Letters

Functional compensation by duplicated genes in mouse

Han Liang[1*] and Wen-Hsiung Li[1,2]

[1] Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA
[2] Biodiversity Research Center, Academia Sinica, Taipei, Taiwan 115

It has been thought that the functional loss of a gene because of null mutation can often be compensated by its paralog(s). Indeed, the genome-wide single-gene knockout or knockdown data in yeast and worm showed that the proportion of essential genes (P_E) in singletons is substantially greater than that in duplicates [1,2]. We consider a gene 'essential' if its deletion leads to lethality or sterility. However, the mouse knockout data [3] collected from individual experimental studies showed similar P_E values for singletons and duplicates [4,5]. This puzzling observation has attracted much attention [6,7]. Here, we propose an explanation.

Essentialities of singletons and duplicates in the mouse genome

Recently Makino *et al.* [6] found that developmental genes tend to be more essential than other genes and are highly enriched in the mouse knockout dataset. Here, we show that this enrichment does not cause a significant bias in the relative P_E values for singletons and duplicates at the genome level because the enrichment exists for both singletons and duplicates. From the dataset of Makino *et al.*, we calculate the P_E values for singletons and duplicates in the mouse genome after adjusting the bias of developmental genes (supplementary materials online). Interestingly, although both genome-wide P_E values become substantially lower than those in the knockout dataset (singletons from 42.2 to 35.6% and duplicates from 41.4 to 32.8%; Table S1), they are still similar ($p = 0.09$, $\chi^2 = 2.9$, Table S2). Next, we consider another bias in the knockout dataset [6], namely the enrichment of duplicate genes from whole-genome duplications. After adjusting for the functional bias, this factor has a less than 1% effect on the genome-wide P_E estimate (supplementary materials).

Higher network centrality of developmental duplicates than developmental singletons

Why is the P_E for mouse duplicates similar to that for singletons even at the genome level? One possible reason is the unequal functional partition between singletons and duplicates (i.e. a higher proportion of developmental genes and a lower proportion of unannotated genes in duplicates than in singletons). However, the observation [6] that among mouse developmental genes the P_E for duplicates is even higher than that for singletons further suggests there are other confounding factors.

From a systems biology perspective, the centrality of a gene in a biological network can affect gene essentiality [8]. Previously, we found that mouse duplicates tend to have more interacting partners in the protein interaction network and that genes encoding hub proteins are more likely to be essential [4]. Thus, is the higher essentiality of developmental duplicates due, in part, to their higher centrality in the network? We used the high quality protein interaction dataset [9] from a systematic examination of all binary interactions among ~7200 human proteins; this dataset is less biased than those collected from individual studies. For mouse developmental genes with phenotypic data, we used their human orthologs and the human protein interaction data to calculate the connectivity and betweenness of mouse proteins, which are two frequently used indexes for quantifying the centrality of a node in a network. Indeed, we found that developmental duplicates have higher network centrality than developmental singletons; the same trend also holds for all the duplicates and singletons in the dataset (Table 1). We obtained similar results using other protein interaction datasets (data not shown).

Unbiased estimation of functional compensation by duplicates

Thus, both functionality and network centrality can influence P_E estimation. To control these two factors, we compared the P_E values for the sets of singletons and duplicates that have the same functionalities and connectivities. In total, there are 1847 mouse genes with both interaction and knockout phenotypic data, and we classified them into three centrality groups: low-, median- and high-connectivity (centrality refers to connectivity here and in the remaining text). We calculated the P_E for each group of duplicates with the same functional classification and centrality classification. To obtain the averaged P_E of mouse duplicates in the dataset, we weighted them according to the proportions of their corresponding groups in singletons (supplementary materials).

Although the P_E values for singletons and duplicates are similar in the original dataset (45.7% vs. 42.4%, $p = 0.22$, $\chi^2 = 1.5$), after controlling for both functionality and centrality biases the adjusted P_E for duplicates is 39.0% (Figure 1), ~7% lower than that for singletons ($p = 0.01$, $\chi^2 = 6.4$, Table S6). This number implies that ~15% of the single-gene deletions that otherwise would be lethal (or infertile) are viable (or fertile) owing to duplicate functional compensation. Thus, the contribution of functional compensation by duplicates seems significant. Su and Gu [7] recently showed that young duplicates

Corresponding author: Li, W.-H. (whli@uchicago.edu)..
* Present address: Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

Table 1. Network centralities of singletons and duplicates in the mouse knockout dataset[a]

	Developmental singletons mean (± SD)	Developmental duplicates mean (± SD)	*P value* (Wilcox test)
Connectivity	3.06 (± 0.45)	4.90 (± 0.40)	0.013
Betweenness	3724 (± 872)	7897 (± 1019)	0.006
	All singletons mean (± SD)	**All duplicates mean (± SD)**	***P value*** **(Wilcox test)**
Connectivity	2.42 (± 0.22)	3.50 (± 0.20)	8×10^{-5}
Betweenness	2817 (± 416)	5206 (± 552)	1×10^{-5}

[a]The 'connectivity' of a protein is defined as the number of interacting partners of the node. The 'betweenness' of a given node is defined as the number of 'times' that a node in the network needs to go through the given node to reach another node by the shortest path.

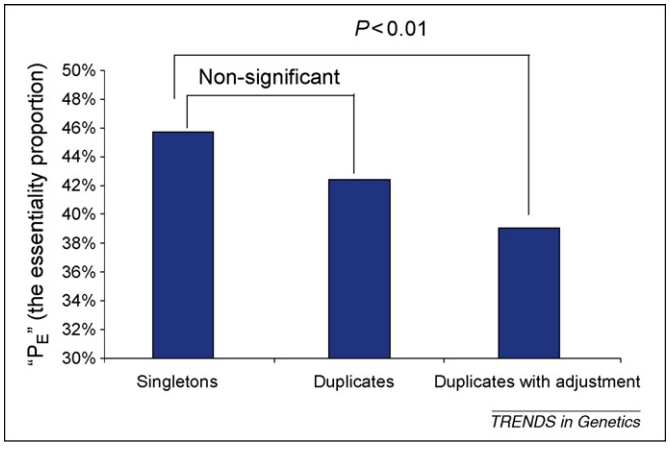

Figure 1. Proportions of essential genes (P_E) in duplicates with and without adjusting for confounding factors.

are under-represented in the knockout dataset. Because the backup role of young duplicates is likely to be stronger than old duplicates, the functional compensation by duplicates for the genome is likely to be higher than we estimated.

Concluding remarks

We have provided a system-level explanation for the observation that developmental genes are more essential [6]. Our results highlight the importance of controlling confounding factors in studying the role of duplicates in genetic robustness. Conventionally, the contribution to functional compensation by duplicates is inferred by directly comparing the P_E value for duplicates with that for singletons, and similar P_E values are usually taken as evidence that there is no contribution to compensation from duplicate genes. However, the functional partitioning and network centrality for duplicates might be different from those for singletons. It should be emphasized that even when genome-wide phenotypic data of single-gene deletion are available, correcting for such intrinsic differences remains necessary.

Our analyses only represent an initial effort to adjust confounding factors. Because current protein interaction datasets are incomplete, an estimation of functional compensation by duplicates in the whole mouse genome remains unfeasible. Moreover, other potentially confounding factors remain to be explored. Nevertheless, our study provides a general framework for estimating the contribution of duplicate genes to functional compensation by integrating functional genomic data.

Acknowledgements

We thank Dr Aoife McLysaght for providing us with her dataset and for her valuable comments on our manuscript. This study was supported by NIH grants (GM30998 and GM081724) to W.H.L.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.08.001.

References

1 Conant, G.C. and Wagner, A. (2004) Duplicate genes and robustness to transient gene knockdowns in *Caenorhabditis elegans*. *Proc. Biol. Sci.* 271, 89–96
2 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66
3 Eppig, J.T. *et al.* (2005) The mouse genome database (MGD): from genes to mice – a community resource for mouse biology. *Nucleic Acids Res.* 33, D471–475
4 Liang, H. and Li, W.H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23, 375–378
5 Liao, B.Y. and Zhang, J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.* 23, 378–381
6 Makino, T. *et al.* (2009) The complex relationship of gene duplication and essentiality. *Trends Genet.* 25, 152–155
7 Su, Z. and Gu, X. (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J. Mol. Evol.* 67, 705–709
8 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42
9 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178

0168-9525/$ – see front matter
doi:10.1016/j.tig.2009.08.001 Available online 25 September 2009