Research Focus

Evolution of duplicate genes versus genetic robustness against null mutations

Xun Gu

Department of Genetics, Development and Cell Biology; Centre for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA

There are two proposed mechanisms for the emergence of gene network robustness: (1) 'genetic buffering' from redundant gene networks (i.e. alternative metabolic or regulatory or signal pathways), and (2) functional complementation from duplicate genes. Their relative significance is a subject to debate, but recent studies in functional genomics provide some interesting insights. In particular, experiments in yeast using whole-genome libraries of single-gene-deletion mutants and worm whole-genome RNAi show that both mechanisms are important for the genetic robustness. Yet, many questions remain.

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 [4–9].

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. Later, Kitami and Nadeau [5] made a similar inference based solely on sequence analysis. They claimed that (1) genes with redundant or alternative metabolic pathways evolved more quickly than did genes without redundant networks, and (2) no significant difference in evolutionary rate was detected between single-copy genes and duplicate genes. Unfortunately, it has since been found that the first result was false, caused by a computational error [5]. Consequently, their analysis turns out to be inconclusive.

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

Corresponding author: Xun Gu (xgu@iastate.edu).

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.

Future experiments

These whole-genome approaches that targeting single genes in yeast [7] and *C. elegans* [9] are only the first-order approaches to investigating the pattern of genetic robustness. The limitation is that gene–gene interactions cannot be measured directly from these datasets. One (naïve) solution is to generate whole-genome libraries of yeast multi-gene deletion strains or *C. elegans* multi-gene RNAi clones. The obstacle is the magnitude of experiments. In the case of yeast genome with ~6000 genes, for instance, ~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. For example:

(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 $P(k) : k$ (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.

(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.

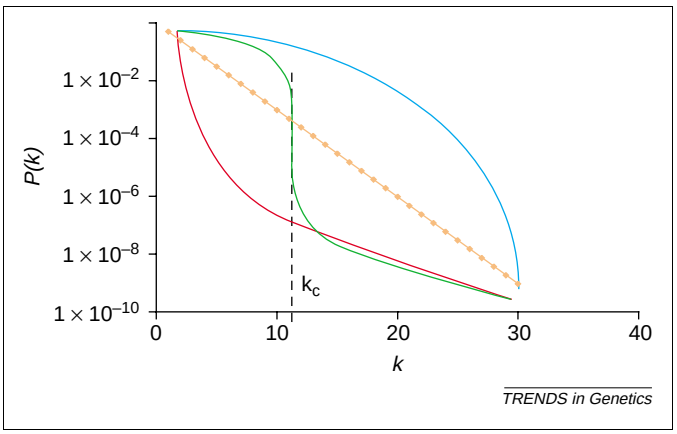

Fig. 1. The hypothetical semi-log plotting for $P(k) : k$; where $P(k)$ is the probability of having phenotype when a random k genes are deleted from the genome. The orange curve is expected by an independent model that the pattern of robustness is determined by individual gene effect (i.e. $P(k) = q^k$). 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.

Conclusions

In summary, recent genome-wide studies [6–9] 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, although the existence of a weak correlation is still controversial [12,13]. Historically, it is interesting to note that in 1969 M. Nei [14] wrote, '… [T]here are likely to be many duplicate genes which have similar biological functions, and the function of one gene may be compensated by the other genes. … [T]he effect of deleterious mutations or the so-called mutation load may be greatly reduced, and they would be expected to interact synergistically.'

However, one fundamental problem, the evolutionary mechanism for the emergence of genetic robustness, remains largely unsolved. The following speculations might be worth further study. First, 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'

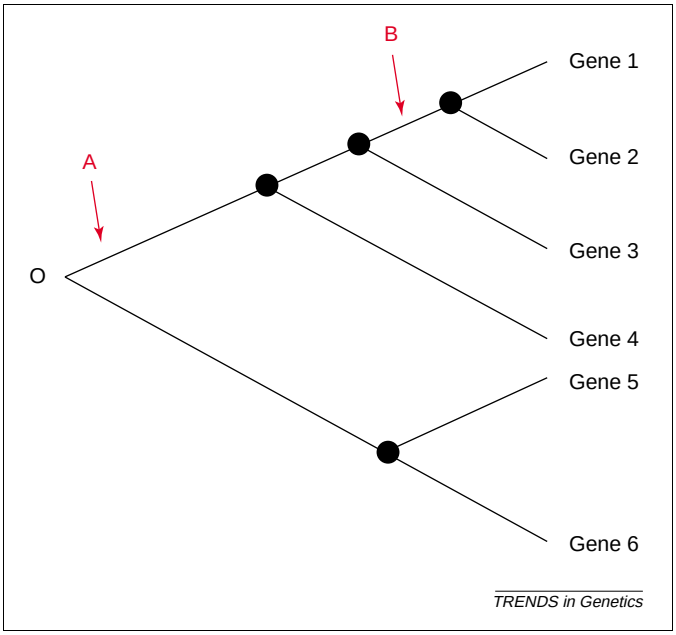

Fig. 2. Phylogeny-based design for loss-of-function phenotypes of a gene family: For partition A, the two gene groups are (1,2,3,4) and (5,6), respectively. Then, the gene deletion patterns are Δ1Δ2Δ3Δ4, and Δ5Δ6, respectively. For partition B, the two gene groups are (1,2) and (3,4,5,6), respectively. For an n-gene phylogeny, there are $2n - 3$ different partitions.

process could still exist (e.g. at the protein structure level), even if the sequence similarity is too little to be detected. This implies that a certain amount of genetic buffering originally comes from gene duplication [7].

Second, 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.

Third, 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.

Acknowledgements

The author thanks Wen-Hsiung Li and Zhenglong Gu for critical comments. The work is supported by the NIH grant.

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 Barabási, 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

0168-9525/03/$ - see front matter
doi:10.1016/S0168-9525(03)00139-2

Genome Analysis

Evolutionary diversification of mitochondrial proteomes: implications for human disease

Erik Richly[1], Patrick F. Chinnery[2] and Dario Leister[1]

[1]Abteilung für Pflanzenzüchtung und Ertragsphysiologie, Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné Weg 10, D-50829 Köln, Germany
[2]Neurology, The Medical School, Framlington Place, Newcastle upon Tyne, UK NE2 4HH

By combining comparative genomics and computational identification of protein targeting, we predicted the size and composition of the mitochondrial proteome for ten species. Functional mitochondria could harbour from a few hundred to more than 3000 gene products, and protein relocation from and to mitochondria occurred during evolution. Although each genome studied contains lineage-specific mitochondrial proteins, conserved

Corresponding author: Dario Leister (leister@mpiz-koeln.mpg.de).