Formal Comment

Comments on "Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication"

Wei-Hua Chen[1]⁹, Xing-Ming Zhao[2]⁹, Vera van Noort[1¤], Peer Bork[1,3]*

[1] European Molecular Biology Laboratory (EMBL) Heidelberg, Heidelberg, Germany, [2] Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai, China, [3] Max-Delbrück-Centrum für Molekulare Medizin (MDC), Berlin, Germany

We previously showed that monogenic disease genes (MDs) are enriched in duplicates and hypothesized that functional redundancy among duplicates underlies this enrichment [1]. In their comment, Singh et al. refine this enrichment to genes resulting from whole genome duplications (WGDs) [2]; they, furthermore, "could not find any significant enrichment in duplicates in support of possible functional compensation for essential genes" [2] by using gene essentiality data from mouse (transferred to human through orthology).

We appreciate the scientific argument, but we would like to point out that confounding factors and data biases can lead to seemingly opposing conclusions. For example, we carefully considered the duplication age of genes, which is a known confounder in such analyses [3,4], as well as the use of gene subsets that have known biases such as the mouse essentiality data [3], which, in addition, have issues when conclusions are being transferred to human genes.

First, when using the data of Singh et al. [2] and stratifying small-scale duplicates (SSDs) into old and young groups according to the duplication age relative to WGD, we found that MDs are enriched in old SSDs; limiting this analysis to recessive MDs produced similar results (Figure 1A). In contrast, MDs are depleted in young SSDs (Figure 1B), which is consistent with our hypothesis and with our findings that coexpression decreases with increased duplication age. Thus, when the duplication is old, the ability of the functional copy to compensate for the mutation-carrying malfunctioning copy could be easily disrupted because of random fluctuation in gene expression in a subpopulation; consequently, the gene is associated with a disease, but it will not be purged from the whole population. Therefore, functional compensation can promote the spreading of disease genes in duplicates. However, in young duplicates, the fluctuation in gene expression among duplicates may not be that huge; thus, deleterious mutations could be tolerated, and the corresponding genes are unlikely to associate with any diseases.

Second, mouse essentiality data are biased [5], e.g., towards developmental genes; i.e., they do not correspond to the full spectrum of MDs. Dividing the tested mouse genes into subgroups, the proportion of essential genes in young SSDs is significantly lower than that of singletons (Figure 1C), consistent with functional redundancy among duplicates; however, the opposite is found in old SSDs (Figure 1D). The latter has led to the somewhat counterintuitive conclusion that "duplicates are as essential as singletons" [6], which has been argued against by several follow-up studies [3–5]. These results, again, highlight the importance of taking duplication age into consideration. As previous studies suggested, it is not trivial to correct the biases [3–5], and hence, conclusions from this data regarding duplications have to be taken with caution. Furthermore, the essentiality status of mouse genes cannot be reliably transferred to human and vice versa. For example, using data from OGEE [7], an online gene essentiality database, 2,322 mouse essential genes have one-to-one orthologs in human; only 476 out of the 2,322 human genes (approximately 20%) were essential according to a genome-wide small interfering RNA (siRNA) experiment [8].

Finally, only less than 30% of the MDs we collected [1] were used in the analyses by Singh et al.; the intersection with the essentiality dataset is even smaller (approximately 18.6% of the MDs used in [1]) because, so far, only less than one-third (approximately 6,400) of mouse genes has been tested for essentiality [9]. Thus, extrapolating any observations on these data to the whole genome would be difficult; for example, some functional signals might only become statistically significant in larger datasets.

Elucidating the molecular basis of human genetic disorders is one of the most important tasks in medical biology. With the relevant data, such as those from genome-wide association studies (GWAS), accumulated at an astonishing speed, integrative and comparative analyses through bioinformatics are much needed. In this regard, Singh et al. did provide an important contribution by refining the enrichment of dominant MDs in duplicates to those derived from WGD. However, we don't believe that they nullified our functional compensation hypothesis with the analyses performed, but they certainly encouraged further studies on more complete datasets, hopefully to be available in the near future.

Citation: Chen W-H, Zhao X-M, van Noort V, Bork P (2014) Comments on "Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication". PLoS Comput Biol 10(7): e1003758. doi:10.1371/journal.pcbi.1003758

Editor: Alon Keinan, Cornell University, United States of America

Published July 31, 2014

Funding: The authors received no specific funding for this work.

Competing Interests: This Formal Comment is a response to Singh et al., "Human Dominant Disease Genes are Enriched in Paralogs Originating from Whole Genome Duplication," by the authors of the original research article "Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs."

* Email: bork@embl.de

⁹ These authors contributed equally to this work.

¤ Current address: Department of Microbial and Molecular Systems (M2S), KU Leuven, Leuven, Belgium

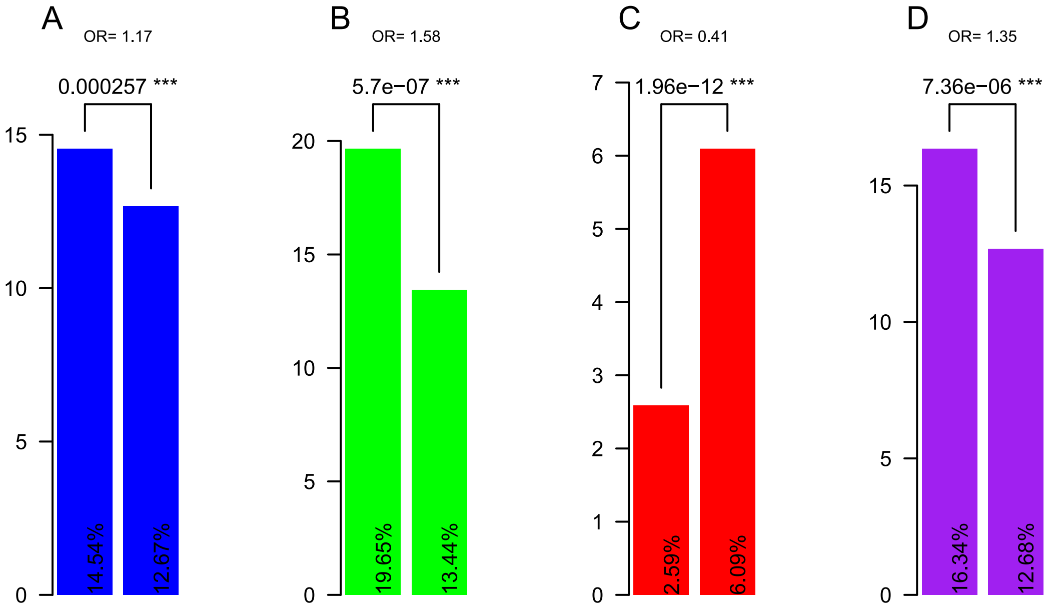

Figure 1. Enrichment of MDs in old SSDs and distinct characteristics of the old SSDs as compared with the young ones. Statistics using data from Singh et al. *P*-values and ORs (odd ratios) are calculated using Fisher's Exact Test (see Dataset S1 for the R code). A. MDs genes are enriched in old duplicates; left: percentage of old SSDs in MDs, right: percentage of old SSDs in all genes. B. Recessive MDs genes are enriched in old SSDs; left: percentage of old SSDs in recessive MDs, right: percentage of old SSDs in all MDs. C. Essential genes are depleted in young SSDs; left: percentage of young SSDs that are essential, right: percentage of young SSDs in tested genes. D. Essential genes are enriched in old SSDs; left: percentage of old SSDs that are essential, right: percentage of old SSDs in tested genes.
doi:10.1371/journal.pcbi.1003758.g001

Supporting Information

Dataset S1 Raw data and an R script used in this study are available in Dataset S1 as an archive file; readers can use these materials to reproduce our results, including the statistical tests and the figure.
(ZIP)

Author Contributions

Conceived and designed the experiments: PB. Performed the experiments: WHC XMZ. Analyzed the data: WHC XMZ. Wrote the paper: WHC XMZ PB VvN.

References

1. Chen W-H, Zhao X-M, van Noort V, Bork P (2013) Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. PLoS Comput Biol 9: e1003073.
2. Singh PP, Affeldt S, Malaguti G, Isambert H (2014) Human Dominant Disease Genes are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Comput Biol 10: e1003754.
3. Chen W-H, Trachana K, Lercher MJ, Bork P (2012) 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. Mol Biol Evol 29: 1703–1706.
4. 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.
5. Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. Trends Genet 25: 152–155.
6. Liao B-Y, Zhang J (2007) Mouse duplicate genes are as essential as singletons. Trends Genet 23: 378–381.
7. Chen WH, Minguez P, Lercher MJ, Bork P (2012) OGEE: an online gene essentiality database. Nucleic Acids Res 40: D901–906.
8. Silva JM, Marran K, Parker JS, Silva J, Golding M, et al. (2008) Profiling essential genes in human mammary cells by multiplex RNAi screening. Science 319: 617–620.
9. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT, et al. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res 39: D842–848.