Personalized Monograph

2023v1

Evolution of Genetic Robustness

Xun Gu

Table of Contents

Unit-1

Two Mechanisms of Genetic Robustness: Duplicate Compensation versus Genetic Robustness

Molecular biologists know that mutations without a phenotype are not exceptional. Yet in many 'wet' laboratories, natural or laboratory-generated null mutations are still used routinely to explore the function of individual genes. The wisdom of this approach is challenged in the post-genomics era because complex networks ranging from biological systems to the Internet show extraordinary robustness against random perturbations (e.g. deleterious mutations [1]). Meanwhile, this observation is greatly increasing interest in the emergence of gene network robustness. Currently, two possible mechanisms are proposed: (1) 'genetic buffering' from redundant gene networks (i.e. alternative metabolic or regulatory/signal pathways), and (2) functional complementation from duplicate genes 2, 3. Despite the fact that the relative importance of the two mechanisms in genetic robustness is still a matter of debate, recent genome-wide studies have provided a tremendous amount of information that could shed some light on this controversy 4, 5, 6, 7, 8, 9.

1. Genetic buffering or functional complementation?

Wagner [4] studied 45 yeast duplicate genes to explore the relationship between sequence evolution and the fitness of yeast when a single gene is deleted. The fitness (f) is measured by the growth rate relative to the wild type, ranging from normal (f=1) to lethal (f=0). The premise is that if duplicate genes are functionally compensated, there would be a positive correlation between sequence similarity of duplicate genes and the fitness; for example, a duplicate gene pair with 99% sequence identity is expected to have $f \approx 1$ (i.e. causes a normal phenotype) when either gene is deleted. Although there was a correlation between sequence and fitness, it had no statistical significance, and Wagner [4] has virtually dismissed the role of duplicates on the genetic robustness.

With the nearly complete dataset of fitness effects for yeast mutant strains with single genes deleted, new studies 6, 7 show that Wagner's inference [4] might not be correct, probably because he used a limited dataset. Indeed, Gu *et al.* [7] have provided several lines of evidence to show the significant role of duplicate genes on genetic robustness. They found a significantly higher probability of functional compensation for a duplicate gene than for a single-copy gene, a high correlation between the frequency of compensation and the sequence similarity of two duplicates, and a higher probability of having a severe fitness effect when the duplicate copy with a higher expression level is deleted. Overall, it has been estimated that in *Saccharomyces cerevisiae* at least a quarter of gene deletions that have no phenotype are compensated for by duplicate genes [7].

The effect of functional complementation by duplicate genes has also been observed recently in a systematic analysis of *Caenorhabditis elegans* genome using double-stranded RNA interference (RNAi) [9]. In this study, Kamath *et al.* [9] screened the loss-of-function RNAi phenotypes for ~86% of predicted genes of *C. elegans*, and identified 1722 genes (~10% of all genes) that have nonviable/lethal, growth defect or post-embryonic phenotypes. They observed that *C. elegans* genes with an orthologue in another eukaryote (i.e. the genes are conserved and therefore supposed to have essential function) are much more likely (~3.5 fold) to have a detectable RNAi phenotype than all other genes; Furthermore, of these conserved genes, genes that have only a single-copy in *C. elegans* are more likely (~2.6 fold) to have an RNAi phenotype than those that have at least one duplicate.

The role of duplicates in genetic robustness is supported by the pattern of RNAi phenotype clustering in *C. elegans* chromosomes [9]. The five autosomes of *C. elegans* have a central 'cluster' with low rates of recombination, which is flanked by chromosome 'arms' with 10-fold high recombination rates. Kamath *et al.* [9] discovered that genes with RNAi phenotypes are enriched twofold in the cluster regions relative to the arms. Because of the increased gene duplications in arm regions (thanks to the high rate of recombination), it is apparently just another observation of the same effect, namely deletion of a gene with a duplicate has less effect than deletion of a singleton.

2. Future experiments

These whole-genome approaches that targeting single genes in yeast [7] are only the first-order approaches to investigating the pattern of genetic robustness. The limitation is that genetic interactions cannot be measured directly from these datasets. One (naıve) solution is to generate whole-genome libraries of yeast multi-gene deletion strains. The obstacle is the magnitude of experiments. For yeast genome with ~6000 genes, ~18 million two-gene-deletion strains are required to cover all possibilities! Obviously, a blind data-driven approach is no longer efficient, and a hypothesis-driven experimental design should be used. Two speculations are follows.

(1) Let q be the probability of a single-copy gene having no phenotype when it is deleted. Thus, under the assumption of independence, the probability of no phenotype after k genes are deleted is given by $P(k)=q^k$. The pattern of genetic buffering against null mutations can be investigated by the semi-log plot of $P(k)$ against k (Fig. 1); a similar study has been reported for non-biological systems [10]. To obtain the $P(k)$:k curve experimentally, one actually only needs to select randomly N sets of single-copy genes for each k, say, in a range of 500–1000.

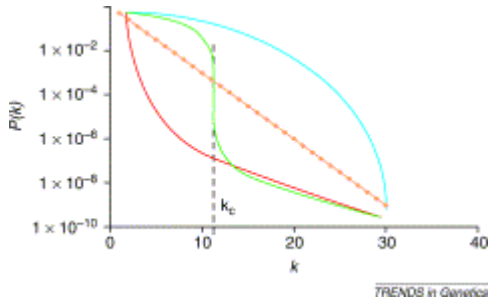

Fig. 1. The hypothetical semi-log plotting for $P(k)$ against k; where $P(k)$ is the probability of having phenotype when a random k genes are deleted from the genome. The orange curve is expected by an independent model that the pattern of robustness is determined by individual gene effect (i.e. $P(k)=q^k$). The green curve shows the case where the genetic robustness remains until k_c genes are deleted; where k_c is the critical value at which robustness collapses. The blue curve is the case showing positive gene interactions for robustness but no critical change. And the red curve is for the case that when more genes are deleted, the genetic robustness collapses faster than predicted by the random model, which seems unlikely.

(2) A detailed characterization of functional compensation of duplicates can be obtained from a complete set of gene deletions of the gene family [11]. The total combination number of a gene family with n genes is 2^n, which is feasible when n is not very large. If the phylogenetic tree of the gene family is known, one could use a phylogeny-based partition: member genes can be partitioned into $2n-3$ various two-group sets along the tree (Fig. 2). For each set, two complementary multi-deletion strains are designed so that only $2(2n-3)$ deletion strains are required for a gene family.

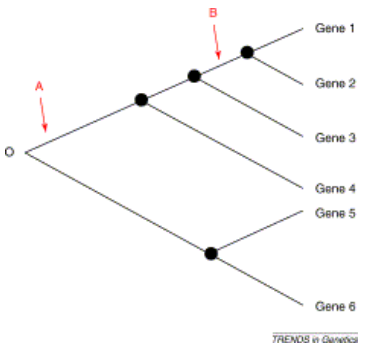

Fig. 2. Phylogeny-based design for loss-of-function phenotypes of a gene family: For partition A, the two gene groups are (1,2,3,4) and (5,6), respectively. Then, the gene deletion patterns are D1D2D3D4, and D5D6, respectively. For partition B, the two gene groups are (1,2) and (3,4,5,6), respectively. For an n-gene phylogeny, there are $2n$-3 different partitions

3. Conclusions

In summary, genome-wide studies 6, 7, 8, 9 have indicated that duplicate genes and genetic buffering are both important in genetic robustness. Because neither of them is related directly to the functional constraints of

genes, no strong correlation is expected between the sequence conservation and the fitness effect when the gene is deleted. However, the fundamental problem, that is, the evolutionary mechanism underlying the emergence of genetic robustness, remains largely unsolved. The following speculations might be worth further study.

(1) *A certain amount of genetic buffering originally comes from gene duplication*. Functional compensation by duplicates might be a by-product of functional divergence after gene duplication. Their recruitment into novel gene network (or other types of functional divergence) sets up a boundary to stop further divergence. For some sub-functional components, such a 'frozen' process is actually the part of functional divergence after gene duplication. The overlap in function due to the 'frozen' process could still exist (e.g. at the protein structure level), even if the sequence similarity is too little to be detected.

(2) *Genetic robustness is selectively nearly-neutral*. Genetic robustness effectively removes lethal mutants from the population so the risk of extinction can be reduced. It is possible that the capability of genetic buffering could be lost when a buffered mutant spreads over the population (fixation) by genetic drifts. This can be illustrated by the case of two alternative pathways that are mutually functionally compensated. Therefore, if one of pathways is inactive, the individual is still 'normal', but the function is then no longer robust against any further null mutation. Nevertheless, an individual carrying a buffered mutant might have a subtle cost in fitness, for example with a coefficient of coefficient (s) as small as 0.01 [15], which means that the fitness of this individual is relatively 1% less than that of the wild type. Obviously, such tiny difference in fitness is not distinguishable under the laboratory conditions, but during the course of evolution, the chance of fixation is very small for a buffered mutant when the effective population size (N_e) is above 100, or $N_e s>1$. In other words, the capability of genetic robustness can be maintained by the stabilizing selection.

(3) *Genetic robustness maintained by continuous gene duplication events.* According to some theoretical models 10, 16, the emergence of genetic buffering against null mutations requires a continuous input of new genes during the course of evolution. Therefore, small- and large-scale gene (domain) duplications, being major mechanisms for the origin of new genes 17, 18, are a prerequisite for the emergence of genetic robustness. Indeed, many examples have shown that functional divergence among duplicates has increased the complexity of molecular pathways [19], supported by a recent estimate that 98% of the human proteome evolved by domain duplication [20]. Of course, these views remain to be validated by further research.

Further-reading

Gu, X*. (2003) Genetic buffering, gene duplication, and evolution. Trends in Genetics. 19:354-356.

References

1 Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. Science 296, 910–913
2 Gibson, T.J. and Spring, J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. 14, 46–49
3 Nowak, M. et al. (1997) Evolution of genetic redundancy. Nature 388, 167–171
4 Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. Nat. Genet. 24, 355–361
5 Kitami, T. and Nadeau, J.H. (2002) Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. Nat. Genet. 32, 191–194
6 Winzeler, E.A. et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901–906
7 Gu, Z. et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421, 63–66
8 Giaever, G. et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418, 387–391
9 Kamath, R.S. et al. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421, 231–236
10 Albert, R. et al. (2000) Error and attack tolerance of complex networks. Nature 406, 378–382
11 Beh, C. et al. (2001) Overlapping functions of the yeast oxysterol-binding protein homologues. Genetics 157, 1117–1140
12 Hirsh, A.E. and Fraser, H.B. (2001) Gene dispensability and rate of evolution. Nature 411, 1046–1049
13 Papp, B. et al. (2003) Gene dispensability does not determine the rate of evolution. Nature 421, 496–497
14 Nei, M. (1969) Gene duplication and nucleotide substitution in evolution. Nature 221, 40–42
15 Tautz, D. (2000) A genetic uncertainty problem. Trends Genet. 16, 475–477
16 Baraba´ si, A. and Albert, R. (1999) Emergence of scaling in random networks. Science 286, 509–512
17 Gu, X. et al. (2002) Age-distribution of human gene families showing equal roles of large and small-scale duplications in vertebrate evolution. Nat. Genet. 31, 205–209
18 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155
19 Gerhart, J. and Kirschner, M. (1997) Cells, Embryos, and Evolution,
Blackwell Science
20 Muller, A. et al. (2002) Structural characterization of the human proteome. Genome Res. 12, 1625–1641

Unit-2

Role of Yeast Duplicate Genes in Genetic Robustness against Single-Gene Deletions

Introduction

Deleting a gene in an organism often has little phenotypic effect[1–5], owing to two mechanisms of compensation[4–10]. The first is the existence of duplicate genes: that is, the loss of function in one copy can be compensated by the other copy or copies. The second mechanism of compensation stems from alternative metabolic pathways, regulatory networks, and so on. The availability of fitness data for a nearly complete set of single-gene-deletion mutants of the *Saccharomyces cerevisiae* genome[11] has enabled us to carry out a genome-wide evaluation of the role of duplicate genes in genetic robustness against null mutations.

Fitness measurements

Fitness measurements were obtained from a high-throughput study[11] that measured the growth of each strain of a nearly complete collection of yeast single-gene-deletion mutants under both fermentable and non-fermentable (respiratory) growth conditions. We studied five growth media: YPD (1% Bacto-peptone (Difco), 2% yeast extract and 2% glucose), YPDGE (0.1% glucose, 3% glycerol and 2% ethanol), YPE (2% ethanol), YPG (3% glycerol) and YPL (2% lactate). Each strain contained the precise homozygous diploid deletion of 1 of 4,706 ORFs in the yeast genome. We calculated the fitness values for each media condition as the extent of survival and reproduction (fitness) of the deletion strain relative to the pool of all strains grown and measured collectively. Fitness values (f) of 1.0 indicate no difference between a single strain and the pool average for that condition; $f<1$ indicates that the strain is less fit, and $f>1$ indicates that the strain is more fit than the pool average. In addition, we added to our analysis 1,060 ORFs that each had a lethal effect when deleted and assayed in YPD; we used only lethal deletions that could be inferred as lethal from both of the two studies conducted[11,19]. We divided all genes into four groups according to the f value as follows: (1) if $f>0.95$ for all five media conditions[11], the deletion has a weak or no fitness effect in all conditions; (2) if $0.8<f_{min}<0.95$, where f_{min} is the smallest f value for all five growth conditions, the deletion has a moderate effect; (3) if $0<f_{min}<0.8$, the deletion has a strong effect; and (4) if the deletion is lethal, we set $f=0$.

Higher probability of functional compensation for a duplicate gene

From 5,766 yeast open reading frames (ORFs) for which we had a fitness measure of strains with a corresponding single-gene deletion[11], we excluded functionally unknown genes from further study. This yielded 1,275 singleton genes, and 1,147 duplicate genes that had at least one paralogue elsewhere in the genome. We compared the frequency distribution of fitness for duplicate genes with that for singletons (Fig. 1a). We classified genes into four groups on the basis of the fitness value for a strain across the five different growth conditions tested (Methods). The two distributions were significantly different ($P<0.001$): duplicate genes had a significantly lower proportion of genes with a lethal effect of deletion (12.4% versus 29.0%) and a significantly higher proportion of genes with a weak or no effect of gene deletion (64.3% versus 39.5%). This comparison indicates that there is a significantly higher probability of functional compensation for a duplicate gene than for a singleton. We emphasize that 'compensation' here does not imply that the gene is dispensable in long-term evolution but means that the gene is dispensable in an individual under the conditions tested[6].

To see whether the above conclusion held for different growth environments, we compared the distributions of fitness (f) for duplicate and singleton genes under each of the five growth conditions studied[11]. The empirical cumulative distributions of f under the YPD growth condition (Methods) for duplicate genes and for singletons are shown in Fig. 1b. The Kolmogrov–Smirnov test indicated that the two distributions were significantly different ($P<10^{-10}$). The same conclusion held for the other four growth conditions.

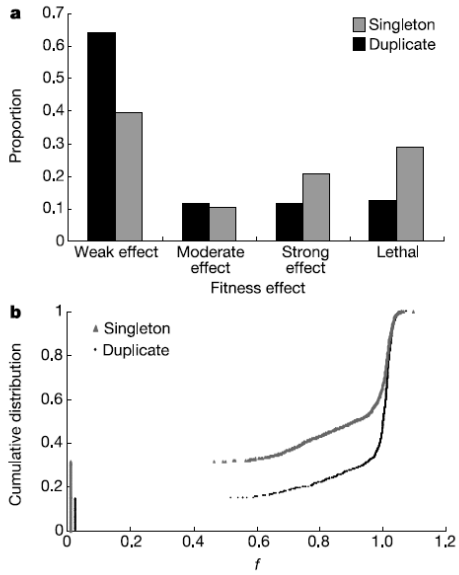

Figure 1 Distributions of fitness (*f*). *a*, Discrete distributions of *f* for singleton genes and for duplicate genes. The difference for the two distributions is significant (a contingency table test, $P<0.001$) under YPD growth conditions. *b*, Empirical cumulative distributions of *f* for singleton genes and for duplicate genes. The Kolmogrov–Smirnov test shows that the two distributions are significantly different ($P<10^{-15}$).

Similar fitness effects between duplicates

If duplicate genes tend to compensate for each other's function, then a testable prediction is that deletions of duplicate genes should tend to show similar fitness effects. To avoid multiple comparisons within a gene family, independent (non-overlapping) duplicate gene pairs were selected randomly from each gene family. For each of the 418 duplicate gene pairs selected (with both copies having been previously studied), we computed the difference between the fitness effects of duplicate genes i and j (D_{ij}) and then obtained the mean of all D_{ij} values (D*) for each growth condition. For comparison, 418 protein pairs were selected ran- domly from all previously studied genes and the D* value was calculated as above. This procedure was repeated 100,000 times to derive a frequency distribution of the mean difference in fitness effects between genes for each studied condition. The mean value (D*=0.193) for duplicate genes was far lower than the mean value for random gene pairs ($P<10^{-5}$) under the YPD growth condition (Fig. 2), confirming that duplicates tend to show more similar fitness effects of deletion than do random gene pairs. The same results held for the four other growth conditions.

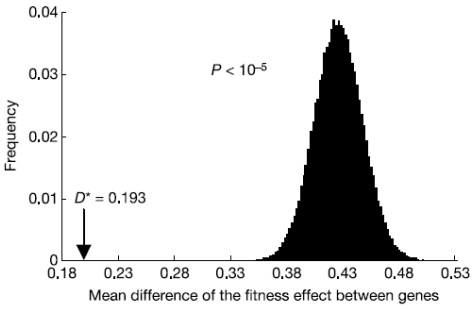

Figure 2 Distribution of mean fitness differences between randomly selected gene pairs (100,000 replicates each with 418 gene pairs) under the YPD growth condition. Arrow indicates the mean difference (D*=0.193) between duplicate genes

High correlation between duplicate compensation and the sequence similarity

The two genes derived from a duplication should initially have the same function. In long-term evolution, the accumulation of mutations in both copies over time results in either functional loss in one copy or functional divergence between the two copies[15]. If gene duplication is important for genetic robustness, genes with close paralogues should be compensated for deletion more often than genes with only distant paralogues. To test this hypothesis, we focused on a set of duplicate genes that excluded ribosomal proteins because of their unusual properties, such as strong codon usage bias, very high expression, and severe fitness

effects of null mutations. This set of duplicate genes was divided further into different groups on the basis of the nonsynonymous distance (K_A) of each gene to its most similar gene in the genome (defined as the gene with the smallest K_A value to the studied duplicate gene). Within each K_A interval, the frequencies of genes with different values of f were calculated. As expected, the proportion of genes with a weak or no effect decreased with K_A (correlation coefficient, R=-0.95; P<0.001), whereas the proportion of genes that are lethal when deleted increased with K_A (R=0.94; P<0.001; Fig. 3).

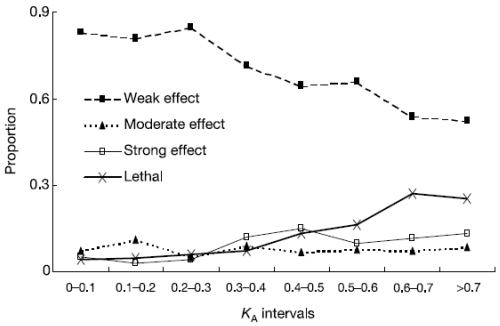

Figure 3 Relationship between protein distance and fitness effect of deletion. Protein distance is measured by the K_A of each gene to its most similar paralogue in the genome. The proportion of genes with a weak effect of deletion decreases with K_A (R =0.95, P<0.001), whereas the proportion of genes with lethal effect increases with K_A (R=0.94, P<0.001).

As the sequence similarity between duplicated genes decreases, their frequency of compensation will approach that for singletons. But even among duplicate genes with a K_A greater than 0.7 from their most similar paralogues in the whole genome, about 53% still had a weak effect or no effect when deleted (Fig. 3), which was significantly higher than the 39.5% of singletons that showed a weak or no effect of deletion (Fig. 1a, χ^2-test, P<0.01), implying that the compensation effect might exist even for ancient duplicate genes. Nevertheless, the decreasing compensation effect between duplicate genes with K_A suggests that the functional divergence of duplicate genes is coupled to some extent with the divergence of their protein sequences.

Correlation between severe fitness effect and expression level

Because expression level is used frequently to infer the activity and function of gene products, we tested whether deleting the copy of a duplicate gene that is more highly expressed would have a stronger fitness effect than deleting the copy with lower expression. We considered only duplicate gene pairs with different fitness effects of deletion (that is, those in which the relative fitness difference (as defined by $(f_i-f_j)/[(f_i+f_j)/2]$ for genes i and j was larger than 5%, or those for which one of the two duplicate copies is essential and the other is not). Expression (absolute transcript abundance) was estimated using available data measured by Affymetrix microarray experiments[16]. We used the fitness effects of the duplicate genes measured under the YPD growth condition in this analysis because the expression levels were also measured under the YPD growth condition[16]. Deleting the duplicate gene that has higher expression indeed tended to have a larger fitness effect (Table 1). For example, in 72 of the 98 gene pairs where the deletions had different fitness effects, the stronger fitness effect was seen in the more highly expressed gene.

Table 1 **Relationship between expression and fitness effect of null mutation in duplicate genes***

Relative expression	Number of genes with a larger fitness effect of deletion		
	Different effect†	One lethal	Similar effect†
High	72‡	50‡	125
Low	26	12	108
Total	98	62	233

*Only gene pairs for which both copies have been studied previously were included.
†Two duplicate genes i and j were said to have different fitness effects if $|(f_i - f_j)/[(f_i + f_j)/2]| > 0.05$, but similar fitness effects if otherwise.
‡Significant at $P < 0.001$.

Relative contributions

The analyses in this study provide strong evidence for the importance of duplicate genes in genetic robustness against null mutations. Yet, the high frequency of genes that have weak or no fitness effects of deletion among single genes, as well as among duplicate genes (Fig. 1a), indicates that the fitness effect of a gene deletion is also affected by factors other than copy redundancy. Whereas some genes may show null-mutation phenotypes only under experimental conditions that differ from the five growth conditions tested here, a fraction of weak, null-

mutation phenotypes among singletons might be due to compensation through alternative pathways or network branching, as suggested previously[10]. The relative importance of the compensating mechanism through functionally redundant duplicate genes can be estimated roughly as follows. If we assume that the extra proportion of genes with a weak or no fitness effect of deletion in duplicate genes when compared with the proportion for singleton genes is due to copy redundancy (64.3% for duplicates, 39.5% for singletons; difference 24.8%; Fig. 1a), this will give the lower bound of the contribution of gene duplication to genetic robustness.

Although our estimates are compatible with the view that interactions among unrelated genes rather than duplicate genes are the main cause of genetic robustness against mutations[10,18], two additional factors need to be considered. First, because we have considered only five growth conditions, it is possible that when a gene deletion showed no effect in any of these conditions it was not due to compensation by other genes but was because the gene deleted was not related to the growth conditions used. For this reason, our lower bound of 25% for the relative contribution of duplicate genes to compensation for null mutations is likely to be an underestimate. Second, a singleton could actually have one or more paralogues in the genome that cannot be detected by the criteria used but still overlap in function. Thus, gene duplication might be the ultimate origin of functional compensation for some 'singletons'. In conclusion, whether the contribution of gene duplication to genetic robustness is really less important than interactions among unrelated genes is an issue that remains to be resolved by further studies.

Further-reading

Gu, Z, Steinmetz, LM, Gu, X, Scharfe, C., Davis RD, Li, WH (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

References

1. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906 (1999).
2. Cadigan, K. M., Grossniklaus, U. & Gehring, W. J. Functional redundancy: the respective roles of the 2 sloppy paired genes in *Drosophila* segmentation. *Proc. Natl Acad. Sci. USA* 91, 6324–6328 (1994).
3. Saga, Y., Yagi, T., Ikawa, Y., Sakakura, T. & Aizawa, S. Mice develop normally without tenascin. *Genes Dev.* 6, 1821–1831 (1992).
4. Gibson, T. J. & Spring, J. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* 14, 46–49 (1998).
5. Normanly, J. & Bartel, B. Redundancy as a way of life—IAA metabolism. *Curr. Opin. Plant Biol.* 2, 207–213 (1999).
6. Brookfield, J. F. Y. Can genes be truly redundant? *Curr. Biol.* 2, 553–554 (1992).
7. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* 388, 167–171 (1997).
8. Tautz, D. Redundancies, development and the flow of information. *BioEssays* 14, 263–266 (1992).
9. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* 296, 910–913 (2002).
10. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nature Genet.* 24, 355–361 (2000).
11. Steinmetz, L. M. *et al.* Systematic screen for human disease genes in yeast. *Nature Genet.* 31, 400–404 (2002).
12. Seoighe, C. & Wolfe, K. H. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2, 548–554 (1999).
13. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155 (2000).
14. Conant, G. C. & Wagner, A. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30, 3378–3386 (2002).
15. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
16. Causton, H. C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* 12, 323–337 (2001).
17. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA* 85, 2444–2448 (1988).
18. Kitami, T. & Nadeau, J. H. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nature Genet.* 32, 191–194 (2002).
19. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391 (2002).
20. Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P. & Li, W.-H. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* 19, 256–262 (2002).
21. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680 (1994).
22. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43 (2000).
23. Holstege, F. C. P. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728 (1998).

Unit-3

Biased Mouse Knockout Genes Overestimates the Proportion of Essential Genes in Mouse

Introduction

Functional compensation of duplicate (paralogous) gene has been thought to be an important factor in the genetic robustness (Conant and Wagner 2004; Dean et al. 2008; Gu 2003; Gu et al. 2003; Harrison et al. 2007; Ihmels et al. 2007; Kamath et al. 2003; Winzeler et al. 1999). This is because the existence of a close paralogue in the same genome could increase the chance that null mutation of a given gene, by gene deletion, knockout, or RNAi technology, has little effect on organismal fitness, as called nonessential. Duplicate genes are indeed less essential than single-copy genes in both the yeast and the nematode (Gu et al. 2003; Conant and Wagner 2004; Kamath et al. 2003); two recent papers, however, reported that the proportion of essential genes (P_E) for singletons is similar to that for duplicates in mouse, based on the currently available mouse knockout phenotypes (Liang and Li 2007; Liao and Zhang 2007). However, they came up with different explanations. Liang and Li (2007) suggested that the potential compensatory role of gene duplication may have been counteracted by another factor—the more intrinsic importance of the duplicated genes. Instead, Liao and Zhang (2007) argued that duplicated genes may have a negligible role in mouse genetic robustness. In this study, we revisit this issue by analyzing the sampling bias for ancient duplicates in knockout genes, leading to an overestimation of the P_E of duplicates in mouse.

Methods

Mouse knockout data

The mouse phenotype and genotype association file (MGI_PhenoGenoMP.rpt) was downloaded from Mouse Genome Informatics (MGI 3.54; ftp://ftp.informatics.jax. org; release 10/23/2007). Only the phenotypic annotations of null mutation homozygotes that were generated by target deletion or gene trap were extracted for further analysis, excluding all phenotypic annotations due to multiple gene knockout experiments. In total 4123 genes with phenotypic information were extracted from this file. We then classified these genes into 1921 essential genes and 2202 nonessential genes. We defined an essential gene as a gene whose knockout phenotype is annotated as lethality (including embryonic, perinatal, and postnatal lethality) or infertility.

Dating duplication age of mouse duplicate genes

We followed the method of Gu et al. (2002) to identify duplicate genes and single-copy genes, and chose the FASTA not-self best hit of a duplicate gene as its closest parologue (Pearson 2000). We have developed an analytical pipeline to estimate the duplication age (time) between each mouse duplicate gene and its cloest paralogue on a large scale, using the split time between mouse and zebrafish (430 million years ago [mya]) as a calibration. By this method, the duplicate ages between each of 9503 mouse genes and its closest paralogues were estimated (the whole-genome set). Among these, 2260 genes were knockout target genes (knockout set). In addition, we used several other speciation events as calibrations, e.g., the splits of mammal-bird (310 mya) and primate-rodent (80 mya), and found that our main results are robust (not shown). To be concise we present here the results based on the mammal-zebrafish split time calibration.

Predicting P_E in the Mouse Genome: Bias Correction

In statistics, the proportion of essential genes in duplicates (P_E) estimated from a sample of mouse knockout genes can be considered an unbiased estimate for the genome value, provided that the assumption of random

sampling holds. While the expectation of P_E for singletons is $1-B$, where B is the proportion of genetic buffering, the expectation of P_E for duplicates is $(1-B)A$, where A is the mean effect of duplicate compensation. Since mouse knockout genes were apparently not randomly selected with respect to the age distribution of duplicates, it is necessary to examine whether the potential sampling bias may have affected our estimation.

We suspect that the underrepresentation of young duplicates in mouse knockout phenotypes could mislead our understanding of the duplication effect (A) on mouse genetic robustness. To test this argument we designed the following analytical procedure: on the basis of duplicate age (t), M duplicate genes in the knockout sample are grouped into several (n) bins, each of which has a 100 million-year interval. For each bin i we calculate the gene frequency ($f_i = M_i/M$, where M_i is the number of [knockout] duplicate genes in bin i) and the proportion of essential genes by $P_{E,i} = m_i/M_i$, where m_i is the number of essential genes in bin i. Assume that the mouse genome frequency for each age bin is known, denoted g_i. If the correlation between P_E and age t in the knockout sample is statistically significant, differences in gene frequency between the sample (f_i) and the genome (g_i) point to the potential resource that may cause the bias in P_E estimation. Under the assumption that the same P_E-t correlation holds in the mouse genome, the bias-correcting predicted P_E in the mouse genome is therefore calculated by

$$P_E^* = g_1 P_{E,1} + g_2 P_{E,2} \cdots + g_n P_{E,n} \qquad (1)$$

For each bin i, the expectation of $P_{E,i}$ is $(1 - B)A_i$, where A_i is the mean effect of duplicate compensation at this age interval. Hence, the expectation of P^*_E is $(1 - B)A$ if the discrete genome frequency (g_i) is an unbiased representation of age distribution of duplicates so that $A = g_1A_1 + g_2A_2 \cdots + g_nA_n$. Note that the observed P_E in the sample of knockout duplicates can be written

$$P_E = f_1 P_{E,1} + f_2 P_{E,2} \cdots + f_n P_{E,n} = m/M \qquad (2)$$

where $m = m_1 + \cdots + m_n$ is the number of essential duplicate genes in the knockout sample. Obviously, if the knockout sample is large enough to cover the whole mouse genome, the bias in P_E estimation would be trivial, which can be estimated by m/M without knowing the ages of duplicate genes. The observed P_E is expected to be equal to P^*_E only if the duplicate gene frequency of each bin is the same between the knockout sample and the genome, i.e., $f_i = g_i$. Finally, from Eq. 1, one can calculate the sampling variance of P^*_E.

Bias-corrected P_E in mouse duplicates

Because most mouse knockout experiments have been carried out by individual laboratories for finding detectable knockout phenotypes, one may suspect that recently duplicated genes have been purposely avoided to minimize the experimental cost of negative-phenotype results. Consequently, recently duplicated genes may have been underrepresented in the mouse knockout database. If this is the case, the sampling bias could be one of the most obvious reasons to explain why there was no statistical difference in P_E between mouse singletons and mouse duplicates.

We conducted a direct comparison of duplication age (mya) of mouse duplicate genes between the whole-genome set and the knockout set (Table 1). Apparently, the ages of most duplicates in the mouse knockout dataset were dated at about 500–700 mya, and recently duplicated genes, say, <100 mya, were seriously underrepresented in the mouse knockout set: 1.4% in the knockout set versus 19.6% in the mouse genome set. In other words, the sampling bias toward ancient duplicates in the currently available mouse knockout target genes is nontrivial.

Since young duplicates are expected to have high degrees of functional compensation between them, resulting in a low proportion of essential genes (PE), degrees of functional compensation between them, result- ing in a low proportion of essential genes (P_E), the age bias in mouse knockout duplicates may cause an overestimation of P_E in mouse duplicates. To avoid this bias, we calculated P_E in each age interval of 100 million years (age bin), respectively. As reported in Table 1, we found a significantly positive P_E-age(t) correlation ($p < 0.001$, χ^2 test). Apparently, the ancient duplicates may have undergone substantial functional divergence so that they have lost the capacity for functional compensation. In contrast, the young duplicates, those duplicated around the mammalian

radiation or during the rodent lineage, are expected to make significant contributions to gene robustness in the current mouse genome. Therefore, the proportion of essential genes (P_E) in young duplicates is much lower than that of singletons. When young duplicates were considerably underrepresented in the mouse knockout dataset, it is actually not very surprising that the observed P_E in this biased sample of mouse knockout duplicates was close to that for single-copy genes.

Table 1 Duplicate age (t) distribution of mouse knockout gene set and genome set

	$t < 100$	$100 \le t < 200$	$200 \le t < 300$	$300 \le t < 400$	$400 \le t < 500$	$500 \le t < 600$	$600 \le t < 700$	$700 \le t < 800$	$800 \le t < 900$	$900 \le t < 1000$	$t \ge 1000$
No. of knockout genes	32	34	64	75	256	419	424	310	153	106	387
Frequency in knockout set (f_t)	0.014	0.015	0.028	0.033	0.113	0.185	0.188	0.137	0.068	0.047	0.171
No. of essential genes	7	10	18	24	99	196	210	166	77	59	216
P_E	0.219	0.294	0.281	0.32	0.387	0.468	0.495	0.535	0.503	0.557	0.558
Frequency in whole-genome set (g_t)	0.196	0.057	0.043	0.042	0.074	0.131	0.134	0.092	0.054	0.034	0.143

The duplication age (t) shows statistically significant differences between the sample and the genome, and the positive P_E-age correlation provides a theory for the potential resource of sampling bias in P_E. We used the duplicate age (t) to obtain a corrected P_E in duplicates of the mouse genome. Finally, we predicted that P_E=41.7% for all duplicate genes, which is impressive compared to the P_E=46.3% observed in sample duplicates ($p<10^{-4}$, χ^2-test) and P_E = 47% in sample singletons ($p< 10^{-4}$) (Fig. 2). We also found that P_E in young duplicates is much lower than in all duplicates and sample singletons. For instance, we estimated P_E = 25.2% and 24.2% for those duplicates with a duplicate age no greater than 400 and 300 mya (Fig. 2), respectively, indicating that those young duplicates that duplicated after the tetrapod-teleost split (about 430 mya) have a significant effect on gene robustness in the mouse genome.

In short, our analysis indicates that the sampling bias for ancient duplicate age of knockout genes caused the overestimation of PE in mouse duplicates. Therefore, the similar proportions of essential genes between singletons and duplicates in the currently biased sample should not be taken as evidence supporting the claim that the role of functional compensation for duplicate genes in the mouse is negligible. Besides, it is intriguing that very ancient duplicate genes have an even higher percentage of essential genes than singletons (Table 1), suggesting that duplicate genes might be more important.

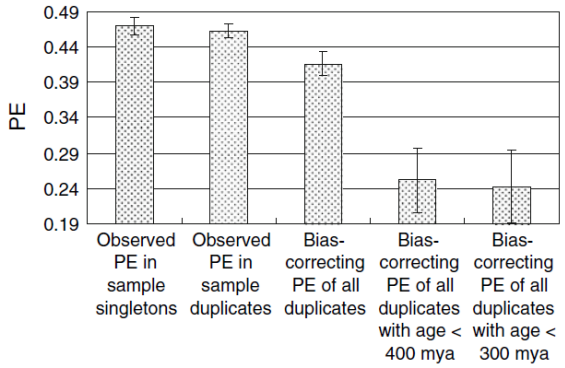

Fig. 2 Proportion of essential genes (P_E) in different gene categories in the mouse genome. Error bars show 1 SE

Further-reading

Z. Su, and X. Gu* (2008) Predicting the Proportion of Essential Genes in Mouse Duplicates Based on Biased Mouse Knockout Genes. Journal of Molecular Evolution 67:705-709.

References

Conant GC, Wagner A (2004) Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271:89–96

Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4(7):e1000113

Gu X (2003) Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet 19:354–356

Gu Z, Cavalcanti A, Chen F-C, Bouman P, Li W-H (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol 19:256–262

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63–66

Harrison R, Papp B, Pal C, Oliver SG, Delneri D (2007) Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci USA 104:2307–2312

Ihmels J, Collins SR, Schuldiner M, Krogan N, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3:86

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–237

Liang H, Li W-H (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet 23:375–378

Liao B-Y, Zhang J (2007) Mouse duplicate genes are as essential as singletons. Trends Genet 23:378–381

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219

Winzeler EA, et al (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285:901–906

Unit-4

Effect of Duplicate Genes on Mouse Genetic Robustness: An Update

Introduction

Functional compensation of duplicate (paralogous) genes has been thought to play an important role in genetic robustness [1–7]. Indeed, existence of a close paralog in the same genome could result in null mutations of the gene with little effect on the organismal fitness (nonessential gene), as observed in both yeast and nematode [1–4]. However, the role and magnitude of the duplicate genes contributing to genetic robustness in mammals (mouse) remain controversial [8–13]. Since knockout mice have been widely used as animal models for human diseases, resolving this issue may have a significant impact on biomedical sciences. There are several hypotheses proposed in the literature.

The Duplicability Hypothesis. By combining the protein- protein interaction data into the analysis, Liang and Li [9] found that mouse duplicate genes tend to have much higher protein connectivity than those for singletons. Since high connectivity means high functional centrality in the gene network, they proposed that mouse duplicates probably are more important than singletons and that this factor could compromise the contribution of duplicate compensation. In other words, functionally important genes may have more chance to be duplicated. It remains unexplained why more important mouse genes tend to be duplicated, while yeast genes may have the opposite trend [14].

The No-Role Hypothesis. Liao and Zhang [10] argued that the compensational role of duplicates in mouse genetic robustness is negligible. After examining a number of genomic factors, they discussed several possibilities that may result in similar proportion of essential genes between single- tons and duplicates. It implies that most recently duplicated mouse genes, for example, 26 rodent-specific prolactin-like proteins [15], may have lost functional compensations to each other. This prediction seems to be counterintuitive and does not receive much experimental evidence for supporting.

Age-Distribution Hypothesis. Su and Gu [11] have noticed the effect of sampling bias: recently duplicated genes, for example, after the mammalian radiation, are severely underrepresented in the current mouse KO database. Because most of the mouse gene knockouts were generated by individual laboratories for finding knockout phenotypes, recently duplicated genes may have been purposely avoided to minimize the experimental cost due to negative-phenotype results. In other words, the age distribution of duplicates in the data sample is upwardly biased, resulting in underestimation of the overall duplicate effect on the genetic robustness.

The Functional Importance Hypothesis. Makino et al. (2009) reported that there is a strong sampling bias towards the duplicated genes generated by whole genome duplication (WGD) in current mouse KO phenotype dataset [12]. Hsiao and Vitkup [8] suggested an important role in robustness against deleterious mutations of duplicate genes in human [8].

In this study, we use an updated mouse KO dataset to carry out an extensive analysis. To facilitate the study, we proposed an empirical evolutionary model of gene essentiality— the A&B model (Age of duplication and genetic Buffering)— to explain knockout duplicate puzzle. Our results suggest that duplication age and genetic buffering determine the essentiality of mouse duplicates.

Sampling bias in mouse KO genes toward ancient duplicates

Of the 4123 mouse genes with available phenotypic data, 1921 were identified as essential genes. Meanwhile, we identified 2479 duplicate genes and 464 singleton genes and calculated proportions of essential genes (P_E), respectively. Consistent with previous studies [9–12], the updated mouse KO dataset shows no statistical difference of P_E between singletons and duplicates (44.8% versus 46.3%). Based on a more broad definition of gene essentiality (Materials and Methods), that is, genes with premature death or induced morbidity phenotype were considered as essential genes, we found the same pattern.

We estimated the duplication times between 2260 mouse knockout genes and their closest paralogs and found that the age distribution of KO duplicate pairs differs significantly from the whole-genome duplicate pairs ($P<10^{-16}$, χ^2-test). The histograms in Figure 1 clearly show that mouse KO experiments have been designed to avoid recently duplicated genes, for example, only 1.4% for those duplicated within 100 mya (around or after the mammalian radiation) in the KO set, compared to 19.6% in the mouse genome set. Consequently, the ages of duplicate genes in the mouse knockout dataset are typically around 500 to 700 mya (in early vertebrates), with a long-tail toward even more ancient ones (>1000 mya). In other words, the sampling bias toward ancient duplicates in the currently available mouse KO target genes has been nontrivial. These ancient duplicates may have undergone substantial functional divergence so that they have lost the capacity of functional compensation. In contrast, recent gene duplications, those duplicated around the mammalian radiation or in the rodent lineage, are expected to have significant contributions to the gene robustness in the current mouse genome. While these young duplicates were considerably underrepresented in the mouse knockout dataset, the observed proportion of essential duplicate genes is upwardly biased close to the value of singletons.

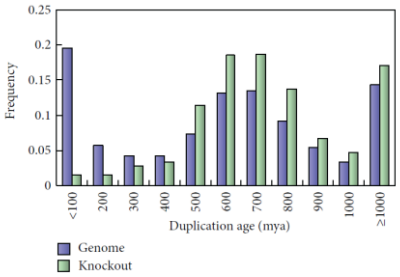

Figure 1: Duplication age distribution of mouse genome set (blue bars) and knockout gene set (green). The *x*-axis indicates the duplication age (*t*) between a duplicated gene and its closest paralog. The *y*-axis indicates the frequency of the duplicates in each duplication age category

The Duplication-Age and Buffering Model (Age-Buffering Model) of Gene Essentiality

Since initially duplicated genes were completely compensated, the loss process of duplicate compensation is apparently time dependent, during which the outcome can be influenced by many gene-specific factors. To have a complete understanding of gene essentiality in duplicates and singletons, an evolutionary model is needed. We formulate a simple *A&B* model as follows, short for *Age of duplication and genetic Buffering*. Without genetic buffering, we assume that the probability of a duplicate remains nonessential, that is, functionally compensated by another duplicate copy in the same genome, and decayed exponentially with the time t (the age of gene duplication), that is, $e^{-\lambda t}$, where λ is the loss rate of duplicate compensation by mutations. Next, let g be the probability that a gene is genetically buffered. Together, the *A&B* model demonstrates that a gene to be essential depends on two mechanisms: the effect of genetic buffering (g) and the age-dependent effect of duplication compensation ($e^{-\lambda t}$). Obviously, the probability of a duplicate gene being essential is the probability for both mechanisms failure, that is,

$$P_E = (1 - g)(1 - e^{-\lambda t}) \quad (1)$$

In fact, Eq.(1) suggests that three parameters, t (duplication age), g (genetic buffering), and λ (loss rate of functional compensation), together determine the gene essentiality of mouse duplicates. In particular, we have two arguments: (*i*) the proportion of essential genes in mouse duplicates (P_E) is age dependent on gene duplications; (*ii*) gene essentiality correlates to sequence conservation or protein connectivity in either duplicates or singletons largely because these two factors affect the efficiency of genetic buffering (g), rather than the functional compensation between duplicates. Moreover, our models suggested that, for sufficient time, P_E approaches to a level that is roughly equal to P_E of singleton. However, it does not mean that all these ancient duplicates are subject to the genetic buffering. A likely situation is that genetic buffering and duplication coevolve. In other words, the reason why some duplicates can remain dispensable for a long time is because they were integrated into existing or novel genetic buffering mechanisms.

Chen et al. (2010) found that in *Drosophila* new genes could become essential rapidly after the gene duplications [23]. This mechanism is also likely to exist in mammals. To take this factor into account, we modify (1) as follows:

$$P_E = (1 - g)[1 - (1 - \rho)e^{-\lambda t}], \quad (2)$$

where the parameter $\rho > 0$ indicates the process of rapid essentiality in the early stage after gene duplication. Because the number of mouse KO genes is small for very young duplicates, a further investigation requires when the data are available.

Proportions of Essential Genes (P_E) in Mouse Duplicates Are Age Dependent.

A simple solution to correct this knockout sampling bias is to calculate P_E under a given age bin. We implemented several approaches to minimize the noise effect in time estimation. First, we used three time calibration points to date mouse duplication events: the mammal- zebrafish split (430 mya), the mammal-bird split (310 mya), and the primate-rodent split (80 mya), respectively, and calculated P_E for every age bin of 100 million years. As shown in Figure 2, in all cases we observed that P_E increases from a low value in young duplicates with the increasing of duplication ages; this P_E-age (t) correlation is statistically significant ($P<10^{-4}$, χ^2-test). To be concise, in the following of this paper, we mainly present the results based on the mammal- zebrafish split time calibration. Noticeably, we found that PE in ancient duplicates, say, >700 mya, is unexpectedly higher than that of singletons; $P_E = 0.542 \pm 0.016$, $P < 0.001$. Next, we inferred the phylogenetic locations of mouse KO duplication events in three intervals: after the mammal-zebrafish split, after the mammal-bird split, and after the primate-rodent split. In each interval we calculated P_E, which is compatible to the proportion of essential genes, with respect to the three major speciation events in vertebrates: P_E is ~23% for those duplicated after the mammalian radiation, ~31% for those duplicated after the bird-mammal split, and close to ~39% for those duplicated after the teleost-tetrapod split. Although a decreasing P_E in younger duplicates is biologically intuitive, it is subject to the statistical uncertainty due to small sample size. Under a more broad age category, such as before the split of land animals and fishes versus the more ancient duplicates, the difference is statistically significant ($P < 0.01$).

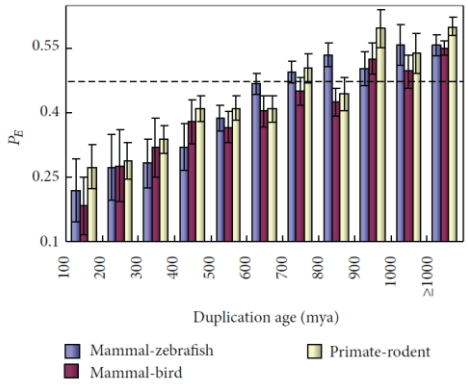

Figure 2: Relationship between P_E in duplicate genes and the duplication age. Error bars show one standard error. The dashed line indicates the P_E level of single-copy genes.

Age Dependence of P_E in Mouse Duplicates and Sequence Conservation

Though a simple interpretation for the P_E-t correlation is that the capability of duplicate compensation decays with the evolutionary time since the duplication [11], some other alternatives cannot be ruled out, which were based on the correlation of gene essentiality with, for instance, sequence conservation or protein connectivity [9, 10, 24].

To measure the sequence conservation, we used the conventional ratio of the number of nonsynonymous substitutions per site (d_N) to the number of synonymous substitutions per site (d_S), which was estimated from the mouse gene and its human ortholog. Consistent with previous studies [10, 25], we showed that essential mouse genes tend to be more conserved: P_E decreases with the increase of d_N/d_S for both duplicates (Spearman rank $\rho = -0.23$, $P<10^{-15}$) and singletons ($\rho = -0.18$, $P<10^{-15}$); see Figure 3(a) for binned results. After calculating the mean d_N/d_S ratio for each age bin of mouse duplicates, we unexpectedly found that sequence conservation is actually positively correlated with the duplication age (t) (Figure 3(b), $P<10^{-10}$). This unexpected inverse age-d_N/d_S relationship raises the possibility that the observed P_E-t (age) correlation could be confounded by theP_E-d_N/d_S correlation conjugated with the age-d_N/d_S correlation.

We first claim that the P_E-d_N/d_S correlation is the consequence of the inverse relationship between the genetic buffering (g) and the sequence conservation (d_N/d_S). Hence, the inverse age-d_N/d_S relationship in mouse duplicates suggests less effect of genetic buffering in ancient duplicates than that in recent duplicates, implying that the genetic buffering of duplicates g could be age dependent. One possible evolutionary mechanism for the age-g inverse relationship could be the neo-functionalization in the late stage after the gene duplication so the preexisting (ancestral) genetic buffering systems did not work for the newly acquired functions.

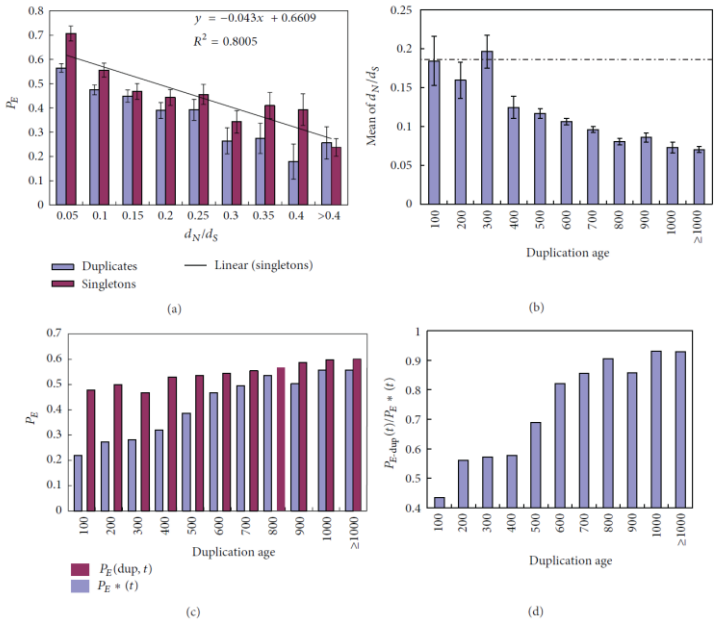

Figure 3: The effect of sequence conservation on the relationship between P_E and duplication age. (a) Relationship between P_E in duplicate genes (blue) or singletons (purple) and the evolutionary conservation of the gene, measured by the ratio of the nonsynonymous (dN) to synonymous (dS) nucleotide distances between the target gene and its human ortholog. Linear regression line and regression equation between dN/dS ratio and PE in knockout single-copy genes are presented on the panel. (b)Mean dN/dS ratio for each age bin of duplicates. Dashed line denotes the mean dN/dS ratio of singleton mouse knockout genes. (c) PE in each age bin of duplicates—PE (dup, t)—and that of singletons with the same dN/dS ratio—$P*E(t)$. $P*E(t)$ is calculated based on the mean dN/dS ratio for duplicates in each age bin (panel b) and the linear regression equation (panel a). (d) Ratio of PE (dup, t) and $P*E$ (t) in each age bin of duplicates. Error bars show one standard error.

Age Dependence of P_E in Mouse Duplicates and Protein Connectivity.

The proportion of essential genes is positively correlated with protein connectivity in mouse [9]. In our updated mouse KO dataset, we compiled 211 singleton mouse KO targeted genes with available protein connectivity data, as well as 845 mouse KO duplicates [26]. Consistent with [9], we confirmed a weak but significant positive correlation between protein connectivity and P_E in both duplicates (Spearman rank $\rho = 0.11$, $P = 0.001$) and singletons ($\rho = 0.11$, $P = 0.003$; see Figure 4(a) for binned results). Similar to the effect of sequence conservation, the A&B model interprets this finding as genes with high connectivity may have low genetic buffering. Due to the small sample size, we further group the 845 genes into seven age groups. We then calculated the mean of protein interaction number for duplicated genes in each age bin and found no correlation of the mean protein connectivity with the duplication age (t) (Spearman rank $\rho = 0.04$, $P = 0.19$, Figure 4(b)).

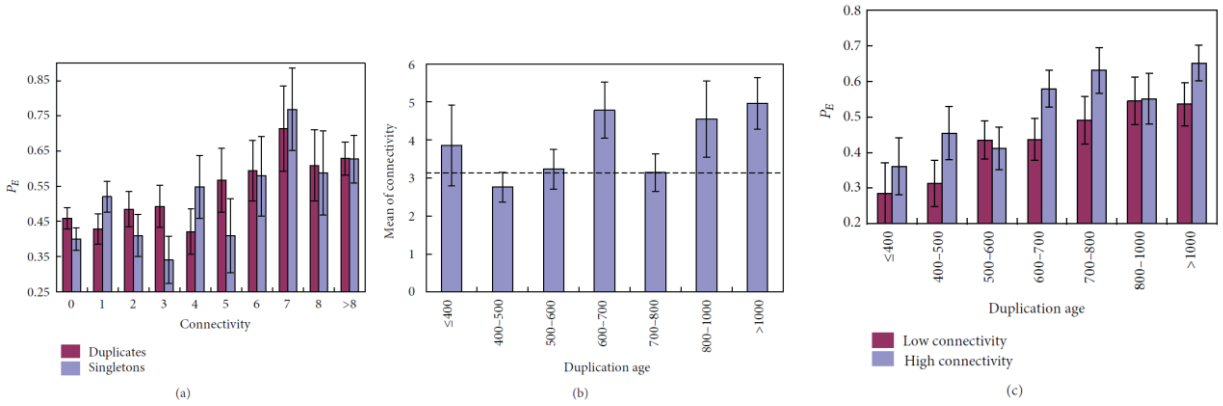

Figure 4: The effect of protein connectivity on the relationship between PE and duplication age. (a) Relationship between PE in duplicate genes (blue) or singletons (purple) and the protein connectivity of the gene. (b) Mean interaction number for each age bin of duplicates. Dashed line denotes the mean interaction number of singleton mouse knockout genes. (c) Relationship between PE in duplicate genes and the duplication age for high connectivity genes and low connectivity genes. Error bars show one standard error.

We thus hypothesize that P_E-connectivity and P_E- age correlations reflect two independent underlying mechanisms.

To further test this hypothesis, we divided duplicate genes with interaction data into two groups, those with high connectivity (larger than the median interaction, i.e., >2 interactions) and those with low connectivity (otherwise). The proportion of essential genes in the high- connectivity group is apparently higher than that in the low-connectivity group ($P < 0.001$). But, as shown in Figure 4(c), the inverse relationship between P_E and the age of duplicates holds in both gene groups. We thus conclude that age dependence of the proportion of essential genes (P_E) in duplicates is unlikely to be confounded by the effect of protein connectivity.

What Determines Duplicate Compensation: Evolutionary Time (Age) or Sequence Conservation?

The protein sequence divergence between duplicate genes, or the evolutionary distance (d), was widely used as a proxy measure of the age of duplicates. In our study we used the Poisson-corrected method to estimate the protein sequence distance (d). Figure 6(a) shows no correlation between P_E and d, as claimed in [10]. A straightforward explanation is that the sequence distance between duplicates (d) is determined by $d=2vt$, where v is the evolutionary rate of the protein sequence and t is the age of duplicates. As shown in Figure 3(b), an ancient duplicate gene (a large t) tends to be conserved (low v as measured by low d_N/d_S ratio) so that the P_E-d independence could be the result of canceled P_E-t and P_E-d_N/d_S correlations. Our conclusion that the P_E-d relationship is not fundamental differs from Liao and Zhang [10]. Assuming that it is the protein sequence similarity, not the age of gene duplication, which determines the likelihood of compensation between duplicates, the authors of [10] argued that the lack of correlation between P_E and d may indicate the negligible role of duplicate genes in the mouse genetic robustness. Here, we conduct a simple case-study to show that it may not be the case. We divided 135 mouse KO duplicate pairs with $d<0.2$ (corresponding to 82% sequence identity between KO duplicates and their paralogs) into the "young" group (age<310 mya, after the bird-mammal split) or the "old" group (≥310 mya). Strikingly, we found $P_E = 0.39$ for the young group and $P_E = 0.58$ for the old group ($\chi^2=4.56$, $P = 0.03$) (Figure 6(b)). Moreover, we calculated the mean sequence conservation (the d_N/d_S ratio) in both groups: $d_N/d_S = 0.12$ for young duplicates and 0.02 for ancient duplicates. Does this mean that different P_E in young and old groups is caused by the difference in sequence conservation? From theP_E-d_N/d_S regression in singletons (Figure 3(a)), we predict that, if there is no functional compensation between duplicates, the young group should have the $P_E = 0.56$ versus the old group $P_E = 0.64$ (Figure 6(b)), which is contradictory to our observation. We therefore conclude that, for these duplicate pairs with >82% protein sequence identity, recent duplicate pairs are functionally more compensated than ancient pairs.

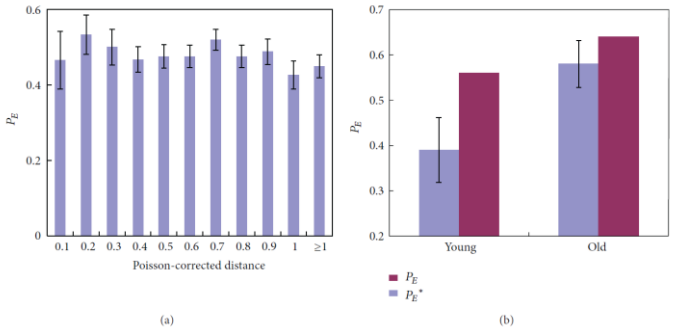

Figure 6: Relationship between PE and protein sequence divergence. (a) PE in duplicate genes is not correlated with the Poisson-corrected distance (d) between the target gene and its closest paralog in the genome. Error bars show one standard error. (b) PE and $P*E$ of mouse knockout duplicate pairs with sequence divergence $d < 0.2$. "Young" group represents the knockout genes with duplication age <310mya, and "old" group represents the knockout genes with duplication age ≥310mya. $P*E$ is calculated based on the mean dN/dS ratio for each group and the linear regression equation of Figure 3(a).

The A&B model we proposed suggests that the age of gene duplication plays an important role in functional compensation between duplicates, while the sequence conservation indicates the likelihood of a duplicate gene actually genetically buffered by other (non-homologous) genes, as supported by recent double deletions of yeast duplicate pairs [29, 30]. Noticing that, in many cases, the sequence similarity and functional similarity between paralogs may not be strongly correlated [31], we tentatively propose the transient hypothesis for the observed P_E-age correlation.

That is, because only a few nucleotide substitutions are responsible for the compensation loss between duplicates, the time interval for maintaining the effective compensation between duplicates mainly depends on the "waiting time" for these substitutions to occur.

Technical comments

Genetic robustness in difference species

It is interesting to find that P_E seems to increase with organismal complexity. That is, though a greater fraction of genes in complex organisms may have been essential to ensure viability and fertility than that in simple organisms, for example, under laboratory conditions, P_E is ~7% in *Escherichia coli* [32], 17% in yeast [8, 33], and >46% in mouse. The effect of gene duplications on genetic robustness depends on the distribution of young duplicate genes in the current genome. Therefore, its impact varies among species, mainly because each species has its unique age distribution of gene duplications. In the yeast *Saccharomyces cerevisiae*, the most recent WGD event occurred relatively recently (in the last ~100 million years) [34], much younger than the vertebrate genome duplication (in the last ~500 millions). Therefore, the majority of the yeast duplicated genes are quite young. For example, we found that only 13.1% of the yeast duplicates were generated 500 mya, whereas 58.9% of the mouse duplicates were created 500 mya. By contrast, due to recent polyploidizations (normally in the range from 1-50 mya), duplicate genes may dominate the genetic robustness in plant genomes [47]. Nevertheless, the age-dependent effect of duplicates on gene robustness remains similar from simple to complicated organisms, as shown by the yeast and mouse.

Ancient duplicate gene tends to be more essential

We observed that ancient duplicates tend to be more conserved, and the ancient duplicate gene tends to be more essential than an average single-copy gene. It is somewhat puzzling because it is generally believed that duplicated genes may have experienced a relaxed evolution due to the functional redundancy. One possible explanation invokes to the positive selection in the follow-up neo-functionalization, which ultimately imposes a stronger functional constraint on the protein sequence. Though it stands as an interesting hypothesis, we offer a much simpler explanation: for those ancient duplicate genes originated over 500 mya, only highly conserved duplicate pairs can be detected by the standard homologous search. In other words, sequence similarity between ancient duplicate genes with relatively low sequence conservation may be too low to be detected.

Ancient functional compensation

In our study, the duplication age was estimated between the mouse KO gene and its closest paralog. Many mouse KO genes have more than one paralog, consisting of a large gene family. In such cases the pattern of functional compensation is complex, which cannot be revealed because most members have no KO phenotype information. Our approach is based on the premise that the closest paralog is the major determinant of functional compensation. Of course our treatment could be biased, and the future study should be gene-family based. The bottleneck still is the lack of sufficient KO genes. We indeed conducted a preliminary survey of the distribution of KO genes in a family but the dataset is too small to be useful at the current stage. Another technical issue is about the age of singleton. While we use the common procedure to determine singletons, the age of gene does affect P_E in both duplicate and singleton genes. One may see Chen et al. (2012) for details [44].

Further-reading

Su Z, Wang J, Gu X* (2014) Effect of duplicate genes on mouse genetic robustness: an update. BioMed research international 2014:758672. doi:10.1155/2014/758672.

References

[1] G. C. Conant and A. Wagner, "Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans," *Proceedings of the Royal Society B: Biological Sciences*, vol. 271, no. 1534, pp. 89–96, 2004.
[2] E. J. Dean, J.C.Davis, R.W.Davis, and D. A. Petrov, "Pervasive and persistent redundancy among duplicated genes in yeast,"*PLoS Genetics*, vol. 4, no. 7, Article ID e1000113, 2008.

[3] Z. Gu, L.M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, andW. Li, "Role of duplicate genes in genetic robustness against null mutations," *Nature*, vol. 421, no. 6918, pp. 63–66, 2003.
[4] Y. Guan, M. J. Dunham, and O. G. Troyanskaya, "Functional analysis of gene duplications in *Saccharomyces cerevisiae*," *Genetics*, vol. 175, no. 2, pp. 933–943, 2007.
[5] X. Gu, "Evolution of duplicate genes versus genetic robustness against nullmutations," *Trends in Genetics*, vol. 19,no. 7, pp. 354–356, 2003.
[6] R. S. Kamath,A.G.Fraser, Y. Dong et al., "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi," *Nature*, vol. 421, no. 6920, pp. 231–237, 2003.
[7] E. A. Winzeler, D. D. Shoemaker, A. Astromoff et al., "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, 1999.
[8] T.-L. Hsiao and D. Vitkup, "Role of duplicate genes in robustness against deleterious human mutations," *PLoS Genetics*, vol. 4, no. 3, Article ID e1000014, 2008.
[9] H. Liang and W. Li, "Gene essentiality, gene duplicability and protein connectivity in human and mouse," *Trends in Genetics*, vol. 23, no. 8, pp. 375–378, 2007.
[10] B.-Y. Liao and J. Zhang, "Mouse duplicate genes are as essential as singletons," *Trends in Genetics*, vol. 23, no. 8, pp. 378–381, 2007.
[11] Z. Su and X. Gu, "Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes,"*Journal ofMolecular Evolution*, vol. 67, no. 6, pp. 705–709, 2008.
[12] T. Makino, K. Hokamp, and A. McLysaght, "The complex relationship of gene duplication and essentiality," *Trends in Genetics*, vol. 25, no. 4, pp. 152–155, 2009.
[13] K. Hannay, E. M. Marcotte, and C. Vogel, "Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation," *BMC Genomics*, vol. 9, article 609, 2008.
[14] A. Prachumwat andW. Li, "Protein function, connectivity, and duplicability in yeast," *Molecular Biology and Evolution*, vol. 23, no. 1, pp. 30–39, 2006.
[15] D. O. Wiemers, L. J. Shao,R.Ain,G.Dai, andM. J. Soares, "The mouse prolactin gene family locus," *Endocrinology*, vol. 144, no.1, pp. 313–325, 2003
[16] X. Gu, Y. Wang, and J. Gu, "Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution," *Nature Genetics*, vol. 31, no. 2, pp. 205–209, 2002.
[17] G. Panopoulou, S. Hennig, D. Groth et al., "New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes," *Genome Research*, vol. 13, no. 6, pp. 1056–1066, 2003.
[18] K. Vandepoele, W. de Vos, J. S. Taylor, A. Meyer, and Y. van de Peer, "Major events in the genome evolution of vertebrates: paranome age and size differ considerably between rayfinned fishes and land vertebrates," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1638–1643, 2004.
[19] W. H. Li, *Molecular Evolution*, Sinauer Associates, Sunderland, Mass, USA, 1997.
[20] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.
[21] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, Germany, 1970.
[22] A. McLysaght, K. Hokamp, and K. H. Wolfe, "Extensive genomic duplication during early chordate evolution," *Nature Genetics*, vol. 31, no. 2, pp. 200–204, 2002.
[23] S. Chen, Y. E. Zhang, and M. Long, "New genes in Drosophila quickly become essential," *Science*, vol. 330, no. 6011, pp. 1682–1685, 2010.
[24] H. Jeong, S. P.Mason, A. L. Barab´asi, andZ. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
[25] A. E. Hirsh and H. B. Fraser, "Protein dispensability and rate of evolution," *Nature*, vol. 411, no. 6841, pp. 1040–1049, 2001.
[26] J.-F. Rual, K. Venkatesan, T. Hao et al., "Towards a proteomescale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.
[27] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer, "The gain and loss of genes during 600 million years of vertebrate evolution," *Genome Biology*, vol. 7,
no. 5, article R43, 2006.
[28] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita, "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates," *Genome Research*, vol. 17, no. 9, pp. 1254–1265, 2007.
[29] J. Ihmels, S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, "Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss,"*Molecular Systems Biology*, vol. 3, 2007.
[30] R. Harrison, B. Papp, C. P´al, S. G. Oliver, and D. Delneri, "Plasticity of genetic interactions in metabolic networks of yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2307–2312, 2007.
[31] H. H. Gan, R. A. Perlow, S. Roy et al., "Analysis of protein sequence/structure similarity relationships," *Biophysical Journal*, vol. 83, no. 5, pp. 2781–2791, 2002.
[32] T.Baba, T. Ara,M.Hasegawa et al., "Constructionof *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection,"*Molecular Systems Biology*, vol. 2, article 0008, 2006.
[33] X.He and J. Zhang, "Why do hubs tend to be essential in protein networks?" *PLoS Genetics*, vol. 2, no. 6, p. e88, 2006.
[34] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.
[35] A.M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K. J. Peterson, "MicroRNAs and the advent of vertebrate morphological complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2946–2950, 2008.
[36] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind, "The role of lineage-specific gene family expansion in the evolution of eukaryotes," *Genome Research*, vol. 12, no. 7, pp. 1048–1059, 2002.
[37] N. Lopez-Bigas, S. de, and S. A. Teichmann, "Functional protein divergence in the evolution of *Homo sapiens*," *Genome Biology*, vol. 9, no. 2, article R33, 2008.
[38] A. Prachumwat and W. Li, "Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes," *Genome Research*, vol. 18, no. 2, pp. 221–232, 2008.
[39] C. Vogel and C. Chothia, "Protein family expansions and biological complexity," *PLoS Computational Biology*, vol. 2, no. 5, p. e48, 2006.
[40] G. Liu, Y. Zou, Q. Cheng, Y. Zeng, X. Gu, and Z. Su, "Age distribution patterns of human gene families: divergent for Gene Ontology categories and concordant between different subcellular localizations,"*MolecularGenetics andGenomics*, vol. 289, no. 2, pp. 137–147, 2014.

[41] X.Wang, W. E. Grus, and J. Zhang, "Gene losses during human origins," *PLoS Biology*, vol. 4, no. 3, article e52, 2006.
[42] J. Zhu, J. Z. Sanborn, M. Diekhans,C. B. Lowe, T. H. Pringle, and D. Haussler, "Comparative genomics search for losses of longestablished genes on the human lineage," *PLoS Computational Biology*, vol. 3, no. 12, e247, 2007.
[43] B. Yngvadottir, Y. Xue, S. Searle et al., "A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 224–234, 2009.
[44] W. H. Chen, K. Trachana, M. J. Lercher, and P. Bork, "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," *Molecular Biology and Evolution*, vol. 29, no. 7, pp. 1703–1706, 2012.
[45] B. Liao and J. Zhang, "Null mutations in human and mouse orthologs frequently result in different phenotypes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, pp. 6987–6992, 2008.
[46] X. Gu, "Evolutionary framework for protein sequence evolution and gene pleiotropy," *Genetics*, vol. 175, no. 4, pp. 1813–1822, 2007.
[47] J. F.Wendel, "Genome evolution in polyploids," *PlantMolecular Biology*, vol. 42, no. 1, pp. 225–249, 2000.
[48] Z. Zhang, N. Carriero, and M. Gerstein, "Comparative analysis of processed pseudogenes in the mouse and human genomes," *Trends in Genetics*, vol. 20, no. 2, pp. 62–67, 2004.
[49] J. T. Eppig, C. J. Bult, J. A. Kadin et al., "The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D471–D475, 2005.
[50] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.
[51] Z. Gu, A. Cavalcanti, F. C. Chen, P. Bouman, and W. H. Li, "Extent of gene duplication in the genomes of Drosophila, nematode, and yeast," *Molecular Biology and Evolution*, vol. 19, no. 3, pp. 256–262, 2002
[52] S. F.Altschul,W. Gish,W.Miller, E.W. Myers, andD. J. Lipman, "Basic local alignment search tool," *Journal ofMolecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
[53] M. Remm, C. E.V. Storm, and E. L. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1041–1052, 2001.
[54] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

Unit-5

Evolution of Genetic Robustness after Duplication of an Essential or Dispensable Gene

Introduction

The role of functional compensation by duplicate genes has been examined in diverse organisms by comparing the proportion (P_E) of essential genes in duplicates to P_E in singletons (Wagner 2000; Gu et al. 2003; Conant and Wagner 2004; Hanada et al. 2009). Technically, a gene is called '*essential*' if the single-gene deletion phenotype is severe or lethal, or '*dispensable*' if its deletion phenotype is normal or nearly-normal (Ihmels et al. 2007; Hsiao and Vitkup 2008; Su et al. 2014; Kabir et al. 2017; Cacheiro et al. 2020). One may see Rancati et al. (2018) for a comprehensive review of gene essentiality. Intuitively, one may speculate that if duplicates play a significant role in functional compensation, the P_E for duplicates should be significantly lower than that of singletons. In other words, duplicate genes have major contributions to the genetic robustness at the organismal level. While this is indeed the case in yeasts, worms, and plants (Gu 2003; Kamath et al. 2003; Qian et al. 2010; Hanada et al. 2011), no significant difference in P_E was found between mouse single-copy and duplicate genes (Liang and Li 2007; Liao and Zhang 2007). A number of explanations were proposed (Li et al. 2010; Makino and McLysaght 2010; Vandersluis et al. 2010; Mendonça et al. 2011; Plata and Vitkup 2014; Zhang et al. 2015). Some studies showed that functional and protein connectivity bias between essential and dispensable duplicate genes may be the cause (Liang and Li 2009; Makino et al. 2009).

Though it is universal in almost all life forms, the pattern of duplicate compensation is complex (Szklarczyk et al. 2008; Hahn 2009; Chen et al. 2012; Keane et al. 2014; Saito et al. 2014; Diss et al. 2017; Teufel et al. 2018; Láruson et al. 2020; Mallik and Tawfik 2020). When an essential gene is duplicated (termed *ancestral essentiality*), duplicate compensation is the only mechanism to keep two resulting copies dispensable. On the other hand, when a dispensable gene is duplicated (termed *ancestral dispensability*), the ancient genetic buffering and duplicate compensation together keep both duplicate copies dispensable. Note that almost all previous P_E-related analyses in the literature did not distinguish between these two possibilities. Indeed, duplication of dispensable genes virtually result in no change of P_E, except for being essential duplicate by neo-functionalization. By contrast, after sufficiently long time, duplication of essential genes would be ultimately back to no change of P_E, except for long-term functional compensation maintained.

This paper will address this issue as follows. We first develop a statistical model to analyze duplicate pairs with three possible states: double-dispensable (DD), semi-dispensable (one dispensable one essential, DE) or double-essential (EE). With some biologically reasonable assumptions, a probabilistic model is then developed to estimate the proportion of essential genes duplicated from essential genes, and that from dispensable genes, respectively. Exemplified by the yeast and mouse duplicate pairs from their own whole genome duplications (WGD), respectively, some new insights about the evolutionary pattern of genetic robustness after gene duplication are discussed.

Genetic robustness between duplicate genes

A gene is called '*essential*' (denoted by d^-) if the single-gene deletion phenotype is severe or lethal, or '*dispensable*' (denoted by d^+) if its deletion phenotype is normal or nearly-normal (Ihmels et al. 2007; Hsiao and Vitkup 2008; Su et al. 2014; Kabir et al. 2017; Cacheiro et al. 2020). Consider two paralogous genes (*A* and *B*) duplicated from a common ancestor (*O*) *t* time units ago. There are four combined states, denoted by (d_A, d_B), representing double-dispensable (d^+,d^+), semi-dispensable (d^+,d^-) or (d^-,d^+), or double-essential (d^-,d^-), respectively.

We are interested in the derivation of $Q_t(d_A, d_B)$, the probability of any joint states (d_A, d_B) at time t since the duplication. To this end, one should distinguish between the duplication of an essential gene (ancestral essentiality, denoted by O^-) and the duplication of a dispensable gene (ancestral dispensability, denoted by O^+). Let $Q_t(d_A, d_B|O^-)$ be the probability of being (d_A, d_B) after t time units since gene duplication, conditional of the ancestral essentiality (O^-), and $Q_t(d_A, d_B|O^+)$ be the probability conditional of the ancestral dispensability (O^+). Since the ancestral state (dispensable or essential) for a duplicate pair is usually unknown, a mixture model is then implemented: let $R_0=P(O^+)$ be the probability of a gene pair duplicated from a dispensable gene, and $1-R_0=P(O^-)$ be that from an essential gene (Liang and Li 2007; Liao and Zhang 2007; Su and Gu 2008). Together, one can write

$$Q_t(d_A, d_B) = (1 - R_0)Q_t(d_A, d_B|O^-) + R_0 Q_t(d_A, d_B|O^+) \quad (1)$$

where $(d_A, d_B)= (d^+, d^+)$, (d^+, d^-), (d^-, d^+) or (d^-, d^-), respectively. Note that the process of non-functionalization of one duplicate copy was not conceptualized in the model under the assumption that the rate of non-functionalization was the same between dispensable and essential genes.

Table 1. A summary of mathematical notations and biological interpretations

Notation	Interpretation
d^+	State of '*dispensable*' if the single-gene deletion phenotype is normal
d^-	State of '*essential*' if the single-gene deletion phenotype is severe or lethal
O^+	Duplication of an dispensable gene (ancestral dispensability)
O^+*-duplicates*	Duplicates from ancestrally dispensable genes
O^-	Duplication of an essential gene (ancestral essentiality)
O^-*-duplicates*	Duplicates from ancestrally essential genes
$Q(d_A, d_B\|O^+)$	Probability of duplicates *A* and *B* being (d_A, d_B) after *t* time units since duplication, conditional of ancestral dispensability (O^+); $d_A, d_B= d^+$or d^-
$Q_t(d_A, d_B\|O^-)$	Probability of duplicates *A* and *B* being (d_A, d_B) after *t* time units since duplication, conditional of ancestral essentiality (O^-); $d_A, d_B= d^+$or d^-
$Q_t(d_A, d_B)$	Probability of duplicates *A* and *B* being (d_A, d_B) after *t* time units since duplication; $d_A, d_B= d^+$or d^-
R_0	Probability of a gene pair duplicated from a dispensable gene, i.e., $R_0=P(O^+)$
P_E	Proportion of essential genes in duplicates
$P_E(O^+)$	Proportion of essential genes in O^+-duplicates
$P_E(O^-)$	Proportion of essential genes in O^--duplicates

Duplication of essential gene: the sub-functionalization

When an essential gene was duplicated, the process of *sub-functionalization*, has been thought to be the major evolutionary mechanism for duplicate preservation (Force et al. 1999; Stoltzfus 1999; Prince and Pickett 2002; Innan and Kondrashov 2010; Stark et al. 2017). As a result, both duplicate copies can be preserved without invoking positive selection. Suppose a duplicate pair has *m* independent functional components, each of which is either 'active' (denoted by '1') or 'inactive' (denoted by '0'). Let U_{11} be the probability of a component being active in both genes; U_{01} (or U_{10}) is that of being inactive in gene *A* but active in gene *B* (or active in *A* but inactive in *B*); and U_{00} be the probability of a component being inactive in both genes. Without loss of generality, it is assumed that $U_{01}=U_{10}$. According to the *not-all-inactive constraint*, i.e., each component is functionally active at least in one duplicate copy, we claim U_{00}=0, leading to U_{11}=1-2*U* and $U_{10}=U_{01}=U$, respectively. That is, with a probability of *2U*, a functional component is active in one duplicate but inactive in another one, and with a probability of 1-2*U*, a component is active in both duplicates.

If these functional components of a gene are statistically independent and identical, $Q_t(d_A, d_B|O^-)$ can be derived in terms of the component parameter (*U*) and the number (*m*) of functional components, that is,

$$\begin{aligned} Q_t(d^+, d^+|O^-) &= (1-2U)^m \\ Q_t(d^+, d^-|O^-) &= (1-U)^m - (1-2U)^m \\ Q_t(d^-, d^+|O^-) &= (1-U)^m - (1-2U)^m \\ Q_t(d^-, d^-|O^-) &= 1-2(1-U)^m + (1-2U)^m \end{aligned} \quad (2)$$

The rationale of Eq.(2) is follows. Under the *m*-component model (*m*>1), two duplicate copies remain both dispensable only when each component is active in both duplicates (with a probability of 1-2*U*), which leads to the derivation of $Q_t(d^+, d^+|O^-)$ directly. Next we consider the (marginal) probability of dispensability (d^+) conditional of the ancestral essentiality (O^-), denoted by $Q_t(d^+|O^-)$. It appears that $Q_t(d^+|O^-)=(1-U)^m$ because the probability of a component for being active in one duplicate is given by (1-*U*). Since $Q_t(d^+|O^-)=Q_t(d^+, d^+|O^-)+Q_t(d^+, d^-|O^-)$, it is straightforward to obtain the second and third equations of Eq.(2). The last equation of Eq.(2) is derived by the sum of probabilities to be one.

Duplication of dispensable gene: the rare neo-functionalization

When a dispensable gene was duplicated, gene dispensability can be maintained through ancient genetic buffering and/or duplicate compensation (Prince and Pickett 2002; Innan and Kondrashov 2010; Stark et al. 2017). As a result, sub-functionalization becomes ineffective approach for the retention of duplicate genes, because the process of complementation between functional components (Force et al. 1999) is difficult to achieve. While the neo-functionalization has been suggested for the duplicate preservation in the case of genetic buffering (Chen et al. 2010; Vankuren and Long 2018; Lee and Szymanski 2021), it is unlikely that both copes acquire new functions simultaneously. In this sense one can assume that

$$Q_t(d^-, d^-|O^+) = 0 \quad (3)$$

This assumption holds well except for very ancient duplicates that may acquire new functions in the later stage.

Analysis of genetic robustness model between duplicates

Model formulation and estimation

Together with Eq.(2) and Eq.(3), the model of genetic robustness between duplicates formulated by Eq.(1) can be further specified as follows

$$\begin{aligned} Q_t(d^+, d^+) &= (1-R_0)(1-2U)^m + R_0\left[1-2Q(d^-|O^+)\right] \\ Q_t(d^+, d^-) &= (1-R_0)\left[(1-U)^m - (1-2U)^m\right] + R_0 Q(d^-|O^+) \\ Q_t(d^-, d^+) &= (1-R_0)\left[(1-U)^m - (1-2U)^m\right] + R_0 Q(d^-|O^+) \\ Q_t(d^-, d^-) &= (1-R_0)\left[1-2(1-U)^m + (1-2U)^m\right] \end{aligned} \quad (4)$$

where $Q(d^-|O^+)$ is the probability of an O^+-duplicate being essential (d^-); under Eq.(3), one can show $Q(d^-|O^+)=Q(d^-,d^-|O^+)+Q(d^-,d^+|O^+)=Q(d^-,d^+|O^+)$. Note that there are four unknown parameters, R_0, U, m and $Q(d^-|O^+)$ in two independent equations. A practically feasible approach is therefore implemented to solve this difficulty, as shown below.

(i) Suppose we have a set (*N*) of duplicate pairs; all *2N* genes have single-gene deletion phenotypes (dispensable or essential). Three types of duplicate pairs are considered, that is, *DD* for (d^+,d^+), *DE* for (d^+, d^-) or (d^-, d^+), and *EE* for (d^-, d^-), and their frequencies are denoted by f_{DD}, f_{DE} and f_{EE}, respectively.

(ii) R_0, the (prior) probability of a gene being dispensable before gene duplication can be replaced by the proportion of single-copy dispensable genes in the current genome as a proxy, under the assumption that R_0 remained a rough constant during the long-term evolution.

(iii) The parameter U can be estimated by replacing $Q_t(d^-, d^-)$ in the last equation of Eq.(4) by f_{EE}, that is,

$$1 - 2\left(1 - \hat{U}\right)^m + \left(1 - 2\hat{U}\right)^m = \frac{f_{EE}}{1 - R_0} \tag{5}$$

where m, the number of functional components, is treated as a known integer, i.e., m=2, 3,….

(iv) The proportion of essential O^--duplicates, i.e., those duplicated from essential genes, is given by $Q(d^-|O^-)=Q(d^-,d^-|O^-)+Q(d^-,d^+|O^-)$. When U is estimated by Eq.(5) (for any fixed m), according to Eq.(2) one can estimate $Q(d^-|O^-)$ by

$$\hat{Q}(d^-|O^-) = 1 - \left(1 - \hat{U}\right)^m \tag{6}$$

(v) After replacing $Q_t(d^+, d^+)$ in the first equation of Eq.(4) by f_{DD}, one can show that the proportion of essential O^+-duplicates can be estimated by

$$\hat{Q}(d^-|O^+) = \frac{1}{2} - \frac{f_{DD} - (1 - R_0)(1 - 2\hat{U})^m}{2R_0} \tag{7}$$

In short, from the observed frequencies f_{DD}, f_{DE} and f_{EE} with two degrees of freedom, we attempt to estimate two parameters $Q(d^-|O^-)$ and $Q(d^-|O^+)$ by Eqs.(5-7). To this end, we use the proportion of single-copy dispensable genes in the current genome as a proxy of R_0, and m as a constant that may only affect our estimation marginally.

Statistical evaluation

The statistical property of two estimates, $Q(d^-|O^-)$ and $Q(d^-|O^+)$, can be evaluated by two approaches. First, their large-sample variances can be obtained by the delta-method under a multinomial model of f_{DD}, f_{DE} and f_{EE}. The analytical formulas can be approximately obtained though the algebra was tedious (not shown). Second, a bootstrapping approach was implemented to empirically determine the sampling variance, as well as the confidence internals of these estimates.

Effect of the number of functional components (m)

By computer simulations, we examined how the number (m) of (canonical) functional components may affect our analysis. Note that the model of sub-functionalization requires at least two functional components. Hughes and Liberles (2007) suggested that between m=2 and m=12 regulatory regions would be biologically realistic. Our main results are follows: (*i*) the estimate of $Q(d^-|O^-)$ tends to decrease slightly when m is increased from 2 to 5 (about 20%), whereas that of $Q(d^-|O^+)$ tends to increase slightly; (*ii*) in both cases little difference was observed for m=5 or more; and (*iii*) all estimates are virtually the same from m=7 to m=∞. In short, it seems that the effect of variable m is negligible as long as it is reasonably large, say, m=5 or more.

Prediction of joint conditional probabilities

In practice it is desirable to know two types of conditional probabilities, $Q_t(d_A, d_B|O^-)$ and $Q_t(d_A, d_B|O^+)$, from the observed frequencies f_{DD}, f_{DE} and f_{EE}. According to Eq.(2), it is straightforward to calculate the conditional probabilities of (d_A, d_B) after duplication of an essential gene (O^-) as

$$\begin{aligned}
\hat{Q}_t(d^+, d^+|O^-) &= (1-2\hat{U})^m \\
\hat{Q}_t(d^+, d^-|O^-) &= (1-\hat{U})^m - (1-2\hat{U})^m \\
\hat{Q}_t(d^-, d^+|O^-) &= (1-\hat{U})^m - (1-2\hat{U})^m \\
\hat{Q}_t(d^-, d^-|O^-) &= \frac{f_{EE}}{1-R_0}
\end{aligned} \quad (8)$$

where U is the positive solution of Eq.(5). Next, one can predict $Q(d^+,d^+|O^+)$ by equating $Q_t(d^+, d^+)$ with f_{DD} in Eq.(1) in the case of $d_A=d^+$ and $d_B=d^+$, and replacing $Q(d^+,d^+|O^-)$ by its prediction given by the first equation of Eq.(8). They are, respectively, given by

$$\begin{aligned}
\hat{Q}_t(d^+, d^+|O^+) &= \frac{f_{DD} - (1-R_0)(1-2\hat{U})^m}{R_0} \\
2\hat{Q}_t(d^+, d^-|O^+) &= 1 - \frac{f_{DD} - (1-R_0)(1-2\hat{U})^m}{R_0} \\
\hat{Q}_t(d^-, d^-|O^+) &= 0
\end{aligned} \quad (9)$$

As indicated before, for a set of duplicate pairs with observed f_{DD}, f_{DE} and f_{EE}, there are only two degrees of freedoms. Hence, the statistical procedure described above treated R_0 and m as known constants and then estimated U and $Q(d^-|O^+)$. In this sense, Eq.(8) and Eq.(9) are not statistically well-justified to be treated as 'estimates'; instead, they should be regarded as predicted values.

Case study: duplicate pairs from the whole genome duplication (WGD) in yeast or mouse

Data availability

In total 325 yeast duplicate pairs were collected, which were from the yeast WGD about 100 million years ago (Kim and Yi 2006; Guan et al. 2007; Musso et al. 2008). According to the yeast single-gene deletion genomics, it can be further grouped into lethal, the strong effect, the moderate effect and the very weak effect (Gu et al. 2003). From the evolutionary view, a yeast gene is then classified as d^+ if it belongs to the very weak-effect group, or d^- otherwise. Under this classification, the proportion of dispensable single-copy genes (0.605) from Gu et al. (2003) is used as a proxy of R_0.

The second dataset includes 217 mouse duplicate pairs from the WGD occurred (Makino and McLysaght 2010), about 600 million years ago (in the early stage of vertebrates). Each pair were assigned by the mouse knockout phenotypes as follows (Su and Gu 2008): an essential gene was defined as a gene whose knockout phenotype is annotated as lethality (including embryonic, prenatal and postnatal lethality) or infertility.

Analysis

Our analysis is focused on three variables: (*i*) P_E is the observed proportion of essential duplicates; (*ii*) $P_E(O^-)$ is the expected proportion of essential O^--duplicates, i.e., those duplicated from essential genes, as estimated by $\hat{Q}(d^-|O^-)$ in Eq,(6); and (*iii*) $P_E(O^+)$ is the expected proportion of essential O^+-duplicates, i.e., those duplicated from dispensable genes, as estimated by $\hat{Q}(d^-|O^+)$ in Eq,(7). Their relationship is simply given by

$$P_E = (1-R_0)P_E(O^-) + R_0 P_E(O^+) \quad (10)$$

The frequencies of duplicate pairs with DD (double-dispensable), DE (dispensable-essential) and EE (double-essential) are presented in Fig.1(A) (yeast) and Fig.2(A) (mouse), respectively. While there is no empirical information about the number of functional components (m) for mouse and yeast genes, the robustness of the following analysis against various m's is important. Consistent with the simulation result, our analysis was generally

not affected by *m* (the number of functional components); overall it revealed little difference among those cases of *m*=3 or more. Our analysis of yeast WGD duplicate pairs is shown in Fig.1(B), and that of mouse in Fig.2(B) (*m*=6). Roughly speaking, yeast WGD pairs represent the case of recent WGD event, whereas mouse WGD pairs represent the ancient one.

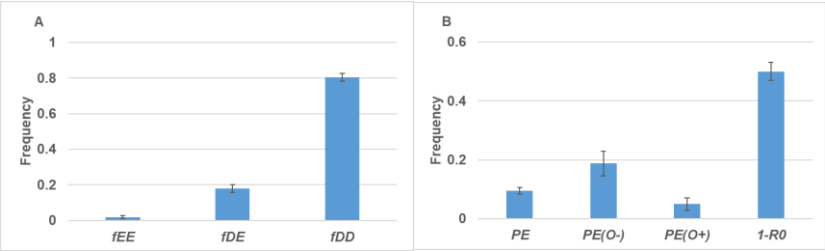

Fig.1. Analysis of yeast 325 WGD pairs. (A) Frequencies of duplicate pairs with DD (double-dispensable), DE (dispensable-essential) and EE (double-essential) are presented. (B) The proportion of essential duplicates (P_E), the estimated P_E in O^--duplicates (duplication of essential genes), $P_E(O^-)$, and the estimated P_E in O^+-duplicates (duplication of dispensable genes), $P_E(O^+)$, are presented. In the analysis, the number of functional components is set to be m=6. For comparison, the proportion of essential genes in single-copy genes (1-R_0) is also presented.

In the case of yeast WGD pairs, the proportion of essential duplicates (P_E=10.3%) is significantly larger than zero (p-value<10^{-6}), yet it is much lower than that of single-copy yeast genes ($P_{E,sin}$=0.395). The new analysis showed that the P_E in O^--duplicates (duplication of essential genes) was $P_E(O^-)$=21.2%, significantly greater than zero (p<0.001), whereas P_E in O^+-duplicates (duplication of dispensable genes) is $P_E(O^+)$=3.0% that was not significant (p>0.05). As expected, f_{EE} (the proportion of double-essential duplicate pairs) is so small that the estimation of U is subject to a large sampling variance. At any rate, one should be cautious to draw any conclusion based on a non-significant result. Nevertheless, it appears that the increase of P_E in the yeast WGD was mainly due to O^--duplicates, those duplicated from dispensable genes. Since the duplication time is the same for all duplicate pairs, one may predict that the rate of essentiality in O^--duplicates through sub-functionalization is about as 7-fold (21.2/3.0) as that in O^+-duplicates through neo-functionalization.

In the case of mouse WGD pairs representing an ancient WGD, we observed P_E=62.2%, virtually the same as P_E in single-copy genes (Liang and Li 2007; Liao and Zhang 2007; Su and Gu 2008). As expected, the estimate of $P_E(O^-)$=86.0% indicated that the majority of O^--duplicates in mouse WGD pairs, i.e., those duplicated from essential genes, may have become essential. Interestingly, the estimate of $P_E(O^+)$=28.4% was significantly greater than zero (p<0.001). Indeed, a nontrivial portion of O^+-duplicates in mice, i.e., those duplicated from dispensable genes, may be essential, which were subjected to neo-functionalization after the gene duplication (Chen et al. 2010; Vankuren and Long 2018; Lee and Szymanski 2021).

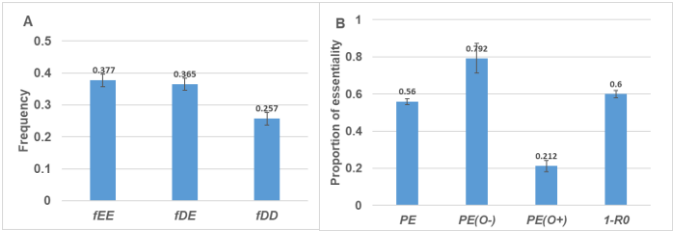

Fig. 2. Analysis of mouse 217 WGD pairs. (A) Frequencies of duplicate pairs with DD (double-dispensable), DE (dispensable-essential) and EE (double-essential) are presented. (B) The proportion of essential duplicates (P_E), the estimated P_E in O^--duplicates (duplication of essential genes), $P_E(O^-)$, and the estimated P_E in O^+-duplicates (duplication of dispensable genes), $P_E(O^+)$, are presented. In the analysis, the number of functional components is set to be m=6. For comparison, the proportion of essential genes in single-copy genes ($1\text{-}R_0$) is also presented.

We observed that, strikingly, $P_E(O^-)>P_E(O^+)$ significantly in both WGD duplicate pairs ($p<0.005$), which can be tentatively interpreted as follows: after the occurrence of WGD, the proportion of essential duplicates (P_E) increases with time t (the same for all duplicate pairs) through two distinct evolutionary routes: a fast process of essentiality in O^--duplicates through sub-functionalization, and a slow process of essentiality in O^+-duplicates through neo-functionalization; the difference is about 3-fold (86/28.4). Finally, Fig.3 shows the predicted conditional probabilities of yeast duplicate pairs: indeed, only marginal differences appeared when m=2, and all estimates were virtually the same between m=5 and m=∞. It was therefore concluded that the effect of variable m is usually negligible.

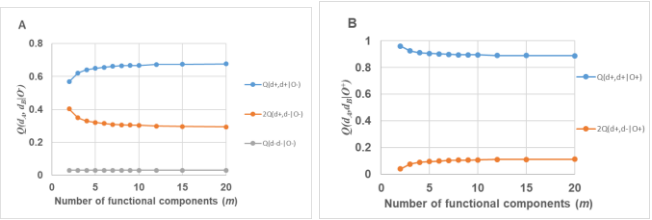

Fig.3. Predicted conditional probabilities of yeast WGD duplicate pairs plotting against the number of functional components m=2,…, 20. (A) $Q(d_A, d_B|O^-)$, probabilities conditional of ancestral essentiality (O^-). (B) $Q(d_A, d_B|O^+)$, probabilities conditional of ancestral dispensability (O^+).

Technical comments

The current model for the evolution of genetic robustness is certainly oversimplified. Due to different gene-silence/knockout technologies, the criteria to determine gene essentiality or dispensability are usually not comparable between species such as yeasts and mice. The concept of gene essentiality is therefore theoretical, depending on different experimental conditions, it has been used as the first-order proxy to study the evolutionary pattern of genetic robustness: how an organism is resilient against the occurrence of null mutations. In yeast, Hillenmeyer et al. (2008) found that 97% of gene deletions exhibited a measurable growth phenotype, suggesting that nearly all genes are essential for optimal growth in at least one condition. Hence, the model of genetic robustness actually depends on a cutoff of fitness effect under a given environmental condition (Nowak et al. 1997; Visser et al. 2003; Flatt 2005). Indeed, dispensable genes in our case studies (yeast or mouse) should be interpreted

as ‘nearly-dispensable’ under ideal experimental conditions, whereas essential genes are likely to be truly ‘essential’ under the wild condition. One may speculate that natural selection may act on those dispensable genes that are only ‘essential’ under certain conditions.

When an essential gene was duplicated, the current model assumed that two duplicate copies evolved under sub-functionalization, neglecting other possibilities such as neo-functionalization. Each functional component is assumed to undergo sub-functionalization independently, which is not biologically realistic. Meanwhile, after the duplication of a dispensable gene, interactions between ancestral genetic buffering, duplicate compensation and neo-functionalization remain largely unknown. In addition, some attributes of genetic mechanisms have not been taken into accounts, such as the effect of dosage balance, or the later-stage functional divergence (Prince and Pickett 2002; Innan and Kondrashov 2010). For instance, a high dosage requirement for a duplicated gene pair could result in both being essential (since loss of expression from either copy would bring the expression below the required threshold). In particular, for WGD-produced duplicates, some evidence showed that much of the duplicate preservation is due to the need of dosage balance (Birchler and Veitia 2012). Indeed, duplicate genes that are subject to dosage selection and constraint tends to be essential, raising an important question how much the estimated neo-functionalization in mouse WGD pairs is actually due to the dosage constraints. Our future study will focus on the development of a more realistic model of gene duplication.

A key assumption in our analyses is Eq.(3), that is, after duplication of a dispensable gene (O^+), the chance for both duplicate copies to be essential is negligible. While it is biologically intuitive, it may cause some bias, especially for some very ancient duplicate pairs. We conducted a simulation study to examine this effect by letting $Q(d^-,d^+|O^+)=q$, where q is a small positive value. Our preliminary result showed that the estimation bias was usually marginal, except for an extremely long evolutionary span after gene duplication (not shown). In addition, the current model does not consider the neo-functionalization after the duplication of an essential gene if the acquired new function would not impair the current functions. Nevertheless, the neo-functionalization after sub-functionalization, or sub-neo-functionalization for short, would not change the status of essentiality.

Dean et al. (2008) demonstrated that yeast duplicated genes can maintain substantial redundancy for extensive periods of time following duplication (over 100 million years). In another study, Vavouri et al. (2008) showed genetic redundancy was not just a transient consequence of gene duplication, but is often an evolutionary stable state; that is why some genes have retained redundant functions since the divergence of the animal, plant and fungi kingdoms (Gu 1997). Though the current study supported the basic idea provided by Vavouri et al. (2008) and Dean et al. (2008), a more careful analysis is required to clarify the difference in the evolutionary time-scale.

Further-reading

Gu, X (2022) A Simple Evolutionary Model of Genetic Robustness After Gene Duplication. Journal of Molecular Evolution. https://doi.org/10.1007/s00239-022-10065-1

References

Birchler JA, Veitia RA (2012) Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proc. Natl. Acad. Sci. U. S. A. 109

Brown SDM, Holmes CC, Mallon AM, et al (2018) High-throughput mouse phenomics for characterizing mammalian gene function. Nat. Rev. Genet. 19

Cacheiro P, Muñoz-Fuentes V, Murray SA, et al (2020) Human and mouse essentiality screens as a resource for disease gene discovery. Nat Commun 11:. https://doi.org/10.1038/s41467-020-14284-2

Chen S, Zhang YE, Long M (2010) New genes in Drosophila quickly become essential. Science (80-) 330:. https://doi.org/10.1126/science.1196380

Chen WH, Lu G, Chen X, et al (2017) OGEE v2: An update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res 45:. https://doi.org/10.1093/nar/gkw1013

Chen WH, 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:. https://doi.org/10.1093/molbev/mss014

Conant GC, Wagner A (2004) Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc R Soc B Biol Sci 271:. https://doi.org/10.1098/rspb.2003.2560

Coop G, Ralph P (2012) Patterns of neutral diversity under general models of selective sweeps. Genetics 192:205–224. https://doi.org/10.1534/genetics.112.141861

de Kegel B, Ryan CJ (2019) Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLoS Genet 15:. https://doi.org/10.1371/journal.pgen.1008466

Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4:. https://doi.org/10.1371/journal.pgen.1000113

Diss G, Gagnon-Arsenault I, Dion-Coté AM, et al (2017) Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. Science (80-) 355:. https://doi.org/10.1126/science.aai7685

Flatt T (2005) The evolutionary genetics of canalization. Q. Rev. Biol. 80

Force A, Lynch M, Pickett FB, et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:. https://doi.org/10.1093/genetics/151.4.1531

Fuller ZL, Berg JJ, Mostafavi H, et al (2019) Measuring intolerance to mutation in human genetics. Nat. Genet. 51

Georgi B, Voight BF, Bućan M (2013) From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. PLoS Genet 9:. https://doi.org/10.1371/journal.pgen.1003484

Gu Z, Steinmetz LM, Gu X, et al (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:. https://doi.org/10.1038/nature01198

Guan Y, Dunham MJ, Troyanskaya OG (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics 175:. https://doi.org/10.1534/genetics.106.064329

Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J. Hered. 100

Hanada K, Kuromori T, Myouga F, et al (2009) Evolutionary Persistence of Functional Compensation by Duplicate Genes in Arabidopsis. Genome Biol Evol 1:. https://doi.org/10.1093/gbe/evp043

Hanada K, Sawada Y, Kuromori T, et al (2011) Functional compensation of primary and secondary metabolites by duplicate genes in Arabidopsis thaliana. Mol Biol Evol 28:377–382. https://doi.org/10.1093/molbev/msq204

Hillenmeyer ME, Fung E, Wildenhain J, et al (2008) The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. Science (80-) 320:. https://doi.org/10.1126/science.1150021

Hsiao T-L, Vitkup D (2008) Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 4:e1000014. https://doi.org/10.1371/journal.pgen.1000014

Hughes T, Liberles DA (2007) The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol 65:. https://doi.org/10.1007/s00239-007-9041-9

Ihmels J, Collins SR, Schuldiner M, et al (2007) Backup without redundancy: Genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3:. https://doi.org/10.1038/msb4100127

Innan H, Kondrashov F (2010) The evolution of gene duplications: Classifying and distinguishing between models. Nat. Rev. Genet. 11

Kabir M, Barradas A, Tzotzos GT, et al (2017) Properties of genes essential for mouse development. PLoS One 12:. https://doi.org/10.1371/journal.pone.0178273

Kamath RS, Fraser AG, Dong Y, et al (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:. https://doi.org/10.1038/nature01278

Keane OM, Toft C, Carretero-Paulet L, et al (2014) Preservation of genetic and regulatory robustness in ancient gene duplicates of Saccharomyces cerevisiae. Genome Res 24:. https://doi.org/10.1101/gr.176792.114

Kim SH, Yi S V. (2006) Correlated asymmetry of sequence and functional divergence between duplicate proteins of Saccharomyces cerevisiae. Mol Biol Evol 23:. https://doi.org/10.1093/molbev/msj115

Láruson ÁJ, Yeaman S, Lotterhos KE (2020) The Importance of Genetic Redundancy in Evolution. Trends Ecol. Evol. 35

Lee Y, Szymanski DB (2021) Multimerization variants as potential drivers of neofunctionalization. Sci Adv 7:. https://doi.org/10.1126/sciadv.abf0984

Li J, Yuan Z, Zhang Z (2010) The cellular robustness by genetic redundancy in budding yeast. PLoS Genet 6:. https://doi.org/10.1371/journal.pgen.1001187

Liang H, Li WH (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 23

Liang H, Li WH (2009) Functional compensation by duplicated genes in mouse. Trends Genet. 25

Liao BY, Zhang J (2007) Mouse duplicate genes are as essential as singletons. Trends Genet. 23

Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. Trends Genet. 25

Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A 107:. https://doi.org/10.1073/pnas.0914697107

Mallik S, Tawfik DS (2020) Determining the interaction status and evolutionary fate of duplicated homomeric proteins. PLoS Comput Biol 16:. https://doi.org/10.1371/journal.pcbi.1008145

Mendonça AG, Alves RJ, Pereira-Leal JB (2011) Loss of genetic redundancy in reductive genome evolution. PLoS Comput Biol 7:. https://doi.org/10.1371/journal.pcbi.1001082

Muñoz-fuentes V, Cacheiro P, Meehan TF, et al (2018) The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. Conserv Genet 19:. https://doi.org/10.1007/s10592-018-1072-9

Musso G, Costanzo M, Huangfu MQ, et al (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res 18:. https://doi.org/10.1101/gr.076174.108

Nowak MA, Boerlijst MC, Cooke J, Smith JM (1997) Evolution of genetic redundancy. Nature 388:. https://doi.org/10.1038/40618

Pengelly RJ, Vergara-Lope A, Alyousfi D, et al (2019) Understanding the disease genome: Gene essentiality and the interplay of selection, recombination and mutation. Brief Bioinform 20:. https://doi.org/10.1093/bib/bbx110

Plata G, Vitkup D (2014) Genetic robustness and functional evolution of gene duplicates. Nucleic Acids Res 42:. https://doi.org/10.1093/nar/gkt1200

Prince VE, Pickett FB (2002) Splitting pairs: The diverging fates of duplicated genes. Nat. Rev. Genet. 3

Qian W, Liao BY, Chang AYF, Zhang J (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. 26

Rancati G, Moffat J, Typas A, Pavelka N (2018) Emerging and evolving concepts in gene essentiality. Nat. Rev. Genet. 19

Saito N, Ishihara S, Kaneko K (2014) Evolution of genetic redundancy: The relevance of complexity in genotype-phenotype mapping. New J Phys 16:. https://doi.org/10.1088/1367-2630/16/6/063013

Smith CL, Blake JA, Kadin JA, et al (2018) Mouse Genome Database (MGD)-2018: Knowledgebase for the laboratory mouse. Nucleic Acids Res 46:. https://doi.org/10.1093/nar/gkx1006

Stark TL, Liberles DA, Holland BR, O'Reilly MM (2017) Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization. BMC Evol Biol 17:. https://doi.org/10.1186/s12862-016-0848-0

Stoltzfus A (1999) On the possibility of constructive neutral evolution. J Mol Evol 49:. https://doi.org/10.1007/PL00006540

Su Z, Gu X (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. J Mol Evol 67:. https://doi.org/10.1007/s00239-008-9170-9

Su Z, Wang J, Gu X (2014) Effect of duplicate genes on mouse genetic robustness: An update. Biomed Res Int 2014:. https://doi.org/10.1155/2014/758672

Szklarczyk R, Huynen MA, Snel B (2008) Complex fate of paralogs. BMC Evol Biol 8:. https://doi.org/10.1186/1471-2148-8-337

Teufel AI, Johnson MM, Laurent JM, et al (2018) Withdrawn as Duplicate: The many nuanced evolutionary consequences of duplicated genes. Mol Biol Evol 35:e1. https://doi.org/10.1093/molbev/msy216

Vandersluis B, Bellay J, Musso G, et al (2010) Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol Syst Biol 6:. https://doi.org/10.1038/msb.2010.82

Vankuren NW, Long M (2018) Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nat Ecol Evol 2:. https://doi.org/10.1038/s41559-018-0471-0

Vavouri T, Semple JI, Lehner B (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet. 24

Visser JAGM, Hermisson J, Wagner GP, et al (2003) PERSPECTIVE: EVOLUTION AND DETECTION OF GENETIC ROBUSTNESS. Evolution (N Y) 57:. https://doi.org/10.1111/j.0014-3820.2003.tb00377.x

Wagner A (2000) Robustness against mutations in genetic networks of yeast. Nat Genet 24:. https://doi.org/10.1038/74174

Wang T, Birsoy K, Hughes NW, et al (2015) Identification and characterization of essential genes in the human genome. Science (80-) 350:. https://doi.org/10.1126/science.aac7041

Zhang W, Landback P, Gschwend AR, et al (2015) New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol 16:. https://doi.org/10.1186/s13059-015-0772-4