J Mol Evol
DOI 10.1007/s00239-008-9170-9

Predicting the Proportion of Essential Genes in Mouse Duplicates Based on Biased Mouse Knockout Genes

Zhixi Su · Xun Gu

Received: 13 May 2008 / Accepted: 22 September 2008
© Springer Science+Business Media, LLC 2008

Abstract In the yeast or nematode, the proportion of essential genes in duplicates is lower than in singletons (single-copy genes), due to the functional redundancy. One may expect that it should be the same in the mouse genome. However, based on the publicly available mouse knockout data, it was observed that the proportion of essential genes in duplicates is similar to that in singletons. The most straightforward interpretation, as claimed in a recent study, is that duplicate genes may have a negligible role in the mouse genetic robustness. Here we show that in the current mouse knockout dataset, recently duplicated genes have been highly underrepresented, leading to an overestimation of the proportion of essential genes in duplicates. After estimating the duplication time of mouse duplication events, we have developed a simple bias-correcting procedure and shown that the bias-corrected proportion of essential genes in mouse duplicates is significantly lower than that in singletons.

Keywords Proportion of essential genes · Genetic robustness · Gene knockout · Gene duplication

Z. Su · X. Gu
Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

X. Gu
School of Life Sciences, Center for Evolutionary Biology, Fudan University, Shanghai 200433, China

X. Gu (✉)
Department of Genetics, Development and Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, 536 Science II, Ames, IA 50011, USA
e-mail: xgu@iastate.edu

Functional compensation of duplicate (paralogous) gene has been thought to be an important factor in the genetic robustness (Conant and Wagner 2004; Dean et al. 2008; Gu 2003; Gu et al. 2003; Harrison et al. 2007; Ihmels et al. 2007; Kamath et al. 2003; Winzeler et al. 1999). This is because the existence of a close paralogue in the same genome could increase the chance that null mutation of a given gene, by gene deletion, knockout, or RNAi technology, has little effect on organismal fitness, as called nonessential. Duplicate genes are indeed less essential than single-copy genes in both the yeast and the nematode (Gu et al. 2003; Conant and Wagner 2004; Kamath et al. 2003); two recent papers, however, reported that the proportion of essential genes (P_E) for singletons is similar to that for duplicates in mouse, based on the currently available mouse knockout phenotypes (Liang and Li 2007; Liao and Zhang 2007). However, they came up with different explanations. Liang and Li (2007) suggested that the potential compensatory role of gene duplication may have been counteracted by another factor—the more intrinsic importance of the duplicated genes. Instead, Liao and Zhang (2007) argued that duplicated genes may have a negligible role in mouse genetic robustness.

Liao and Zhang (2007) examined several factors that might have led to overestimation of the P_E for duplicates. The null hypothesis in their study is that the P_E of the knockout duplicates sample is unbiased. Since this null hypothesis cannot be rejected by the survey of a series of genomic measures, they conclude that singletons and duplicates are equally likely to be essential, implying that the compensational role of duplicates in mouse genetic robustness is negligible. In this study, we revisit this issue by careful treatment of the updated mouse knockout dataset. Our analysis clearly indicates that the sampling bias for ancient duplicates in knockout genes was the main reason for the overestimation of the P_E of duplicates in mouse. Moreover, we used the age distribution of mouse duplicate genes to obtain a bias-corrected P_E and tested whether it is significantly lower than in singletons.

Materials and Methods

Mouse Knockout Data

The mouse phenotype and genotype association file (MGI_PhenoGenoMP.rpt) was downloaded from Mouse Genome Informatics (MGI 3.54; ftp://ftp.informatics.jax.org; release 10/23/2007). Only the phenotypic annotations of null mutation homozygotes that were generated by target deletion or gene trap were extracted for further analysis, excluding all phenotypic annotations due to multiple gene knockout experiments. In total 4123 genes with phenotypic information were extracted from this file. We then classified these genes into 1921 essential genes and 2202 nonessential genes. As described by Liang and Li (2007), we defined an essential gene as a gene whose knockout phenotype is annotated as lethality (including embryonic, perinatal, and postnatal lethality) or infertility.

Dating Duplication Age of Mouse Duplicate Genes

We followed the method of Gu et al. (2002) to identify duplicate genes and single-copy genes, and chose the FASTA not-self best hit of a duplicate gene as its closest parologue (Pearson 2000). We have developed an analytical pipeline to estimate the duplication age (time) between each mouse duplicate gene and its cloest paralogue on a large scale, using the split time between mouse and zebrafish (430 million years ago [mya]) as a calibration (Fig. 1). By this method, the duplicate ages between each of 9503 mouse genes and its closest paralogues were estimated (whole-genome set). Among these, 2260 genes were knockout target genes (knockout set). In addition, we used several other speciation events as calibrations, e.g., the splits of mammal-bird (310 mya) and primate-rodent (80 mya), and found that our main results are robust (not shown). To be concise we present here the results based on the mammal-zebrafish split time calibration.

Predicting P_E in the Mouse Genome: Bias Correction

In statistics, the proportion of essential genes in duplicates (P_E) estimated from a sample of mouse knockout genes can be considered an unbiased estimate for the genome value, provided that the assumption of random sampling holds. While the expectation of P_E for singletons is $1 - B$, where B is the proportion of genetic buffering, the expectation of P_E for duplicates is $(1 - B)A$, where A is the mean effect of duplicate compensation. Given the age distribution of mouse duplicates, denoted $f(t)$, and the age-dependent effect of duplicate compensation, denoted $A(t)$, theoretically we have $A = \int A(t)f(t)dt$. Since mouse knockout genes were apparently not randomly selected with respect to the age distribution of duplicates, it is necessary to examine whether the potential sampling bias may have affected our estimation.

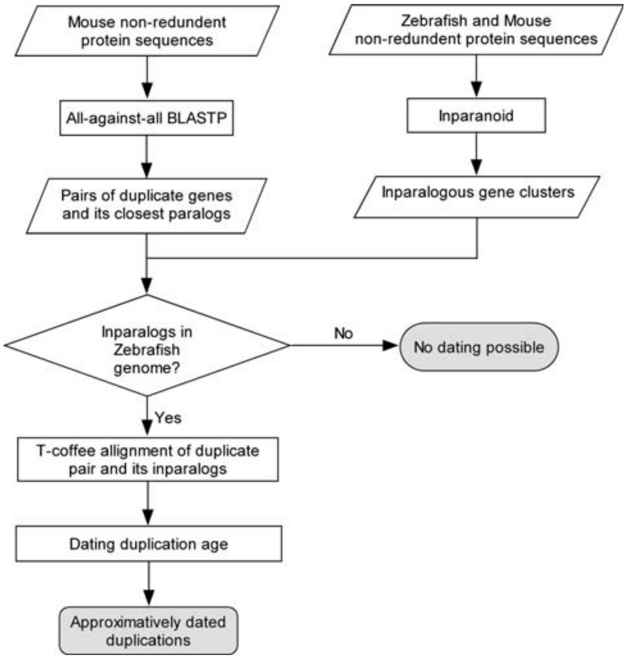
Mouse non-redundent protein sequences
Zebrafish and Mouse non-redundent protein sequences
All-against-all BLASTP
Inparanoid
Pairs of duplicate genes and its closest paralogs
Inparalogous gene clusters
Inparalogs in Zebrafish genome?
No
No dating possible
Yes
T-coffee allignment of duplicate pair and its inparalogs
Dating duplication age
Approximatively dated duplications

Fig. 1 Diagram for determining the evolutionary age of a mouse duplicate pair

We suspect that the underrepresentation of young duplicates in mouse knockout phenotypes could mislead our understanding of the duplication effect (A) on mouse genetic robustness. To test this claim we designed the following analytical procedure: on the basis of duplicate age (t), M duplicate genes in the knockout sample are grouped into several (n) bins, each of which has a 100 million-year interval. For each bin i we calculate the gene frequency ($f_i = M_i/M$, where M_i is the number of [knockout] duplicate genes in bin i) and the proportion of essential genes by $P_{E,i} = m_i/M_i$, where m_i is the number of essential genes in bin i. Assume that the mouse genome frequency for each age bin is known, denoted g_i. If the correlation between P_E and age t in the knockout sample is statistically significant, differences in gene frequency between the sample (f_i) and the genome (g_i) point to the potential resource that may cause the bias in P_E estimation. Under the assumption that the same P_E-t correlation holds in the mouse genome, the bias-correcting predicted P_E in the mouse genome is therefore calculated by

$$P_E^* = g_1 P_{E,1} + g_2 P_{E,2} \cdots + g_n P_{E,n} \tag{1}$$

For each bin i, the expectation of $P_{E,i}$ is $(1 - B)A_i$, where A_i is the mean effect of duplicate compensation at this age interval. Hence, the expectation of P_E^* is $(1 - B)A$ if the discrete genome frequency (g_i) is an unbiased representation of age distribution of duplicates $A(t)$ so that $A = g_1A_1 + g_2A_2 \cdots + g_nA_n$. Note that the observed P_E in the sample of knockout duplicates can be written

$$P_E = f_1 P_{E,1} + f_2 P_{E,2} \cdots + f_n P_{E,n} = m/M \tag{2}$$

where $m = m_1 + \cdots + m_n$ is the number of essential duplicate genes in the knockout sample. Obviously, if the knockout sample is large enough to cover the whole mouse genome, the bias in P_E estimation would be trivial, which can be estimated by m/M without knowing the ages of duplicate genes. For the current knockout sample, which covers only ~15% of mouse genes, P_E^* equals the observed P_E only if the duplicate gene frequency of each bin is the same between the knockout sample and the genome, i.e., $f_i = g_i$. Finally, from Eq. 1, one can calculate the sampling variance of P_E^* as follows:

$$Var(P_E^*) = g_1^2 Var(P_{E,1}) + g_2^2 Var(P_{E,2}) \cdots + g_n^2 Var(P_{E,n}) \tag{3}$$

where the sampling variance in each bin is $Var(P_{E,i}) = P_{E,i}(1 - P_{E,i})/M_i$. It should be noted that in Eq. 3 we consider only the sampling variance of $P_{E,i}$, neglecting the intrinsic dependence between any $A_{,i}$ and $A_{,j}$. As a result, $Var(P_E^*)$ may differ under various grouping cutoffs. Nevertheless, we found that the difference is usually small and does not affect our main result (not shown).

It is practically useful to estimate the proportion of essential genes in young duplicate genes. To this end, we compute $P_E^*(t^*)$, the proportion of essential genes for duplicate genes which are younger than a given age t^* that corresponds to the n^*th age bin. Let a^* be the proportion of 'ancient' duplicates with ages greater than t^* in the mouse genome. Thus, to estimate $P_E^*(t^*)$, Eq. 1 should be modified as follows:

$$P_E^*(t^*) = [g_1 P_{E,1} + g_2 P_{E,2} \cdots + g_{n^*} P_{E,n^*}]/(1 - a^*) \tag{4}$$

Obviously, the sampling variance of $P_E^*(t^*)$ is given by

$$Var[P_E^*(t^*)] = [g_1^2 Var(P_{E,1}) + g_2^2 Var(P_{E,2}) \cdots + g_{n^*}^2 Var(P_{E,n^*})]/(1 - a^*)^2 \tag{5}$$

Results and Discussion

Consistent with previous studies (Liang and Li 2007; Liao and Zhang 2007), we found no statistical difference in P_E between singletons and duplicates (47% vs. 46.3%; $p > 0.05$). Because most mouse knockout experiments have been carried out by individual laboratories for finding detectable knockout phenotypes, one may suspect that recently duplicated genes have been purposely avoided to minimize the experimental cost of negative-phenotype results (Liang and Li 2007; Liao and Zhang 2007). Hence, it was not surprising that recently duplicated genes may have been underrepresented in the mouse knockout database. If this is the case, the sampling bias could be the most obvious reason to explain why there was no statistical difference in P_E between mouse singletons and mouse duplicates.

We conducted a direct comparison of duplication age (mya) of mouse duplicate genes between the whole-genome set and the knockout set (Table 1).

Apparently, the ages of most duplicates in the mouse knockout dataset were dated at about 500–700 mya, and recently duplicated genes, say, <100 mya, were seriously underrepresented in the mouse knockout set: 1.4% in the knockout set versus 19.6% in the mouse genome set. In other words, the sampling bias toward ancient duplicates in the currently available mouse knockout target genes is nontrivial.

Since young duplicates are expected to have high degrees of functional compensation between them, resulting in a low proportion of essential genes (P_E), the age bias in mouse knockout duplicates may cause an overestimation of P_E in mouse duplicates. To avoid this bias, we calculated P_E in each age interval of 100 million years (age bin), respectively. As reported in Table 1, we found a significantly positive P_E/age(t) correlation ($p < 0.001$, χ^2 test). Apparently, the ancient duplicates may have undergone substantial functional divergence so that they have lost the capacity for functional compensation. In contrast, the young duplicates, those duplicated around the mammalian radiation or during the rodent lineage, are expected to make significant contributions to gene robustness in the current mouse genome. Therefore, the proportion of essential genes (P_E) in young duplicates is much lower than that of singletons. When young duplicates were considerably underrepresented in the mouse knockout dataset, it is actually not very surprising that the observed P_E in this biased sample of mouse knockout duplicates was close to that for single-copy genes.

Despite the small and biased sampling problem of mouse knockout genes (only ~15% of the genome), Liao and Zhang (2007) claimed that the estimated P_E in mouse knockout duplicates was almost unbiased, after having examined many genetic measures such as sequence and expression divergence. We explain their result from the view of statistics. For any genetic measure X, the procedure of analysis includes two tests: (i) whether the measure X differs significantly between the (knockout) sample and the mouse genome, and (ii) whether X is significantly correlated with

Table 1 Duplicate age (t) distribution of mouse knockout gene set and genome set

	$t < 100$	$100 \leq t < 200$	$200 \leq t < 300$	$300 \leq t < 400$	$400 \leq t < 500$	$500 \leq t < 600$	$600 \leq t < 700$	$700 \leq t < 800$	$800 \leq t < 900$	$900 \leq t < 1000$	$t \geq 1000$
No. of knockout genes	32	34	64	75	256	419	424	310	153	106	387
Frequency in knockout set (f_i)	0.014	0.015	0.028	0.033	0.113	0.185	0.188	0.137	0.068	0.047	0.171
No. of essential genes	7	10	18	24	99	196	210	166	77	59	216
P_E	0.219	0.294	0.281	0.32	0.387	0.468	0.495	0.535	0.503	0.557	0.558
Frequency in whole-genome set (g_i)	0.196	0.057	0.043	0.042	0.074	0.131	0.134	0.092	0.054	0.034	0.143

P_E (in the knockout sample). As shown in Eqs. 1 and 2, if one of these two tests shows a nonsignificant result, one may exclude X from the potential sources of the bias. In the case where the genetic measure X is the duplication age, both tests show significant results (Table 1), suggesting that the null hypothesis that P_E in a sample of knockout duplicates is unbiased must be rejected.

Liao and Zhang (2007) indeed found that the frequency distribution of protein sequence divergence (d) between a gene and its closest paralogue differs significantly between the (knockout) sample and the mouse genome. That is, duplicates in the knockout dataset are clearly underrepresented in low d ranges, compared with the genome set. Also, a weak negative correlation between P_E and d has been discovered. Hence, they inferred an underestimation of P_E in duplicates due to the sampling bias, which is counterintuitive. We have carefully examined this issue in our updated dataset, and found no correlation between P_E and d (not shown). It seems likely that the negative correlation between P_E and d observed by Liao and Zhang (2007) could be an artifact of the data noise. Indeed, many paralogous genes that were very closely related to mouse knockout genes in the dataset of Liao and Zhang (2007) have been removed from the updated Ensemble database.

Since no other genetic measure examined by Liao and Zhang (2007) showed significant results in both tests, the authors argued that their conclusion of a negligible role of mouse duplicates in genetic robustness was probably not caused by any sampling bias. Their conclusion holds until this null hypothesis can be rejected by some new genetic measures, as shown by the current analysis. The duplication age (t) shows statistically significant differences between the sample and the genome, and the positive P_E-age correlation provides a theory for the potential resource of sampling bias in P_E. We used the duplicate age (t) to obtain a corrected P_E in duplicates of the mouse genome (see Materials and Methods). Finally, we predicted that $P_E = 41.7\%$ for all duplicate genes, which is impressive compared to the $P_E = 46.3\%$ observed in sample duplicates ($p < 10^{-4}$, χ^2 test) and $P_E = 47\%$ in sample singletons ($p < 10^{-4}$) (Fig. 2). We also found that P_E in young duplicates is much lower than in all duplicates and sample singletons. For instance, we estimated $P_E = 25.2\%$ and 24.2% for those duplicates with a duplicate age no greater than 400 and 300 mya (Fig. 2), respectively, indicating that those young duplicates that duplicated after the tetrapod-teleost split (~430 mya) have a significant effect on gene robustness in the mouse genome.

In short, our analysis indicates that the sampling bias for ancient duplicate age of knockout genes caused the overestimation of P_E in mouse duplicates. Therefore, the similar proportions of essential genes between singletons and duplicates in the currently biased sample should not be

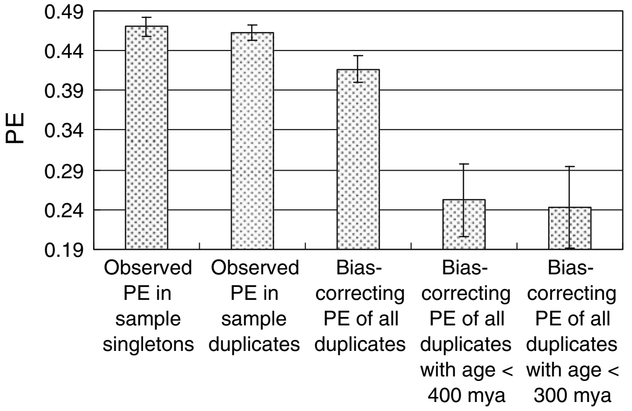

Fig. 2 Proportion of essential genes (P_E) in different gene categories in the mouse genome. Error bars show 1 SE

taken as evidence supporting the claim that the role of functional compensation for duplicate genes in the mouse is negligible. Besides, it is intriguing that very ancient duplicate genes have an even higher percentage of essential genes than singletons (Table 1), suggesting that duplicate genes might be more important (Liang and Li 2007). Indeed, our further extensive analysis indicates that duplicate genes play an important role in mouse genetic robustness (unpublished data).

Acknowledgments This work was supported in part by grants (30428003 to X.G., 30700140 to Z.S.) from the National Natural Science Foundation of China, and grants from Iowa State University and Fudan University (to X.G.), and supported by the Shanghai Leading Academic Discipline Project (Project No. B111). We are grateful to Wen-Hsiung Li, Han Liang, and Jianzhi Zhang for critical comments on the early version of the manuscript.

References

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

Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4(7):e1000113

Gu X (2003) Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet 19:354–356

Gu Z, Cavalcanti A, Chen F-C, Bouman P, Li W-H (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol 19:256–262

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

Harrison R, Papp B, Pal C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci USA 104:2307–2312

Ihmels J, Collins SR, Schuldiner M, Krogan N, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3:86

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–237

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

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

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285:901–906