19 Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584
20 Makova, K.D. and Li, W.H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13, 1638–1645
21 Prince, V.E. and Pickett, F.B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* 3, 827–837
22 Kirschner, M. and Gerhart, J. (1998) Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8420–8427
23 Freilich, S. *et al.* (2006) Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol.* 7, R89
24 Kellis, M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624
25 He, X. and Zhang, J. (2006) Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.* 23, 144–151

0168-9525/$ – see front matter.
doi:10.1016/j.tig.2007.04.005

Mouse duplicate genes are as essential as singletons

Ben-Yang Liao and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan, 1075 Natural Science Building, 830 North University Avenue, Ann Arbor, MI 48109, USA

Duplicate genes in mouse are widely thought to have functional redundancy, and to be less essential than singleton genes. We analyzed nearly 3900 individually knocked out mouse genes and discovered that the proportion of essential genes is ~55% in both singletons and duplicates. This suggests that mammalian duplicates rarely compensate for each other, and that the absence of phenotypes in mice deficient for a duplicate gene should not be automatically attributed to paralogous compensation.

Duplicates, singletons and redundancy

Duplicate genes occur in all organisms [1], especially in multicellular eukaryotes [2]. Because gene duplication is the primary source of new genes [3], there is enduring interest in understanding the function of each duplicate gene [4,5]. However, early mouse studies that 'knocked out' duplicate genes revealed only mild or even no phenotypes [6,7], prompting the hypothesis that many mouse duplicates are functionally redundant and, therefore, that it would be difficult to discern the function of each copy by knocking out individual genes [8–10]. This view was reinforced when genome-wide gene deletion experiments showed that 12.4% of duplicates, compared with 29.0% of singletons, are essential to the viability or fertility of the yeast *Saccharomyces cerevisiae* [11] (Figure 1a). Similarly, in the nematode *Caenorhabditis elegans*, 2.3% of duplicates, but 7.6% of singletons, show lethal phenotypes in genome-wide knock-down experiments by RNA interference (RNAi) [12] (Figure 1a). However, the presumption that removing a mouse duplicate gene generates milder phenotypes than removing a singleton gene was based on anecdotal evidence and has not been systematically verified. Because of the expense and effort required to generate knockout mice and the potential value of such studies in understanding and treating human diseases, this verification is important because it could substantially affect the design and interpretation of mouse knockout experiments.

Proportion of mouse essential genes

We examined the presumption that mouse duplicates are functionally redundant using a list of 3872 genes that have been individually knocked out from the mouse genome. Because there are numerous different mutant phenotypes and it is not easy to compare their severities, we separated all phenotypes into two categories based on the phenotype annotation by Mouse Genome Informatics (MGI 3.51; http://www.informatics.jax.org). If the deletion of a gene leads to either lethality before reproduction or sterility (i.e. fitness reduces to 0), the gene is referred to as essential (see Methods in supplementary material online). All other genes are considered as nonessential, because they are not essential to viability or fertility. With this classification, our dataset includes 2136 essential and 1736 nonessential genes. We also classified the 3872 genes into 3087 duplicate genes (Table S1 in supplementary material online), which have at least one duplicate in the genome, and 785 singleton genes (Table S2 in supplementary material online). Unexpectedly, we found that the proportion of essential genes (P_E) is not significantly different between duplicate genes (55.1%) and singleton genes (55.4%) ($P = 0.89$; χ^2 test; Figure 1a). Use of different criteria to define duplicate and singleton genes does not change this result qualitatively (Table S3 in supplementary material online). There is also no difference in P_E among genes belonging to small gene families and those belonging to large families (Figure 1b). These results differ from yeast and nematode genomic studies (Figure 1a) and contradict the widely held view that mouse duplicate genes are functionally redundant.

Potential data biases

Protein sequence divergence

Because most of the mouse gene knockouts were generated by individual laboratories for different purposes, rather than by a genome-wide systematic effort, it is important to consider potential biases in the data that might have led to

Corresponding author: Zhang, J. (jianzhi@umich.edu).
Available online 7 June 2007.

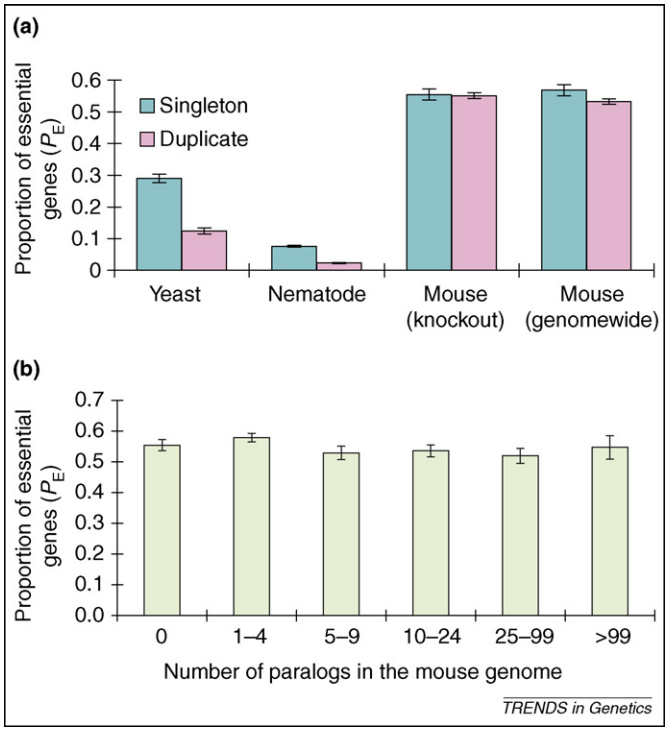

Figure 1. Proportion of essential genes (P_E) among singletons and duplicates. **(a)** Contrary to observations from yeast (*Saccharomyces cerevisiae*) and nematode (*Caenorhabditis elegans*), mouse (*Mus musculus*) duplicates and singletons are equally essential. Error bar indicates one standard error. The yeast results, based on gene deletion experiments, are from ref. [11], whereas the nematode results, based on RNAi knockdown experiments, are from ref. [12]. The mouse results labeled 'knockout' are based on the gene knockout data, whereas those labeled 'genomewide' are predicted from the knockout data after correction for potential data biases (see main text). **(b)** Essentiality of a mouse gene does not depend on how many paralogs it has in the genome (0, singletons; ≥1, duplicates). Error bar indicates one standard error.

overestimation of the P_E for duplicates. In both yeast [11] and nematode [12], P_E increases as the protein sequence divergence between the knocked-out gene and its closest paralog increases, presumably because functional redundancy between duplicate genes decreases as the protein sequence divergence increases [11,12]. If this relationship also exists in mouse, and if mouse geneticists have been avoiding targeting genes with close paralogs, our sample of duplicates might have been biased for a greater P_E compared with the P_E of randomly chosen duplicate genes in the genome. To examine this possibility, we measured the protein sequence divergence between a gene and its closest paralog by the proportion of amino acid positions that differ between the two proteins (*p*-distance) (see Methods in supplementary material online). We plotted the frequency distribution of the *p*-distance for the duplicate genes in the knockout gene dataset and for all duplicates in the mouse genome. Indeed, duplicates in the knockout dataset are clearly underrepresented in low *p*-distance ranges, compared with all duplicates in the genome ($P < 10^{-168}$, χ^2 test; Figure 2a). However, contrary to the observations in yeast [11] and nematode [12], mouse duplicates show a weak negative correlation between *p*-distance and P_E (Spearman rank correlation coefficient $\rho = -0.063$, $P < 10^{-3}$; see Figure 2a for binned results). Thus, the underrepresentation of knockout duplicates with low *p*-distances does not lead to an overestimation of P_E for duplicates.

Expression divergence

It is possible that duplicate genes with similar expression patterns have also been disfavored as knockout targets, because they might be more likely than those with dissimilar expression patterns to compensate for each other. We used $D = 1 - r$ to measure the temporal and spatial expression-profile divergence between a duplicate gene and its closest paralog, where r is the Pearson correlation coefficient between microarray expression signals of the two genes across 61 mouse tissues at different developmental stages (see Methods in supplementary material online). However, we did not observe a significant correlation between D and P_E ($\rho = -0.0031$, $P > 0.5$). Furthermore, there was no difference in the distribution of D between the knockout data and the entire genome ($P > 0.5$, χ^2 test; see Figure 2b for binned results). Thus, there is no need to correct the P_E estimate for duplicates here.

Evolutionary conservation

It is also possible that genes chosen for 'knockout' experiments tend to be evolutionarily more conserved than other genes in the genome, because conserved genes might be involved in fundamental cellular processes that are of particular interest to biologists. A recent mouse study showed that the rate of sequence evolution is on average 25% lower in essential genes than in nonessential genes [13]. Thus, if genes targeted for knockout differ from the genomic average in evolutionary rate, a bias in P_E might result. We measured evolutionary conservation of each mouse gene by computing the ratio of the number of nonsynonymous substitutions per site (d_N) to the number of synonymous substitutions per site (d_S) between the gene and its rat ortholog (see Methods in supplementary material online). A low d_N/d_S indicates that a gene is evolutionarily conserved. Mouse genes without one-to-one rat orthologs were not considered. As expected, P_E decreases as d_N/d_S increases, for both duplicates ($\rho = -0.17$, $P < 10^{-18}$; Figure 2c for binned results) and singletons ($\rho = -0.26$, $P < 10^{-11}$; Figure 2d for binned results). Interestingly, duplicate genes targeted for knockout have lower d_N/d_S values than the genomic average (Figure 2c), whereas the opposite is true for singletons (Figure 2d), suggesting that P_E has been overestimated for duplicates but underestimated for singletons. Using the P_E estimate in each d_N/d_S bin for duplicates (Figure 2c), we estimated that a randomly picked duplicate gene from the mouse genome has an expected P_E of 53.2% (Figure 1a). Similarly, we estimated that a randomly picked singleton gene from the mouse genome has an expected P_E of 56.8% (Figure 1a). These two numbers are not significantly different from each other ($P = 0.10$, χ^2 test).

Gene essentiality, functional compensation and genetic robustness

Estimating the proportion of essential genes (P_E) in a genome is not an easy task. For example, the nematode P_E values shown in Figure 1a are certainly underestimates because of frequent failure of RNAi to suppress gene expression completely [14], although the relative magnitudes of the P_E values for singletons and duplicates are probably correct

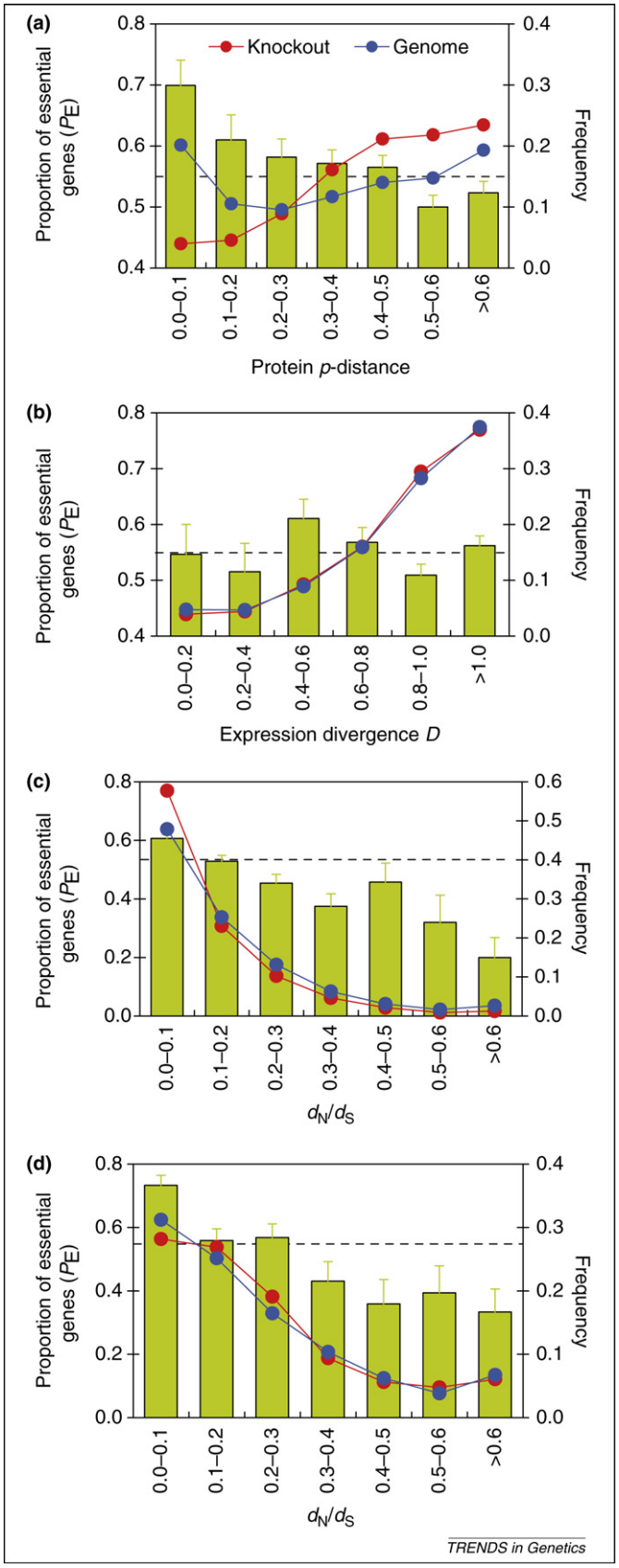

Figure 2. Factors that might influence the proportion of essential genes (P_E) in the mouse knockout data. **(a)** Relationship between P_E in duplicate genes and the protein *p*-distance between the target gene and its closest paralog in the genome. **(b)** Relationship between P_E in duplicate genes and the expression divergence between the target gene and its closest paralog. **(c)** Relationship between P_E in duplicate genes and the evolutionary conservation of the gene, measured by the ratio of the nonsynonymous (d_N) to synonymous (d_S) nucleotide distances between the target gene and its rat ortholog. **(d)** Relationship between P_E in singleton genes and the evolutionary conservation of the gene, measured by d_N/d_S between the target gene and its rat ortholog. In each panel, the green bars indicate P_E, corresponding to the left *y*-axis. Error bars show one standard error. The red and blue lines show the frequency distributions of genes from the knockout dataset and the entire genome, respectively, and the values are indicated on the right *y*-axis. The horizontal black dotted line shows the observed P_E among singleton knockout genes. None of these potential biases alter the relationship between P_E of duplicates and P_E of singletons.

[12]. In the case of mouse, a recent random *N*-ethyl-*N*-nitrosourea (ENU) mutagenesis targeting a 50 million-nt genomic region showed that 13.7% of mouse genes cause embryonic lethality when mutated [15]. The corresponding number from our knockout data is 14.0%. This comparison reconfirms the reliability of our estimate of mouse P_E. It is interesting that P_E seems to increase with organismal complexity, because it is ~7% in *Escherichia coli* [16], 17% in yeast [17,18] and 55% in mouse, when the organisms were examined under relatively benign laboratory conditions. Apparently, complex organisms devote a greater fraction of their genes to ensure viability and fertility than simple organisms do.

Our analyses of the phenotypic data from ~15% of genes in the mouse genome show that singletons and duplicates are equally likely to be essential to the organism. A similar result has been independently obtained by Liang and Li [19]. These findings strongly suggest that functional compensation between duplicates is rare in mouse, refuting the anecdote-based perception. In addition, there is no positive correlation of P_E and the protein sequence divergence between duplicates, providing further evidence for our conclusion. These findings differ from observations in yeast [11] and nematode [12] (Figure 1a). One possible reason is that the expression patterns of duplicates might diverge more rapidly in mouse than in yeast and nematode [20], because of the existence of many more cell types and developmental stages in mouse than in the other two species. Such expressional differences make functional compensation between duplicates unlikely even if the protein sequences of the duplicates are similar. However, we did not observe lower P_E for duplicates with lower expression divergence (Figure 2b), suggesting that the rapid expression divergence is probably not the main reason.

The second possible reason is that factors determining gene duplicability might be different in different species. It has been shown in yeasts that nonessential genes tend to duplicate, rendering P_E lower for duplicates than for singletons [21]. The absence of this mechanism in mouse [19] might explain our observation here. Regardless, our finding of low functional redundancy between mouse duplicate genes shows that there is no reason to avoid targeting duplicate genes in knockout experiments unless there are technical difficulties. Furthermore, the absence of expected phenotypes in mice deficient of a duplicate gene should not be automatically attributed to functional compensation from its paralogs.

Genetic robustness against null mutations seems to be a universal phenomenon of all living systems [22]. Such robustness can arise from the existence of duplicate genes with similar function and expression or from other reasons collectively known as 'distributed robustness' [22,23]. For

example, even in a metabolic network where no two enzymes catalyze the same reactions, some enzymes can be removed without rendering the network nonfunctional [22]. Our results indicate that duplicate genes have a negligible role in the genetic robustness of mouse. Furthermore, P_E of singleton genes is much greater in mouse than in yeast, indicating that the distributed robustness is also lower in mouse than in yeast.

Acknowledgements

We thank David Burke, Wen-Hsiung Li, members of the Zhang laboratory and three anonymous reviewers for valuable comments. This work was supported by a CCMB pilot grant from the University of Michigan and a research grant from the National Institutes of Health to J.Z.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.05.006.

References

1 Zhang, J. (2003) Evolution by gene duplication – an update. *Trends Ecol. Evol.* 18, 292–298
2 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
3 Ohno, S. (1970) *Evolution by Gene Duplication* Springer-Verlag
4 Greer, J.M. *et al.* (2000) Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* 403, 661–665
5 Zhang, J. (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat. Genet.* 38, 819–823
6 Joyner, A.L. *et al.* (1991) Subtle cerebellar phenotype in mice homozygous for a targeted deletion of the En-2 homeobox. *Science* 251, 1239–1243
7 Saga, Y. *et al.* (1992) Mice develop normally without tenascin. *Genes Dev.* 6, 1821–1831
8 Tautz, D. (1992) Redundancies, development and the flow of information. *Bioessays* 14, 263–266
9 Thomas, J.H. (1993) Thinking about genetic redundancy. *Trends Genet.* 9, 395–399
10 Cooke, J. *et al.* (1997) Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet.* 13, 360–364
11 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66
12 Conant, G.C. and Wagner, A. (2004) Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc Biol Sci* 271, 89–96
13 Liao, B.Y. *et al.* (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.* 23, 2072–2080
14 Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237
15 Wilson, L. *et al.* (2005) Random mutagenesis of proximal mouse chromosome 5 uncovers predominantly embryonic lethal mutations. *Genome Res.* 15, 1095–1105
16 Baba, T. *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection.. *Mol Syst Biol* 2, 2006.0008
17 He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet* 2, e88
18 Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906
19 Liang, H. and Li, W-H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23, 375–378
20 Makova, K.D. and Li, W.H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13, 1638–1645
21 He, X. and Zhang, J. (2006) Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.* 23, 144–151
22 Wagner, A. (2005) Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays* 27, 176–188
23 Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.* 24, 355–361

0168-9525/$ – see front matter.
doi:10.1016/j.tig.2007.05.006