Published in final edited form as:

Trends Genet. 2009 October ; 25(10): 441–442. doi:10.1016/j.tig.2009.08.001.

Functional compensation by duplicated genes in mouse

Han Liang^{1,#} and Wen-Hsiung Li^{1,2,*}

¹Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

²Biodiversity Research Center, Academia Sinica, Taipei, Taiwan 115

It has been thought that functional loss of a gene due to 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 to be ‘essential’ if its deletion leads to lethality or sterility. However, the mouse knockout data³ collected from individual experimental studies showed similar P_E values for singletons and for 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 for duplicates in the mouse genome after adjusting the bias of developmental genes (Supplemental 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 dataset⁶, 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 (Supplemental 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 unequal functional partition between singletons and duplicates (i.e. a higher proportion of developmental genes and a lower proportion of un-annotated genes in duplicates than in singletons). However, the observation⁶ that among mouse developmental genes, the P_E for duplicates is even higher than that for singletons further suggests that there were other confounding factors.

© 2009 Elsevier Ltd. All rights reserved.

*Corresponding author: W.-H. Li (whli@uchicago.edu).

#Present address: Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

From the perspective of systems biology, the centrality of a gene in a biological network can affect gene essentiality⁸. 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 essential⁴. Thus, is the higher essentiality of developmental duplicates due in part to their higher centrality in the network? We use the high-quality protein interaction data set⁹ from a systematic examination of all binary interactions among ~7,200 human proteins; this dataset is less biased than those collected from individual studies. For mouse developmental genes with phenotypic data, we use 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 find that developmental duplicates have higher network centrality than developmental singletons; the same trend also holds for all the duplicates and the singletons in the dataset (Table 1). We obtain 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 compare the P_E values for the sets of singletons and duplicates that have the same functionalities and the same connectivities. In total, there are 1,847 mouse genes with both interaction and knockout phenotypic data, and we classify them into three centrality groups: low-, median- and high-connectivity (centrality refers to connectivity here and in the remaining text). We calculate 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 give their weights according to the proportions of their corresponding groups in singletons (Supplemental Materials).

Although in the original dataset the P_E values for singletons and duplicates are similar (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% (Fig. 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) due to duplicate functional compensation. Thus, the contribution of functional compensation by duplicates seems significant. Su and Gu⁷ recently showed that young duplicates 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 above.

Concluding remarks

We have provided a systems-level explanation for the observation that developmental genes are more essential⁶. 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 due to 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 the genome-wide phenotypic data of single-gene deletion is available, correcting for such intrinsic differences is still necessary.

Of course, 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 is still not feasible. Moreover, other potential 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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

References

1. Conant GC, Wagner A. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc Biol Sci* 2004;271:89–96. [PubMed: 15002776]
2. Gu Z, et al. Role of duplicate genes in genetic robustness against null mutations. *Nature* 2003;421:63–66. [PubMed: 12511954]
3. Eppig JT, et al. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* 2005;33:D471–D475. [PubMed: 15608240]
4. Liang H, Li WH. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet* 2007;23:375–378. [PubMed: 17512629]
5. Liao BY, Zhang J. Mouse duplicate genes are as essential as singletons. *Trends Genet* 2007;23:378–381. [PubMed: 17559966]
6. Makino T, et al. The complex relationship of gene duplication and essentiality. *Trends Genet* 2009;25:152–155. [PubMed: 19285746]
7. Su Z, Gu X. Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J Mol Evol* 2008;67:705–709. [PubMed: 19005716]
8. Jeong H, et al. Lethality and centrality in protein networks. *Nature* 2001;411:41–42. [PubMed: 11333967]
9. Rual JF, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437:1173–1178. [PubMed: 16189514]

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

Table 1Network centralities of singletons and duplicates in the mouse knockout dataset^a

	Developmental singletons mean ($\pm$ s.d.)	Developmental duplicates mean ($\pm$ s.d.)	<i>P</i> -value (Wilcoxon test)
Connectivity	3.06 ($\pm$ 0.45)	4.90 ($\pm$ 0.40)	0.013
Betweenness	3724 ($\pm$ 872)	7897 ($\pm$ 1019)	0.006
	All singletons mean ($\pm$ s.d.)	All duplicates mean ($\pm$ s.d.)	<i>P</i> -value (Wilcoxon test)
Connectivity	2.42 ($\pm$ 0.22)	3.50 ($\pm$ 0.20)	8×10^{-5}
Betweenness	2817 ($\pm$ 416)	5206 ($\pm$ 552)	1×10^{-5}

^aThe '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.