensable duplicates			
	N	*n*	%	*N*	*n*	%	FET *P*-value
Essential	5,562	2,437	43.82	1,204	657	54.57	1.2×10^{-11}***
Developmental genes	5,562	1,772	31.93	1,204	329	27.33	0.0019*
WGD	5,562	2,141	38.49	1,204	398	33.05	0.0003***
SSD	5,562	3,421	61.51	1,204	806	66.95	0.0003***

	N	Average	Median	SD	*N*	Average	Median	SD	MWU *P*-value
Gene age	5,524	722.53	454.60	669.09	1,200	768.10	454.6	699.41	0.0235*
Interactions	4,072	17.07	7	53.16	982	16.69	6	59.87	0.1477
Sequence similarity to the closest paralog	5,429	48.92	47.42	19.22	1,170	42.15	40.61	17.13	$<2.2\times10^{-16}$***

FET, Fisher's exact test; MWU, Mann–Whitney *U* test; WGD, Whole-genome duplicates; SSD, Small-scale duplicates; *N*, number of genes with available data for each variable; *n*, number of genes in a particular category.
*$P<0.05$.
**$P<0.01$.
***$P<0.001$.

paralog and the essentiality of duplicated genes (Gu et al. 2003; Conant and Wagner 2004; Hanada et al. 2009), indicating that genes are more likely to compensate for the loss of their paralogs if they encode proteins that are highly similar at the sequence level. In contrast, surprisingly, studies in mice found a negative or no correlation between sequence divergence from the closest paralog and essentiality (Liao and Zhang 2007; Su and Gu 2008; Su et al. 2014).

Our analyses show that compensable duplicates exhibit, on average, a high degree of sequence similarity with their closest paralogs compared with noncompensable duplicates (table 2). Assuming that low divergence between paralogs enables compensation, this observation might potentially explain the low P_E of compensable duplicates. To account for this possible effect, we divided duplicates into four groups (each with 1,649 genes) according to their level of amino acid similarity to their closest paralog, and compared the P_E of compensable and noncompensable duplicates within each group. Within each of the four groups, compensable duplicates exhibited a significantly lower P_E than noncompensable duplicates (fig. 3). Within group 1 (>60.20% sequence similarity), P_E is 45.63% for compensable duplicates and 63.39% for noncompensable duplicates (FET, $P=9.655\times10^{-6}$). Within group 2 (similarity ranging between 60.19% and 46.28%), P_E is 43.02% and 56.13% for compensable and noncompensable duplicates, respectively (FET, $P=0.0001$). Within group 3 (similarity ranging between 46.27% and 35.25%), P_E was 42.60% for compensable duplicates and 44.48% for noncompensable duplicates (FET, $P=0.0035$). Finally, within group 4 (<32.24% sequence similarity), P_E was 42.97% for compensable duplicates and 45.40% for noncompensable duplicates (FET, $P=0.0003$). These results indicate that the similarity with the closest paralog is not the cause of the differences in P_E between compensable and noncompensable duplicates.

The Number of Protein–Protein Interactions does Not Explain the Low Essentiality of Compensable Duplicates

A number of studies in worms, yeast, and mouse indicate that genes involved in a high number of protein–protein interactions tend to be more essential than those with fewer interactions (Jeong et al. 2001; Batada et al. 2006; Reguly et al. 2006; Liang and Li 2007; Su et al. 2014; Kabir et al. 2017). Duplicates tend to be lowly connected in yeast and worm (Hahn and Kern 2005; Prachumwat and Li 2006), whereas the opposite trend has been observed in mammals and plants (Liang and Li 2007; Alvarez-Ponce and Fares 2012; Doherty et al. 2012). In line with previous results in mammals (Liang and Li 2007; Zhu et al. 2012; Su et al. 2014), our analyses suggest a higher number of protein–protein interactions in mouse duplicates compared with singletons (table 1). However, the differences are not statistically significant (MWU test, $P=0.0811$), possibly due to the scarcity of mouse interactome data compared with other organisms.

Even though we did not detect significant differences in the number of protein–protein interactions of compensable and noncompensable duplicates (table 2), we wanted to test the possibility that the number of interactions may account for the differences in the P_E of both kinds of duplicates. To that end, we divided duplicates into four categories according to their number of interactions, and compared the P_E of compensable and noncompensable duplicates within each category. Group 1 contained genes with more than 15 interactions ($n=1,309$), group 2 included genes with 7–15 interactions ($n=1,284$), group 3 contained genes with 3–6 interactions ($n=1,318$), and group 4 included genes with 1–2 interactions ($n=1,143$). These thresholds were chosen so that the resulting connectivity groups had a similar number of genes. In line with previous results (Jeong et al. 2001; Batada et al. 2006; Reguly et al. 2006; Liang and Li 2007), the P_E of

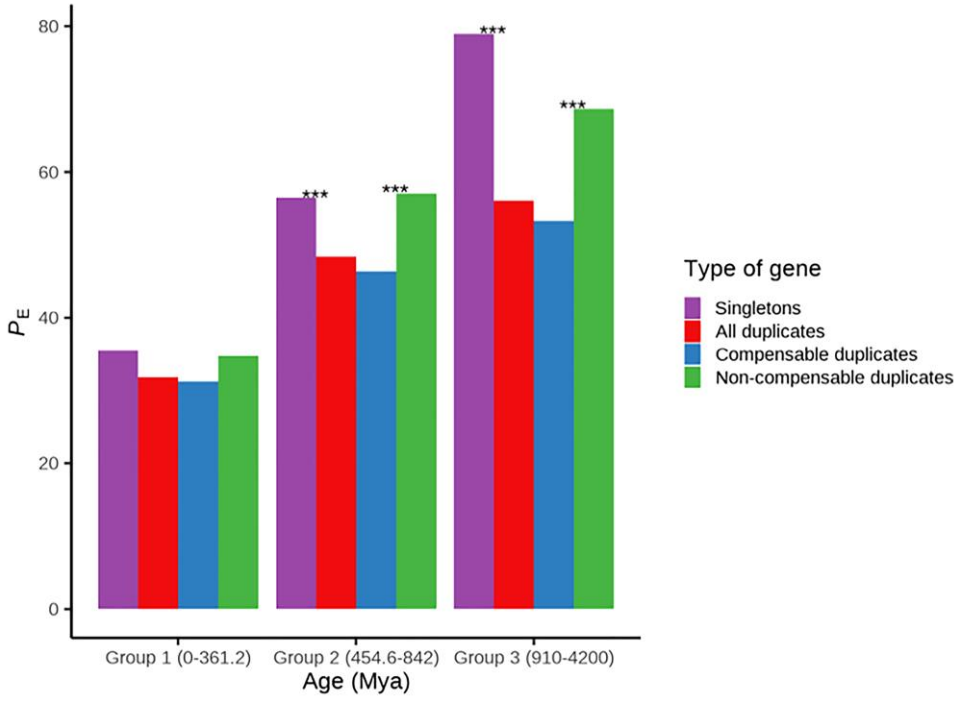

Fig. 2.—Differences in the percentage of essential genes (P_E) between singletons and duplicates, and between compensable and noncompensable duplicates, controlling for gene age. Fisher's exact test significance levels: $*P<0.05$; $**P<0.01$; $***P<0.001$.

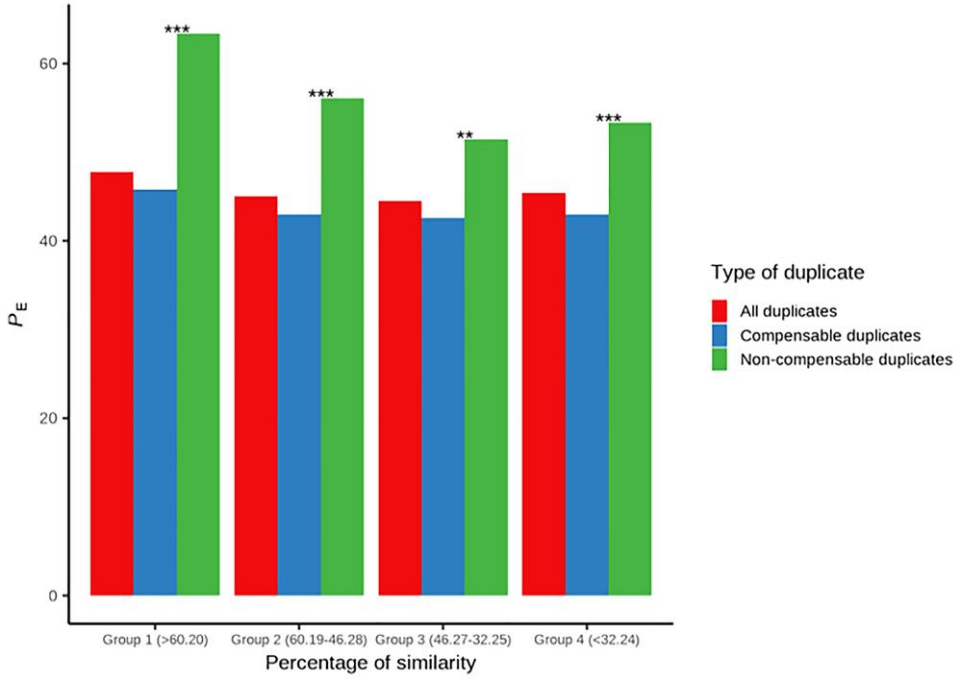

Fig. 3.—Differences in the percentage of essential genes (P_E) between compensable and noncompensable duplicates, controlling for the percent of similarity to the closest paralog. Fisher's exact test significance levels: $*P<0.05$; $**P<0.01$; $***P<0.001$.

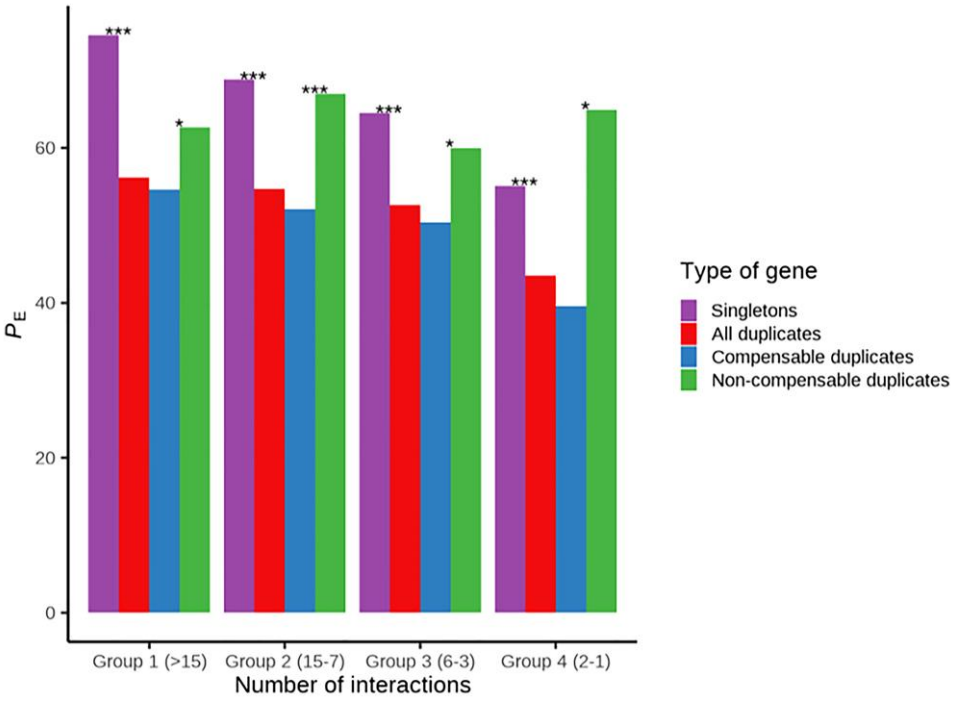

Fig. 4.—Differences in the percentage of essential genes (P_E) between singletons and duplicates, and between compensable and noncompensable duplicates, controlling for the number of protein–protein interactions. Fisher's exact test significance levels: $*P<0.05$; $**P<0.01$; $***P<0.001$.

both singletons and duplicates was higher for proteins involved in a high number of interactions (fig. 4).

P_E was significantly higher for singletons than for duplicates in all four groups, ranging from 74.58% in group 1 to 55.11% in group 4 for singletons and from 56.15% in group 1 to 43.52% in group 4 for duplicates (FET, $P=1.16\times10^{-10}$ for group 1, $P=2.94\times10^{-6}$ for group 2, $P=2.77\times10^{-5}$ for group 3, and $P=0.0001$ for group 4; fig. 4). Moreover, the P_E estimates for compensable and noncompensable duplicates are, respectively, 54.68% and 62.66% within group 1, 52.09% and 66.96% within group 2, 52.61% and 60.00% within group 3, and 39.58% and 64.88% within group 4, with significant differences in all groups (FET, $P=0.0250$ for group 1, $P=4\times10^{-5}$ for group 2, $P=0.0101$ for group 3, and $P=0.0430$ for group 4; fig. 4). These results confirm that the differences between the P_E of singletons and duplicates, and those between the P_E of compensable and noncompensable duplicates, are not due to differences in their number of protein–protein interactions.

Gene Developmental Function does Not Explain the Low Essentiality of Compensable Duplicates

Earlier research revealed that the mouse knockout data set is enriched in developmental genes, which results in biased P_E estimates, especially for duplicated genes (Makino et al. 2009). Developmental genes are more likely to be essential than nondevelopmental genes (Liao and Zhang 2007; Makino et al. 2009; Su et al. 2014; Kabir et al. 2019). The fraction of developmental genes was higher among compensable duplicates than among noncompensable duplicates (table 2). The fraction of developmental genes is thus unlikely to account for the lower essentiality of compensable duplicates. Nonetheless, in order to account for this possible effect, we divided our data set into two categories (developmental and nondevelopmental; $n=2{,}762$ and 5,768, respectively) and repeated our analyses within each category. In line with previous results (Makino et al. 2009; Su et al. 2014; Kabir et al. 2019), we found that developmental genes are more essential than nondevelopmental genes (fig. 5).

We found significant differences in the P_E of singletons and duplicates, and in the P_E of compensable and noncompensable duplicates, both within developmental (FET, $P=0.0061$ and $P=0.0069$, respectively) and within nondevelopmental genes (FET, $P<2.2\times10^{-16}$ and 1.33×10^{-12}, respectively; fig. 5). These results indicate that the differences between these groups cannot be explained by gene ontology (GO) annotation status relating to developmental function.

Compensable and Noncompensable Duplicates Differ in P_E Regardless of Whether They are Ohnologs or Small-Scale Duplicates

Previous studies have revealed a sampling bias in the mouse knockout data set toward ohnologs (duplicates resulting

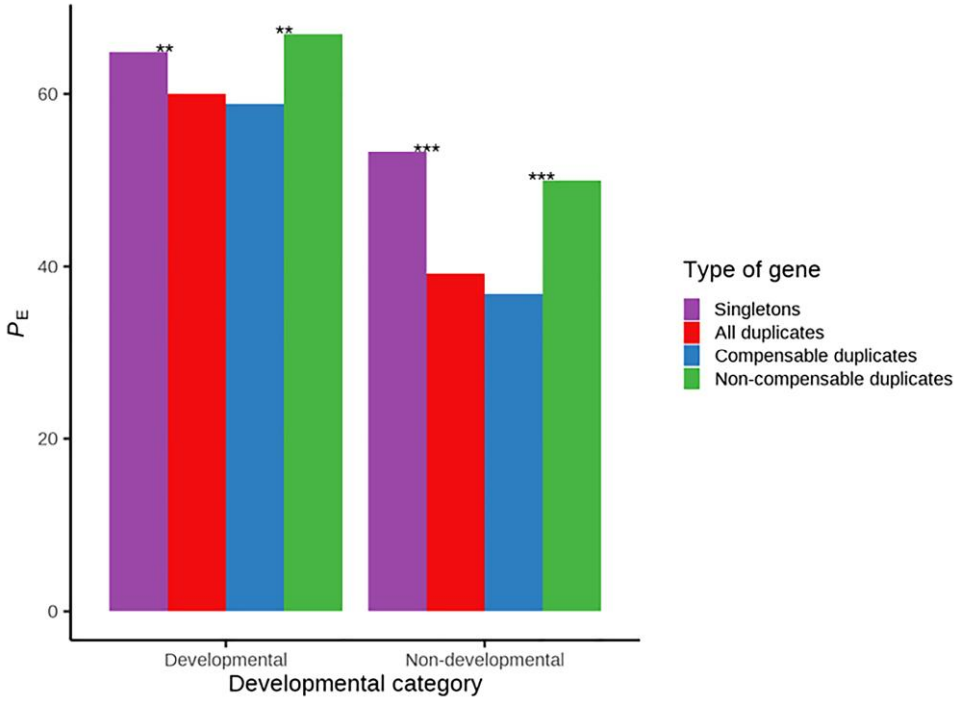

Fig. 5.—Differences in the percentage of essential genes (P_E) between singletons and duplicates, and between compensable and noncompensable duplicates, separately for developmental and nondevelopmental genes. Fisher's exact test significance levels: $*P<0.05$; $**P<0.01$; $***P<0.001$.

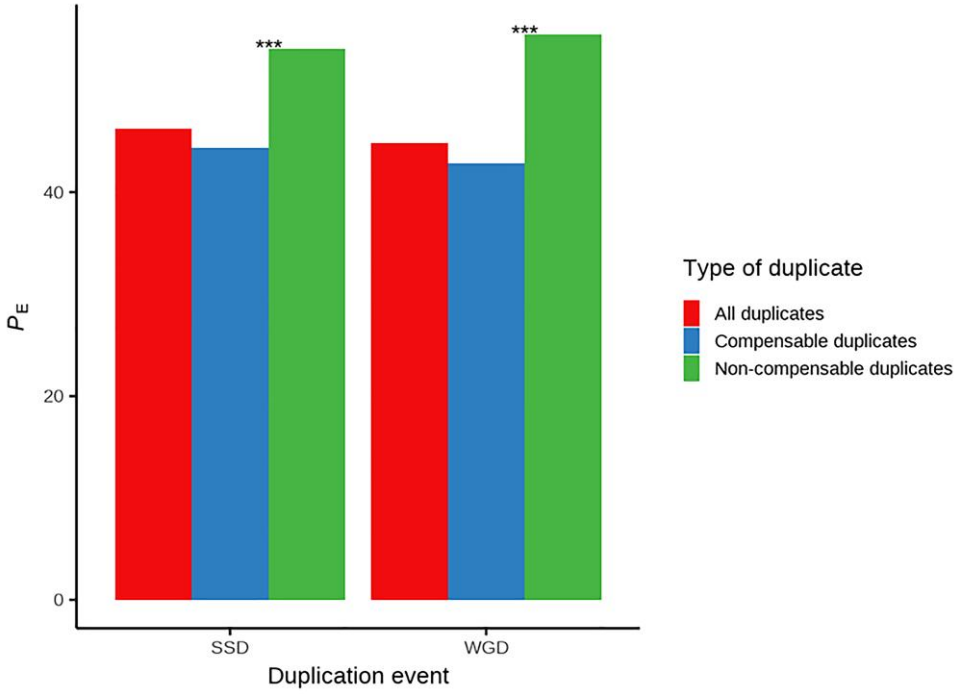

Fig. 6.—Differences in the percentage of essential genes (P_E) between compensable and noncompensable duplicates, separately for ohnologs (WGD) and small-scale duplicates (SSD). Fisher's exact test significance levels: $*P<0.05$; $**P<0.01$; $***P<0.001$.

from WGD events; Makino et al. 2009; Su et al. 2014). Compensable duplicates are enriched in ohnologs compared with noncompensable duplicates (table 2). As ohnologs are more likely to be essential than duplicates resulting from small-scale duplication (SSD) events (Makino et al. 2009), it is unlikely that this observation is driving the differences in P_E between compensable and noncompensable duplicates. Nonetheless, to test this possibility, we compared the P_E of compensable and noncompensable duplicates among WGD and SSD duplicates separately (fig. 6).

We found significant differences in both cases. Among WGD duplicates, compensable duplicates exhibited a P_E of 42.88% and noncompensable duplicates displayed a P_E of 53.44% (FET, $P = 3.79 \times 10^{-6}$). Among SSD duplicates, P_E was 44.40% and 54.09% for compensable and noncompensable duplicates, respectively (FET, $P = 7.33 \times 10^{-7}$). These results indicate that the type of duplication event (WGD vs. SSD) does not explain the differences in the P_E of essential and nonessential genes.

Multivariate Analyses Confirm that P_E Differs Between Compensable and Noncompensable Genes, Regardless of Their Differences in Confounding Factors

Our controls indicate that the differences in P_E between compensable and noncompensable duplicates are not due to differences in any of the controlling factors (gene age, sequence similarity to closest paralog, number of protein–protein interactions, function, or WGD) individually (figs. 2–6). Nonetheless, it is conceivable that these factors, combined, might be producing the observed differences in P_E. To account for potential effect, we conducted a multivariate binary logistic regression, using essentiality as the dependent variable and all potentially confounding factors, as well as the type of duplicate (compensable vs. noncompensable), as explanatory variables. All binary variables were encoded as dummy variables (see Materials and Methods).

Our analyses confirm that compensability has a negative effect on essentiality, independently of all other factors ($\beta = -0.4568$, $P = 1.67 \times 10^{-9}$; supplementary table S1, Supplementary Material online). Thus, the differences in the P_E of compensable and noncompensable genes are not due to the studied confounding factors. Ohnology also has a negative effect on essentiality, whereas developmental function, gene age, and number of protein–protein interactions have a positive effect, and sequence similarity with the closest paralog has no significant effect (supplementary table S1, Supplementary Material online). The two factors with the strongest effect are developmental category ($\beta = 0.8166$, $P < 2 \times 10^{-16}$) and type of duplicate ($\beta = -0.4568$, $P = 1.67 \times 10^{-9}$).

Discussion

Comparing the percent of essential genes (P_E) among singletons and duplicates in the genomes of several organisms has allowed researchers to infer the extent to which duplicates can compensate for the loss of their paralogs. If duplicates can often compensate for the loss of their paralogs, their P_E should be lower than that of singletons. This effect has been observed in all analyzed organisms (Gu et al. 2003; Conant and Wagner 2004; Hannay et al. 2008; Hanada et al. 2009; Wang, Birsoy, et al. 2015). Initial analyses in mouse, based on 2,899 and 3,872 genes, did not observe this effect (Liang and Li 2007; Liao and Zhang 2007), and only later analyses based on larger data sets and controlling for a number of confounding factors uncovered significant differences between the P_E of singletons and duplicates (Liang and Li 2009; Makino et al. 2009; Chen et al. 2012; Su et al. 2014; Acharya et al. 2015; Kabir et al. 2019). Analyzing a total of 8,530 genes with known knockout phenotypes, we found significant differences in the P_E of singletons and duplicates, either when controlling (figs. 2–6; supplementary table S1, Supplementary Material online) or not controlling (fig. 1) for potentially confounding factors. These results indicate that mouse duplicates can, in many cases, compensate for the loss of their paralogs, and that early studies were limited by biases in the mouse knockout data set.

Previous studies in yeast, worms, and plants have shown that the loss of a gene is more likely to be compensated by a paralog if the proteins encoded by both genes are highly similar at the sequence level (Gu et al. 2003; Conant and Wagner 2004; Hanada et al. 2009). Nonetheless, the opposite effect (Liao and Zhang 2007), or no effect (Su and Gu 2008; Su et al. 2014), has been found in mice, potentially due to biases in the data set (Liao and Zhang 2007; Su and Gu 2008; Makino et al. 2009; Su et al. 2014). Indeed, our analyses show that the essentiality of duplicates as a whole is largely unaffected by the amount of similarity with their closest paralogs (fig. 3); however, among noncompensable duplicates, those with low similarity with their closest paralogs exhibited a high P_E (fig. 3), in line with previous results in yeast, worms, and plants (Gu et al. 2003; Conant and Wagner 2004; Hanada et al. 2009).

We hypothesized that, in multicellular organisms, a gene's ability to compensate for the loss of a paralog would depend not only on the sequence of the encoded protein, but also on the tissues in which the gene is expressed. Mouse genes tend to diverge in expression patterns soon after duplication, but after that initial period of divergence expression patterns tend to be stable (Huerta-Cepas et al. 2011). For a gene (gene 1) to be able to compensate for the loss of one of its paralogs (gene 2), gene 1 needs to be expressed at least in the same set of tissues in which gene 2 is functional. For example, experiments in mice

have shown that loss of the *Myf5* gene, which is associated with skeletal muscle development, can be compensated by paralogs of the gene that are expressed in the same tissue (*Myod1*, *Myog*, and *Myf6*; Wang et al. 1996). In contrast, experiments in *Daucus carota* have shown that the function of the *ZDS1* gene, essential for early carrot development, cannot be compensated for by that of its *ZDS2* paralog, because only *ZDS1* is expressed in the developing carrot (Flores-Ortiz et al. 2020). For some examples of mouse genes whose ancestral expression patterns were partitioned among the duplicates after duplication, and thus remain as duplicates (i.e., examples of subfunctionalization at the level of tissue expression), see Lynch and Force (2000).

To test our hypothesis, we divided mouse duplicates into two categories: compensable duplicates (those with paralogs expressed in the same tissues) and noncompensable duplicates (those without paralogs expressed in the same tissues). In agreement with our hypothesis, the P_E of noncompensable duplicates is similar to that of singletons, whereas compensable duplicates exhibit a significantly higher P_E (fig. 1). The differences remain significant after controlling for a number of confounding factors (figs. 2–6, supplementary table S1, Supplementary Material online). These results indicate that, as we hypothesized, a gene's loss can mostly only be compensated by paralogs that are expressed in the same tissues. In addition, our results, combined with the fact that no clear association has been detected between essentiality and sequence similarity with the closest paralog (Liao and Zhang 2007; Su and Gu 2008; fig. 3), indicate that the compensation ability of mice genes is more dependent on the genes' expression patterns than on the sequence similarity of the encoded proteins.

Our analyses thus reveal the existence of two kinds of duplicates: noncompensable duplicates, which behave like singletons in terms of compensation (and thus exhibit a P_E similar to that of singletons), and compensable duplicates, which exhibit a substantially lower P_E. Previous studies have grouped both kinds of duplicates into the same category, which may have hindered the detection of differences between singletons and duplicates.

At least four factors may be homogenizing the P_E of singletons, compensable and noncompensable duplicates, thus hindering the detection of differences. First, proteins encoded by homologous genes do not necessarily carry out the same function, due to sequence divergence (Hahn 2009), which increases the essentiality of duplicates. Second, the loss of a gene can be compensated by mechanisms other than paralogs. For instance, the loss of a certain enzyme can be compensated by another enzyme catalyzing the same reaction (even if it is not encoded by a paralog), or by an alternative pathway (Gu et al. 2003; Papp et al. 2003; Deutscher et al. 2006; Harrison et al. 2007; Hanada et al. 2011), which may reduce the essentiality of both singletons and duplicates. Third, for the loss of a gene to be compensated by one of its paralogs, both genes need to be co-expressed during the same developmental period (Kabir et al. 2019). Last, a gene may not be functional or essential in all the tissues in which it is expressed (e.g., even though DNA methyltransferases Dnmt1, Dnmt3a, and Dnmt3b are ubiquitously expressed and essential in mice, conditional mutants deficient for the enzymes in certain tissues can survive; Dodge et al. 2005; Gao et al. 2011; Barau et al. 2016; Chen et al. 2021; Li et al. 2021), which may result in some genes being compensable even if they do not have paralogs covering all of the tissues in which it is expressed.

Conclusion

In conclusion, we show that mouse duplicates are less likely to be essential than singletons, but only if they have paralogs that are expressed in the same tissues. Thus, the divergence of expression patterns after gene duplication plays a critical role in determining whether genes can compensate for the loss of their paralogs. Indeed, paralog divergence at the level of expression patterns seems to be more determinant than divergence at the level of protein sequence. The existence of two groups of duplicates with significantly different essentialities (unrecognized until now), along with various biases in the mouse knockout data set (particularly in early versions) may have hindered the detection of differences in the essentiality of singletons and duplicates in earlier studies.

Materials and Methods

Genomic Expression and Essentiality Data

For each mouse protein-coding gene ($n = 22{,}175$), we retrieved a list of mouse paralogs from Ensemble's Biomart, release 90 (Kinsella et al. 2011). Genes with one or more paralogs were classified as duplicates ($n = 16{,}208$), whereas those without paralogs were deemed as singletons ($n = 5{,}967$).

We further classified duplicated genes into two categories (compensable duplicates and noncompensable duplicates) based on their patterns of protein expression. For each gene, protein abundances in eight adult organs/tissues (brain, brown adipose tissue, heart, kidney, liver, lung, pancreas, and spleen) were retrieved from the PaxDb database, version 4.0 (Wang, Herrmann, et al. 2015). For each tissue, the "integrated data set" was used. Genes that have, for each tissue in which it is expressed, at least one mouse paralog expressed in that tissue, were considered compensable ($n = 5{,}562$), whereas all other duplicated genes were deemed as noncompensable ($n = 1{,}204$). As an example, let us consider a hypothetical gene family with three members: gene 1 is expressed in the brain and the liver, gene 2 is expressed in the brain and

the muscle, and gene 3 is expressed in the liver. In this example, gene 1 would be compensable (by the combined action of genes 2 and 3), gene 2 would be noncompensable (since neither gene 1 nor gene 3 are expressed in the muscle), and gene 3 would be compensable (by gene 1).

Next, we retrieved phenotype information from all mouse genes from the Mouse Genome Informatics (MGI) database, release 6.10 (Eppig et al. 2017). Using this information, we classified the mouse genes into three categories: essential ($n = 4,215$), nonessential ($n = 4,515$), and genes without essentiality information ($n = 13,645$). Essential genes are those producing lethality (prenatal, perinatal or postnatal) or sterility in mice upon gene knockout. Genes without essentiality information were removed from all our analyses. For each gene category (singletons, duplicates, compensable duplicates, and noncompensable duplicates), we computed the proportion of essential genes (P_E) as the number of essential genes divided by the total number of genes with essentiality information (essential + nonessential).

Confounding Factors Data

For each mouse protein-coding gene, we collected data for five potentially confounding factors to assess whether they account for the difference in P_E between singletons and duplicates, and between compensable and noncompensable duplicates.

i. Similarity with the closest paralog: For each duplicated gene, we retrieved the percentage of sequence similarity with its closest paralog (at the level of amino acid sequence) from Ensembl's BioMart, release 98 (Kinsella et al. 2011).
ii. Gene age: For each gene, we collected the time of origin of its gene family from the ProteinHistorian database (Capra et al. 2012). We used the estimates based on the Wagner parsimony algorithm (Wagner 1961).
iii. Type of duplication event: We classified gene duplicates into two categories according to whether they originated from SSD or WGD events. Mouse genes listed as ohnologs in the OHNOLOGS v2 database (Singh and Isambert 2020) were deemed as WGD genes ($n = 2,539$), whereas all other duplicates were classified as SSD ($n = 4,227$).
iv. Connectivity: For each mouse gene, we obtained the number of protein–protein interactions from the BioGRID database, v4.4 (Oughtred et al. 2021), which lists protein interactions determined experimentally. Only physical interactions were included in our calculations. Protein connectivity information was available for 6,437 of the mouse genes.
v. Gene function: We classified genes as "developmental" or "nondevelopmental" using their GO "slim" annotations (Ashburner et al. 2000). We classified genes as developmental if they were annotated with terms GO:00075225 ("development of multicellular organisms") or GO:0030154 ("cell differentiation"), as in Makino et al. (2009).

Statistical Analyses

We compared a number of variables (essentiality and the potentially confounding factors listed above) between singletons and duplicates, and between compensable and noncompensable duplicates. We used the FET to test for differences in categorical variables (essentiality, type of duplication event, and gene function), and the MWU test to test for differences in noncategorical variables (gene age, connectivity, and sequence similarity with the closest paralog).

In order to test the possibility that the differences in P_E between different kinds of genes (singletons vs. duplicates, and compensable vs. noncompensable duplicates) may be due to differences in potentially confounding factors, we partitioned the genes according to these factors and re-evaluated the differences in P_E in each of the resulting gene groups. Based on their age, we partitioned genes into three groups. The first group contains genes originated between 0 and 361.2 Ma, the second group contains genes originated between 454.6 and 842 Ma, and the third group contains genes originated between 910 and 4,200 Ma. Next, using the number of protein–protein interactions, we partitioned genes into four categories with a similar number of genes (~1,600). Genes in group 1 have >15 interactions, genes in group 2 have 7–15 interactions, genes in group 3 have 3–6 interactions, and genes in group 4 have 1–2 interactions. Finally, based on the percentage of sequence similarity with the closest paralog, we divided genes into four groups with the same number of genes ($n = 1,649$). Group 1 contains genes with a percentage of similarity >60.20%, group 2 contains genes with a percentage of similarity between 60.19% and 46.28%, group 3 contains genes with a percentage of similarity between 46.27% and 32.25%, and group 4 group contains genes with a percentage of similarity below 32.24%. As the percentage of sequence similarity with the closest paralog is only applicable to duplicates, this variable was only used to confirm the differences between compensable and noncompensable duplicates.

To evaluate the effects of the duplicate type (compensable vs. noncompensable) and all five potentially confounding factors on essentiality simultaneously, we conducted a multivariate binomial logistic regression. This kind of regression simultaneously examines the effects of a number of explanatory variables (including binary and nonbinary variables) on a dependent binary variable. In this analysis, we used essentiality as the dependent variable, and all other variables (type of duplicate—compensable vs. noncompensable—plus the five

confounding variables) as explanatory variables. Binary variables were recoded using dummy variables (essential = 1, nonessential = 0, compensable = 1, noncompensable = 0, developmental = 1, nondevelopmental = 0, WGD = 1, and SSD = 0). The analysis was conducted using R (R Core Team 2018).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

This work was supported by grant MCB 1818288 from the National Science Foundation.

Data Availability

All data used in this work are publicly available, as described in Materials and Methods section.

Literature Cited

Acharya D, Mukherjee D, Podder S, Ghosh TC. 2015. Investigating different duplication pattern of essential genes in mouse and human. PLoS One 10:e0120784.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. Genome Biol Evol. 4:1263–1274.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Barau J, et al. 2016. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science 354:909–912.

Batada NN, Hurst LD, Tyers M. 2006. Evolutionary and physiological importance of hub proteins. PLoS Comput Biol. 2:e88.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput Biol. 8:e1002567.

Chen WH, Trachana K, Lercher MJ, Bork P. 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol Biol Evol. 29:1703–1706.

Chen DY, et al. 2021. Dnmt3a deficiency in the skin causes focal, canonical DNA hypomethylation and a cellular proliferation phenotype. Proc Natl Acad Sci U S A. 118(16):e2022760118.

Conant GC, Wagner A. 2004. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. Proc R Soc B Biol Sci. 271:89–96.

D'Antonio M, Ciccarelli FD. 2011. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 7:e1002029.

Deutscher D, Meilijson I, Kupiec M, Ruppin E. 2006. Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nat Genet. 38:993–998.

Dodge JE, et al. 2005. Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. J Biol Chem. 280:17986–17991.

Doherty A, Alvarez-Ponce D, McInerney JO. 2012. Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. Mol Biol Evol. 29:3563–3573.

Eppig JT, et al. 2017. Mouse Genome Informatics (MGI): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. Methods Mol Biol. 1488:47–73.

Flores-Ortiz C, et al. 2020. Differential role of the two ζ-carotene desaturase paralogs in carrot (*Daucus carota*): *ZDS1* is a functional gene essential for plant development and carotenoid synthesis. Plant Sci. 291:110327.

Gao Q, et al. 2011. Deletion of the de novo DNA methyltransferase Dnmt3a promotes lung tumor progression. Proc Natl Acad Sci U S A. 108:18061–18066.

Gu Z, et al. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered. 100(5):605–617.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol. 22:803–806.

Hanada K, et al. 2009. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. Genome Biol Evol. 1:409–414.

Hanada K, et al. 2011. Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. Mol Biol Evol. 28:377–382.

Hannay K, Marcotte EM, Vogel C. 2008. Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation. BMC Genomics 9(1):1–8.

Harrison R, Papp B, Pál C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci U S A. 104:2307–2312.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief Bioinform. 12:442–448.

Hughes AL, Friedman R. 2005. Gene duplication and the properties of biological networks. J Mol Evol. 61:758–764.

Jeong H, Mason SP, Barabási A-L, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Kabir M, Barradas A, Tzotzos GT, Hentges KE, Doig AJ. 2017. Properties of genes essential for mouse development. PLoS One 12(5):e0178273.

Kabir M, Wenlock S, Doig AJ, Hentges KE. 2019. The essentiality status of mouse duplicate gene pairs correlates with developmental co-expression patterns. Sci Rep. 9:1–12.

Kinsella RJ, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.

Li F, et al. 2021. Brown fat Dnmt3b deficiency ameliorates obesity in female mice. Life (Basel) 11(12):1325.

Liang H, Li W-H. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 23:375–378.

Liang H, Li W-H. 2009. Functional compensation by duplicated genes in mouse. Trends Genet. 25:441.

Liao B-Y, Zhang J. 2007. Mouse duplicate genes are as essential as singletons. Trends Genet. 23:378–381.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154(1):459–473.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet. 25:152–155.

Ohno S. 1970. Evolution by gene duplication. New York (NY): Springer Science & Business Media.

Oughtred R, et al. 2021. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 30:187–200.
Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.
Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. Mol Biol Evol. 23:30–39.
R Core Team. 2018. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. www.R-project.org.
Reguly T, et al. 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol. 5: 1–28.
Singh PP, Isambert H. 2020. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. Nucleic Acids Res. 48:D724–D730.
Su Z, Gu X. 2008. Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. J Mol Evol. 67:705–709.
Su Z, Wang J, Gu X. 2014. Effect of duplicate genes on mouse genetic robustness: an update. Biomed Res Int. 2014:758672.
Wagner WH. 1961. Problems in the classification of ferns. Rec Adv Bot. 1:841–844.
Wang T, Birsoy K, et al. 2015. Identification and characterization of essential genes in the human genome. Science 350:1096–1101.
Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15:3163–3168.
Wang Y, Schnegelsberg PN, Dausman J, Jaenisch R. 1996. Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin. Nature 379:823–825.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.
Zhu Y, Du P, Nakhleh L. 2012. Gene duplicability-connectivity-complexity across organisms and a neutral evolutionary explanation. PLoS One 7: e44491.

Associate editor: Yves Van De Peer