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

Wei-Hua Chen,[1] Kalliopi Trachana,[1] Martin J. Lercher,*[,2] and Peer Bork*[,1]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[2]Institute for Computer Science, Heinrich-Heine-University, Düsseldorf, Germany

***Corresponding author:** E-mail: bork@embl.de; lercher@cs.uni-duesseldorf.de.

Associate editor: Arndt von Haeseler

Abstract

Recently duplicated genes are believed to often overlap in function and expression. A priori, they are thus less likely to be essential. Although this was indeed observed in yeast, mouse singletons and duplicates were reported to be equally often essential. This contradiction can only partly be explained by experimental biases. We herein show that older genes (i.e., genes with earlier phyletic origin) are more likely to be essential, regardless of their duplication status. At a given phyletic gene age, duplicates are always less likely to be essential compared with singletons. The "paradoxical" high essentiality among mouse gene duplicates is then caused by different age profiles of singletons and duplicates, with the latter tending to be derived from older genes.

Key words: gene essentiality, yeast, mouse, phyletic age, linking genotype to phenotype.

In model organisms such as mouse and yeast, phenotypic changes caused by single-gene mutations were assayed on a genome-wide scale (Kelly et al. 2001; Blake et al. 2011). Of particular interest are essential genes, whose removal results in death or infertility. Many expressed genes performing important molecular functions are nonessential. In these cases, it is likely that the gene deletion can be partially compensated by another gene with overlapping function and expression.

Gene duplication is believed to be an important source of such functional redundancy (Ohno 1970). Accordingly, the proportion of essential genes (P_E) among duplicates is much lower than among singletons in yeast (Gu et al. 2003). However, this expected trend was not confirmed in mouse, where the proportion of essentials among duplicates is comparable (Liao and Zhang 2007) or even lower (table 1) than among singletons.

The contradicting results in mouse were initially interpreted as evidence against widespread functional redundancy of duplicates (Liao and Zhang 2007); this interpretation was hotly disputed (Su and Gu 2008; Liang and Li 2009; Makino et al. 2009). At that time (Liao and Zhang 2007; Su and Gu 2008; Liang and Li 2009; Makino et al. 2009), only ~5,000 mouse genes had been tested in knockout experiments. Biases were expected in this subset of mouse genes, as genes with known severe mutational phenotypes had been selected with higher priority. Two follow-up studies (Su and Gu 2008; Makino et al. 2009) discovered that the knockout data were further enriched in genes derived from old duplications and in developmental genes; after correcting these biases, the overall P_E in duplicates became statistically significantly lower than that in singletons (Su and Gu 2008; Makino et al. 2009).

However, the authors did not explore two immediate conclusions from their studies (Su and Gu 2008; Makino et al. 2009): 1) genes derived from old duplications are more likely to be essential than singletons and 2) developmental duplicates are more likely to be essential than developmental singletons (or indeed singletons as a whole). Both conclusions hold true in the older as well as the latest versions of the mouse phenotype data sets (table 1). This appears to again contradict the duplication-functional redundancy concept, and we thus consider the issue unresolved.

What factors other than duplication status affect gene essentiality? Developmental genes are more likely to be essential than nondevelopmental genes (Makino et al. 2009), but this should apply to duplicates and singletons alike. It was also suggested that hubs in protein–protein interaction networks are more likely to be essential (Jeong et al. 2001); however, this observation probably reflects biases toward proteins in large essential protein complexes (Zotenko et al. 2008).

Previous studies indicated that the phyletic origin (age) of genes, defined by the evolutionarily most distant species group where homologs can be found (Wolf et al. 2009), is correlated with several gene features (Hao et al. 2010). Genes that originated early tend to be conserved across species, highly and broadly expressed, and broadly useful (Hao et al. 2010). Thus, we hypothesized that knocking out phyletically old genes is more likely to have severe phenotypic effects: old genes should be more often essential.

To test this idea, we classified mouse and yeast genes into different age groups according to their earliest phyletic

Letter

Table 1. Proportions of Essential Genes in Different Gene Categories in the Two Phenotypical Data Sets for Mouse.

	Proportion of Essential Genes (%)	
Categories	Current Data Set[a]	Data from Makino et al. (2009)
All genes	**43.3**	**42.07**
Duplicates	**43.9 (41.6[b])**	**41.92**
Singletons	**41.1**	**42.61**
All developmental genes	**62.53**	**59.51**
Developmental duplicates	**64.75**	**60.9**
Developmental singletons	**53.1**	**53.36**
Old duplications (K_s[c] ≥ 2)	**47.31**	**44.94**

[a] MGI 4.4 (October 2010).

[b] If only genes with valid phyletic ages are used.

[c] If a gene has multiple duplicates, all pairwise K_s (the number of synonymous substitutions per synonymous site) between this gene and its duplicates will be calculated, and the lowest K_s value is used. Synonymous substitutions in most genes with $K_s \geq 2$ will have reached saturation, and hence, the corresponding genes will tend to be older than genes with $K_s < 2$.

origin (see Materials and Methods). We classified genes as specific to one of five taxonomic groups for yeast (fig. 1A) and six broad taxonomic groups for mouse (fig. 1C). Because of the large differences between yeast and mouse, we did not attempt any direct cross-species comparisons and did not attempt to map their histories onto a common timescale.

We found that within each age group, the P_E among singletons is always higher than among duplicated genes; this is true both in mouse and in yeast (fig. 1). Thus, duplicated genes indeed tend to be less likely essential. Furthermore, for both singletons and duplicated genes, the fraction of essential genes increases with increasing age; thus, older genes are indeed more likely to be essential (fig. 1). The trends observed in figure 1 are reproduced when restricting the analysis either to developmental genes or to nondevelopmental genes (Supplementary figs. S1 and S2, Supplementary Material online; for the raw data, see Supplementary table, Supplementary Material online).

Gene duplicates have two ages: the age of the gene family (phyletic age; fig. 1) and the age of the most recent duplication event (duplication age). The effect of phyletic age is likely similar between duplicates and singletons. In addition, functional redundancy is expected to be strongly affected by the age of the duplication event, as duplicates derived from ancient duplications are more likely to be essential than genes derived from recent duplications (Su and Gu 2008). In mouse gene duplicates, essentiality reaches a plateau in the Fungi/Metazoa group and does not increase further in the two older age groups. According to the reasoning above, this plateau might be caused by a comparably young duplication age. We indeed find that the two oldest groups contain higher fractions of younger duplicates than the "Fungi/Metazoa" group (Supplementary fig. S3, Supplementary Material online).

In each phyletic age group, duplicates are less likely to be essential than singletons (fig. 1). Why then is the same

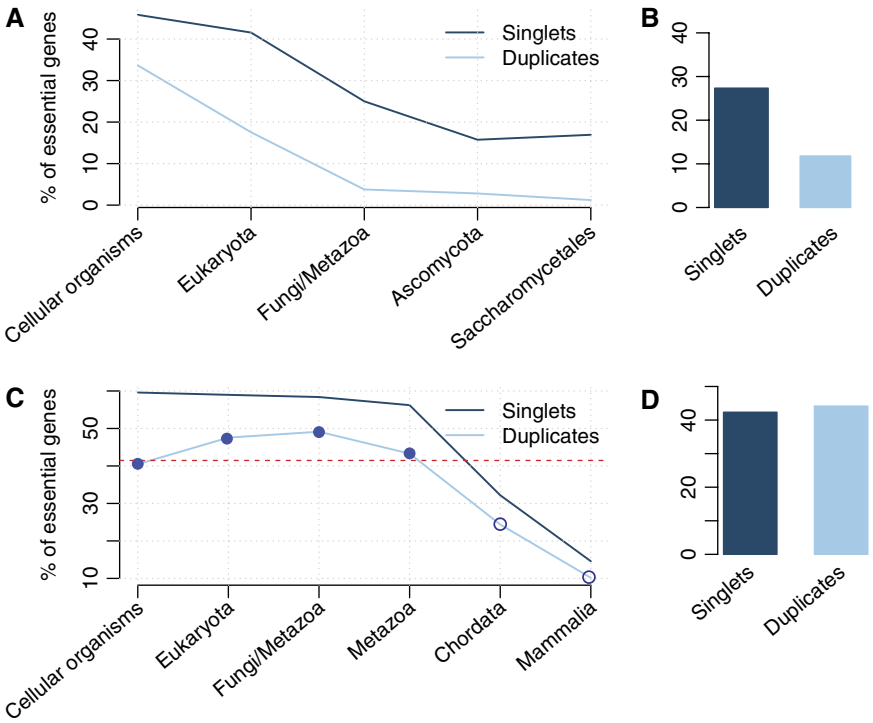

Fig. 1. In both yeast (*A*) and mouse (*C*), genes with more recent phyletic origins are less likely to be essential, as are duplicated genes compared with singletons of the same phyletic age. However, ignoring age, the overall proportion of essential genes in singletons is higher in yeast (*B*) but lower in mouse (*D*) compared with duplications. Filled circles in (*C*) indicate that the proportion of essential genes in the corresponding duplication groups is higher than or closer to the overall P_E in singletons (41.1%; the dashed horizontal line); whereas hollow circles indicate that P_E is lower.

not true when disregarding age, as done in previous studies (Liao and Zhang 2007; Su and Gu 2008; Liang and Li 2009; Makino et al. 2009)? This is in fact an instance of Simpson's paradox (Simpson 1951), which can arise when the dependence of two categorical variables (essentiality and duplication status) on a third variable (phyletic age) is disregarded. To illustrate the mathematics behind this paradox, we divided the six age groups of duplicated genes into two parts reflecting a very coarse definition of age: one including four age groups (the "old part," filled circles in fig. 1C) that mostly have higher P_Es than the overall singletons and other including the remaining two groups with lower P_Es (the "young part," open circles in fig. 1C). This partitioning results in a higher overall proportion of essential genes in the old duplicate part ($P_E^{old}=44.89\%$) are higher compared with the overall singletons ($P_E^{singleton}=41.1\%$), whereas the corresponding proportion in the young duplicate part ($P_E^{young}=22.97\%$) is lower. The overall P_E of duplicates regardless of age can be calculated from this as a weighted average:

$$P_E = f_{old} \times P_E^{old} + f_{young} \times P_E^{young},$$

where f_{old} and f_{young} are the fraction of duplicates contained in the "old" and "young" parts, respectively (with $f_{old}+f_{young}=1$) (for more details, see Supplementary text and Supplementary table, Supplementary Material online). In theory, the overall P_E could be as high as 44.89% or as low as 22.97%, depending on the values of f_{old} and f_{young}. In our study, we found that the vast majority of duplicates was derived from old gene families ($f_{old}=84.66\%$), resulting in an overall P_E of 41.6% for duplicates (see Supplementary table, Supplementary Material online). Thus, the surprising result of a higher essentiality among mouse duplicates compared with singletons is caused by a different age profile of singletons and duplicated gene families.

Our results differ significantly from a recent publication on *Drosophila melanogaster* (Chen et al. 2010). Based on RNAi knockdowns of ~440 genes, Chen et al. found that ~30% of young genes (<35 myr) were essential compared with ~35% of old genes (>40 myr). The authors concluded that "young genes are as essential as old genes in terms of viability" (Chen et al. 2010). We reanalyzed their data using our methods, which differ from those of Chen et al. in age classification and in the separate analysis of duplicates and singletons (for the raw data, see Supplementary table, Supplementary Material online). We found that the proportion of essential genes in both singletons and duplicates in general increases with increasing age in the five age groups, with some fluctuations (Supplementary fig. S4, Supplementary Material online). However, the P_E in duplicates is not always lower than in singletons of similar age, and the differences are not statistically significant (Fisher's exact test, all comparisons $P > 0.05$; Supplementary table, Supplementary Material online), probably due to the small size of the data set. Since only a small number (~3.2%) of *D. melanogaster* genes have been tested (Chen et al. 2010), our findings regarding *D. melanogaster* are not yet conclusive.

Materials and Methods

We determined the phyletic origins of genes from yeast, mouse, and fly using a method described in Wolf et al. (2009) with modifications (for more details, see Supplementary text, Supplementary Material online, and for the results, Supplementary table, Supplementary Material online). We separated genes into singletons and duplicates as previously described (Liao and Zhang 2007; Makino et al. 2009). We grouped duplicates into gene families using a clustering-based method (Markov cluster algorithm [MCL]; Enright et al. 2002) and then used the most ancient origin of all members as the age of the corresponding family.

We obtained the phenotypic data for the three species from online gene essentiality database (Chen et al. 2012), which were originally published by the Saccharomyces Genome Deletion Project (Cherry et al. 1997), the Mouse Genome Informatics (Blake et al. 2011), and the authors of Chen et al. (2010), respectively. We restricted further analyses to genes that were tested in these phenotypic data sets.

Supplementary Material

Supplementary text, Supplementary figures, and Supplementary table are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

SysteMtb grant #241587 to P.B. provided funding for the open access license.

References

Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT. Mouse Genome Database G. 2011. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* 39:D842–D848.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. *Science* 330:1682–1685.

Chen WH, Minguez P, Lercher MJ, Bork P. 2012. OGEE: an online gene essentiality database. *Nucleic Acids Res.* 40:D901–D906.

Cherry JM, Ball C, Weng S, et al. (11 co-authors). 1997. Genetic and physical maps of Saccharomyces cerevisiae. *Nature* 387: 67–73.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575–1584.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63–66.

Hao L, Ge X, Wan H, Hu S, Lercher M, Yu J, Chen W-H. 2010. Human functional genetic studies are biased against the medically most relevant primate-specific genes. *BMC Evol Biol.* 10:316.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411:41–42.

Kelly DE, Lamb DC, Kelly SL. 2001. Genome-wide generation of yeast gene deletion strains. *Comp Funct Genomics.* 2:236–242.

Liang H, Li WH. 2009. Functional compensation by duplicated genes in mouse. *Trends Genet.* 25:441–442.

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

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: Springer-Verlag.

Simpson EH. 1951. The interpretation of interaction in contingency tables. *J R Stat Soc Ser B.* 13:238–241.

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.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. Inaugural article: the universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A.* 106: 7273–7280.

Zotenko E, Mestre J, O'Leary DP, Przytycka TM. 2008. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol.* 4:e1000140.