11 Cui, Q. *et al.* (2006) Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.* 2, 46
12 Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297
13 Alvarez-Garcia, I. and Miska, E.A. (2005) MicroRNA functions in animal development and human disease. *Development* 132, 4653–4662
14 Tirosh, I. *et al.* (2006) A genetic signature of interspecies variations in gene expression. *Nat. Genet.* 38, 830–834
15 Krek, A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.* 37, 495–500
16 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
17 Grun, D. *et al.* (2005) microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS. Comput. Biol.* 1, e13
18 Wittkopp, P.J. *et al.* (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature* 430, 85–88
19 Chen, K. and Rajewsky, N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.* 38, 1452–1456
20 Hornstein, E. and Shomron, N. (2006) Canalization of development by microRNAs. *Nat. Genet.* 38 (Suppl.), S20–S24
21 Farh, K.K. *et al.* (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* 310, 1817–1821
22 Sood, P. *et al.* (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2746–2751
23 Li, Y. *et al.* (2006) MicroRNA-9a ensures the precise specification of sensory organ precursors in *Drosophila*. *Genes Dev.* 20, 2793–2805
24 Cohen, S.M. *et al.* (2006) Denoising feedback loops by thresholding – a new role for microRNAs. *Genes Dev.* 20, 2769–2772

0168-9525/$ – see front matter.

doi:10.1016/j.tig.2007.04.003

Gene essentiality, gene duplicability and protein connectivity in human and mouse

Han Liang and Wen-Hsiung Li

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

It has previously been found that, in yeast, gene essentiality is positively correlated with protein connectivity (number of interaction partners) but negatively correlated with the existence of gene duplicates and that highly connected proteins tend to have a low gene duplicability. Using data from human and mouse, we show here that, in mammals, the first of these relationships holds true, but unlike the second relationship in yeast, highly connected mammalian proteins tend to have a high gene duplicability, and there is no correlation between gene essentiality and gene duplication in mammals.

Introduction

There has been much interest in the relationships among gene function, phenotypic effect of gene deletion or knockout, and gene duplication at the genomic level [1–9]. For this purpose, three terms are often used: (i) protein connectivity, which is defined as the number of links that a protein node has to other nodes in the protein interaction network; (ii) gene essentiality, which is defined using words such as 'the deletion of a gene from the genome has a lethal effect or causes infertility' [10,11]; and (iii) gene duplicability, which describes the likelihood of a gene having one or more paralogs [8]. So far, however, most of our knowledge about the relationships among these three factors comes from yeast. In yeast, a protein that is highly 'connected' to other proteins (i.e. that interacts with many other proteins) tends to result in the death of the organism if it is deleted from the genome [3,12,13]. This is commonly known as the 'centrality–lethality' rule, which either reflects the crucial role of hub proteins (i.e. highly connected proteins) in the architecture of the network [3] or is simply because hub proteins have a higher probability of engaging in essential protein–protein interactions [14]. Furthermore, the proportion of essential (deletion-lethal) genes is significantly higher among singletons than among duplicates, and the deletion of a duplicate gene is, on average, less severe than the deletion of a singleton [2]. Recent studies indicated a negative correlation between protein connectivity and gene duplicability, which implies that genes with a higher protein connectivity tend to have fewer duplicate genes in the yeast genome [15]. Do these relationships hold true in such complex organisms as mammals?

Relationships among gene essentiality, gene duplicability and protein connectivity

First, the available mouse targeted knockout phenotypic annotations were extracted from the Mouse Genome Database (MGD; http://www.informatics.jax.org/) [16], and mouse genes and their orthologous human genes (annotated by MGD) were classified as essential or non-essential genes. Here, we defined an essential gene as a gene whose knockout phenotype is annotated as lethality (including embryonic, perinatal and postnatal lethality) or infertility [10,11]. Second, protein connectivity was calculated based on the human protein–protein interaction data (including both yeast two-hybrid and literature-curated interactions) from the study by Rual *et al.* [17]. Finally, gene family information was obtained (i.e. gene family IDs) in the human and mouse genomes, according to the annotation in the Ensembl Genome Database [18,19].

Corresponding author: Li, W.-H. (whli@uchicago.edu).
Available online 21 May 2007.

Gene essentiality versus protein connectivity

From the 1137 human genes for which protein interaction data from humans and phenotypic data from mice were available, we found that the proportion of essential genes is positively correlated with protein connectivity (Figure 1a). Moreover, in terms of protein connectivity, the distributions for essential and non-essential genes are significantly different (Wilcoxon rank test; $P = 5 \times 10^{-6}$). These results are consistent with the observation in yeast, suggesting the centrality–lethality rule [3] also holds true in mammals. Thus, highly connected proteins tend to be essential for survival or reproduction for both simple and complex organisms.

Protein connectivity versus gene duplicability

From the 5530 human genes for which both protein interaction data and gene family annotation were available, we found that gene duplicability, defined as $1 - F$ (where F is the proportion of unduplicated gene types), is positively correlated with protein connectivity (Figure 1b). Consistent with this, the number of paralogs per gene is positively correlated with protein connectivity (Spearman rank test; $R = 0.26$, $n = 5{,}530$, $P < 10^{-84}$). This trend is opposite to that observed in yeast.

Gene essentiality versus gene duplication

From the 2899 mouse genes for which both phenotypic data and gene family annotation were available, we found that the proportion of essential genes does not differ between singletons and duplicates (48.6% versus 46.2%; $\chi^2 = 1.3$, $P = 0.3$; see the supplementary material online). Moreover, there is no significant difference between essential and non-essential genes in terms of family size distribution (Wilcoxon rank test; $P = 0.1$).

In view of considerable noise in the datasets, the robustness of these results was further tested in two directions. First, the same analyses were performed using literature-curated and multivalidated interaction datasets separately. Second, to examine the effect of the definition of 'essential genes' that we use here, genes whose deletions have a lethal effect and genes whose deletions cause infertility were considered separately. In all of these analyses, we obtained the results obtained as those described earlier (see the supplementary material online). The potential biases and caveats in

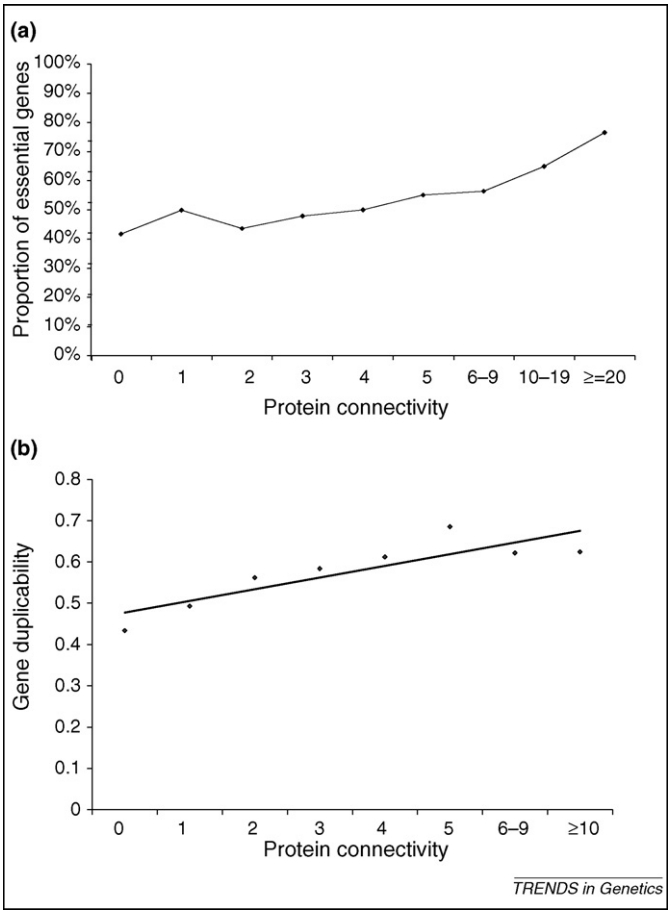

Figure 1. Relationships among gene essentiality, gene duplicability and protein connectivity in mammals. **(a)** A positive correlation between proportion of essential genes and protein connectivity in the human protein–protein interaction network. **(b)** A positive correlation between protein connectivity and gene duplicability in the human protein–protein interaction network. Gene duplicability is defined as $1 - F$, where F is the proportion of unduplicated gene types, and the number of gene types is defined as the number of singletons plus the number of gene families with more than one member. The trend between gene essentiality and protein connectivity holds the same as in yeast, whereas the correlation between protein connectivity and gene duplicability is the opposite of that found in yeast.

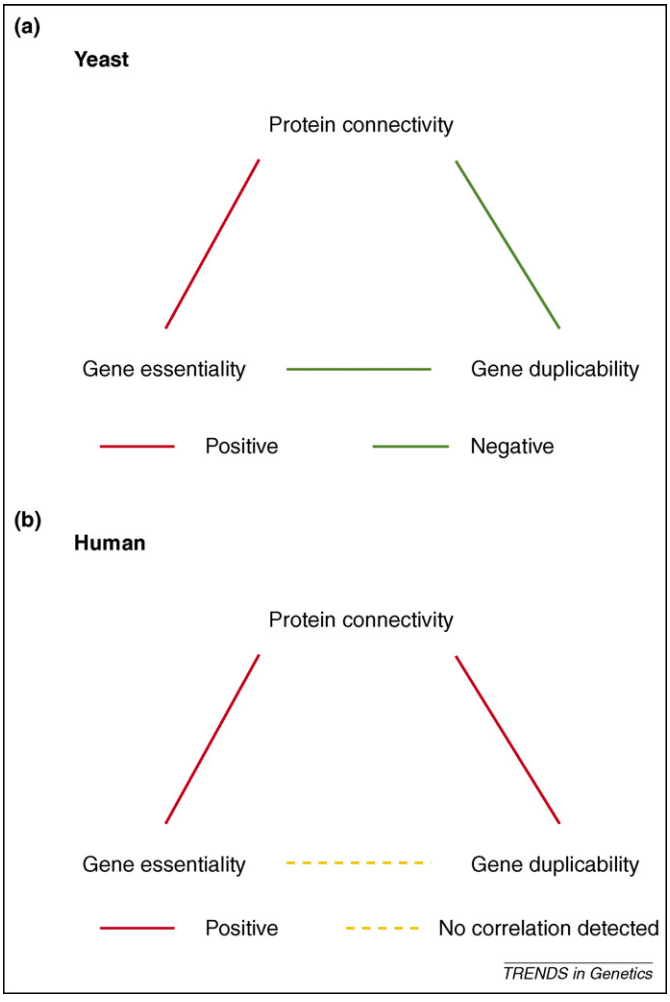

Figure 2. A comparison of the relationships among gene essentiality, gene duplicability and protein connectivity in **(a)** yeast and **(b)** human. There is a fundamental difference among these relationships between yeast and human.

these analyses are discussed in more detail in the supplementary material online. To highlight the differences between yeast and human, the relationships among gene essentiality, gene duplicability and protein connectivity can be demonstrated in the form of a triangle in which the pairwise correlations are represented by its three sides (Figure 2).

Why do highly connected proteins tend to have a higher gene duplicability in humans?

In yeast, an important factor for determining the retention of gene duplicates is whether the duplication causes a deleterious effect as a result of higher protein dosage, which is more sensitive for hub proteins than for non-hub proteins, leading to a negative correlation between protein connectivity and gene duplicability. By contrast, mammals are more robust against a dosage increase caused by gene duplication and have a greater variety of cell types, enabling duplicate genes to diversify in function [20,21]. These two factors have been suggested to account for the higher gene duplicability in mammals than in yeast [8,22], and they might also help to explain the observation that, in mammals, highly connected proteins tend to have a higher gene duplicability than do less connected proteins. We speculate that, in mammals, a highly connected protein might need to be produced in a high dosage, so that a duplicated hub protein might have a better chance of survival than a duplicated non-hub protein. More importantly, a high connectivity might confer a greater chance of functional diversification (e.g. tissue specialization) to duplicated genes in mammals. In comparison, selection for functional diversification in yeast might not be a major factor because of the simplicity of the organism (i.e. it is unicellular). This view is consistent with a recent study showing that only duplicates that arose through post-multicellularity duplication events have a tendency to become more specifically expressed, rather than duplicates that arose in a unicellular ancestor [23]. An alternative explanation for the opposite connectivity–duplicability patterns between yeast and humans is that yeast has undergone a relatively recent whole-genome duplication (in the last ~100 million years) [24], whereas mammals have not.

Why do gene essentiality and gene duplication seem to be uncorrelated in mammals?

The fitness effect of deleting a singleton gene reflects the intrinsic importance of that gene in the organism. For a duplicate gene, the single-deletion fitness effect is also influenced by the compensatory role of its paralog(s) in the genome [2]. In yeast, singleton genes tend to have more interaction partners, suggesting that they are intrinsically more essential for the organism. This is confirmed by He and Zhang *et al.* [25], who focused on *Saccharomyces cerevisiae* singleton genes and examined whether their orthologs have been duplicated in related yeast genomes. They found that the singletons that were duplicated in other yeast species have less severe deletion fitness effects than those that were not duplicated. Thus, both factors – the difference in intrinsic importance between singletons and duplicates and the compensatory role of duplicates – contribute to a less severe fitness effect of deleting a yeast duplicate gene, although the contributions from these two factors cannot be separated. In mammals, duplicate genes, on average, have a higher connectivity than do singletons, suggesting that duplicate genes are intrinsically more essential. Moreover, using a similar approach to study mouse singletons by examining the duplicability of their orthologs in the human genome, it was found that the trend was opposite to that seen in yeast (see supplementary material online). Thus, the potential compensatory role of gene duplication contributes to a less severe fitness effect of gene deletion in mammals; whereas the difference in intrinsic importance between singletons and duplicates might contribute to a more severe fitness effect of deletion in duplicate genes. These two factors might cancel each other out, leading to no detectable difference in gene essentiality between duplicate genes and singletons.

Acknowledgements

We thank Ay Prachumwat, Zhenglong Gu and Jian Lu for helpful discussion. This work was supported by NIH grants to W.H.L.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.04.005.

References

1 Ohno, S. (1970) *Evolution by Gene Duplication* Springer-Verlag
2 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66
3 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42
4 Long, M. and Thornton, K. (2001) Gene duplication and evolution. *Science* 293, 1551
5 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
6 Oltvai, Z.N. and Barabasi, A.L. (2002) Systems biology. Life's complexity pyramid. *Science* 298, 763–764
7 Ge, H. *et al.* (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* 19, 551–560
8 Yang, J. *et al.* (2003) Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15661–15665
9 Boulton, S.J. *et al.* (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* 295, 127–131
10 Graur, D. and Li, W.H. (2000) *Fundamentals of Molecular Evolution* Sinauer Press
11 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
12 Batada, N.N. *et al.* (2006) Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* 2, e88
13 Reguly, T. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* 5, 11
14 He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.* 2, e88
15 Prachumwat, A. and Li, W.H. (2006) Protein function, connectivity, and duplicability in yeast. *Mol. Biol. Evol.* 23, 30–39
16 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–D475
17 Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178
18 Birney, E. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.* 34, D556–D561

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.