

The Distribution of Single-Function Duplicate Genes Among Functional Groups

Gabriel A. Harp
Department of Biology
Indiana University, Bloomington

Introduction

Despite the plurality of mechanisms proposed to explain the origins of duplicated genes, the transient nature of the genome has made the ultimate causes of gene duplication difficult to determine. The origin of duplicated genes is an important evolutionary process from which substantial genetic variation may be generated. It is difficult, however, to distinguish between paralogs that result from whole-genome duplications followed by the individual loss of genes and those that arise from individual gene duplications (Brookfield, 2001). No single mechanism can explain the patterns of duplication observed for all genes, genomes, or organisms. Instead, residual patterns have been used to infer the mechanisms of duplication (e.g. Wolfe and Shields 1997; Vision et al. 2000), and substantial emphasis has been placed on determining the role of selection in the generation of new gene duplications (Brown et al. 1998; Riehle et al. 2001; Kondrashov et al. 2002; reviewed in Anderson and Roth 1977). An emerging model for the role of gene duplication characterizes the response to selection that new gene duplications and their effect on dosage provide to a genome under stress (Kondrashov et al. 2002; Wagner 2002). This study reports a non-random distribution of single-function duplicate genes that demonstrates the importance of genes with environmental stress-related functions in shaping the yeast genome.

A positive heuristic for the study of genome evolution that unifies sets of genes having a common feature with their functional annotations is developing and has been

described in Conant and Wagner (2002). In addition to the sets of duplicated genes offered by Conant and Wagner, other groups of genes could be compared against the null expectation that the distribution is random among functional categories. A cursory comparison using intron-containing genes from the *Saccharomyces cerevisiae* genome showed that many more intron-containing genes are present in duplicate than would be expected by chance. The majority of these genes belong to the protein synthesis functional class and code for ribosomal proteins, an observation consistent with previous analysis (Ares et al. 1999). Another opportunity would be to compare genes with standardized molecular pathway positions for patterns of functional assignment. Any set of genes with common features and functional annotations can be compared to a null distribution consistent with a random sample of the genome. The goal of this approach is to identify patterns in the distribution of genes throughout individual genomes and across multiple genomes and make useful predictions about the cause or causes of such patterns.

This approach has already provided at least one interesting correlation. Of the 25 hottest recombination sites in the yeast genome, 19 contained one or more ORFs in the metabolism functional class. This indicated a significant over-representation of hotspot genes in the metabolism functional class and led to the interpretation that certain functional categories are associated with a particular chromatin structure (Gerton et al. 2000). We discuss some difficulties when categorizing sets of genes by their functional annotation and present new results that are complementary to those of Conant and Wagner (2002). Specifically, we show a significant nonrandom distribution of gene duplicates among functional categories in *Saccharomyces cerevisiae*, and provide an explanation for these patterns.

Methods

A list of protein-coding duplicate gene pairs in *Saccharomyces cerevisiae* was graciously provided by Mike Lynch (Department of Biology, Indiana University) and developed as described in Lynch and Conery (2000). The identity of each open reading frame (ORF) was converted to its yeast systematic ID and the list was organized such that each gene was only listed once. Thus, our focus was on the property of each gene having been duplicated or serving as the template for duplicates.

We compared our list of duplicated genes to an entire listing of *Saccharomyces cerevisiae* ORFs in the Munich Information Center for Protein Sequences (MIPS) database of functional categories (FUNCAT) using a Perl-scripted (Wall et al. 1996), semi-automated approach. Genes that were present in our list and each of the node-1 functional categories were kept in the analysis, while genes that occurred in our list and multiple functional categories were excluded.

For our analysis, we used the MIPS FUNCAT assignments for *Saccharomyces cerevisiae* that grew out of the PEDANT database of functional assignments (Mewes et al. 1997). These assignments are based on similarity to known proteins, presence of indicative sequence patterns, and experimental data from the literature (Mewes et al. 1997). Thus, the functional assignments are both automatically and manually derived. An advantage of using the MIPS FUNCAT scheme is that it was specifically developed for the yeast genome as well as other unicellular eukaryotes. This does, however, limit its applicability to other organisms. The Gene Ontology (GO) scheme by comparison is designed to allow comparisons among many organisms. However, there exists an order of magnitude greater complexity within the GO scheme than other classification schemes

(Rison et al. 2000). There is no evidence that this increased complexity allows a greater understanding of the processes underlying genome evolution, particularly when only a single organism is considered.

Statistical Methods

Chi-square analysis was used to compare the distribution of duplicate genes among functional categories to its null expectation. The null expectation was derived from the distribution of all yeast genes in the MIPS FUNCAT. It is this null distribution that is critical to the interpretation of the data. Comparisons of results are limited when these null distributions show little similarity. The expected number of duplicated genes in each functional category was calculated using a ratio of genes for each category to the total number of genes in all categories multiplied by the total number of duplicated genes observed in all categories. Thus, the null distribution was scaled to the observed numbers of duplicates. After applying the Yates correction for continuity, Chi-square values for each category were compared with a Chi distribution with $df=1$. We used a significance threshold (the probability of making a type I error) of 5% that, after a Bonferroni correction for multiple comparisons ($n=18$) of functional categories, was reduced to 0.27%. Summing the Chi-squared values across all categories allowed us to test the hypothesis that the overall distribution of duplicate genes among categories was different from the null. The result was significant at $p<0.001$, $df=17$.

There were four categories with no observed duplicate genes and at least three others that had low estimations (<5) of the expected value. At least one, transposons, was systematically generated to have no observed duplicates as Lynch and Conery removed

these from the list of gene duplicates. We reached the conclusion that both k (number of categories) and n (number of genes in all categories) were sufficiently large that no bias would be introduced into the analysis despite the non-uniformity of the distribution among categories (Zar 1996). Thus, categories with zero observations and/or small expected values were kept in the analysis.

Results

Six categories showed a non-random distribution of duplicate genes: cellular transport, energy, metabolism, protein synthesis, rescue, defense, and virulence, and unclassified genes (Fig. 1). Cellular transport and unclassified categories had fewer than expected duplicate genes ($p < 0.0027$). Energy, metabolism, protein synthesis, rescue, defense, and virulence genes all exhibited a larger than expected proportion of duplicated genes ($p < 0.0027$). The observed distribution of duplicate genes among all functional categories was significantly different from the expected distribution ($p \ll 0.001$, $df=17$).

Discussion

There is a non-random distribution of duplicate genes with a single function among functional categories. That certain classes of genes are present in duplicate more often than others suggests that energy, metabolism, rescue, defense, and virulence-related genes have historically played a pivotal role in protein dosage regulation and the evolution of the *Saccharomyces cerevisiae* genome. These data support a model of gene duplication that posits the importance of environmental stress to the origin of novel genetic variation through new gene duplications (Kondrashov et al. 2002). This model is

underscored by results indicating that fitness gains are made through alterations in the regulation of central metabolism and functionally related genes (Ferea et al. 1999) and that distinct asymmetry in divergence in gene function is observed for the response to environmental stress (Wagner 2002). The non-random distribution of duplicate genes among functional groups may then be a pattern of the historical response of the highly domesticated *Saccharomyces cerevisiae* genome to environmental stress if 1) the response to stress is garnered through changes in the expression patterns of metabolically-related genes and 2) gene duplications are generated in a fitness-dependent manner (Agarwal in press) that increase the robustness of the genome to deleterious mutations (Wagner 2002). Experiments that focused on hexose-transport genes in yeast showed that after 450 generations of glucose-limited growth, gene duplication was observed at both regulatory and coding loci in cell types exhibiting enhanced fitness relative to their progenitors (Brown et al. 1998). In contrast to our results, transport genes have been cited as having a higher proportion of duplicates among functional groups (Kondrashov et al. 2002; Conant and Wagner 2002). While certain transport genes have been duplicated, a more striking pattern emerges in the overrepresentation of energy, metabolism, rescue, defense, and virulence genes. We conclude that these functions have a greater effect of dosage than transport genes, and that transport related genes are underrepresented in duplicate because of sequence-specific binding properties. The broad classifications of these genes are related to the interaction of the organism with its environment and suggest that environmental stress may be indirectly causing the origin of genetic novelty.

The unclassified genes that are present in fewer than expected numbers may reveal an artifact of the methods used to define gene function when duplicated genes provide experimenters with greater functional information

There are difficulties that make our approach distinct from others. In contrast to the methods taken by Conant and Wagner (2002), the list of pairs of gene duplicates was screened to eliminate those amino-acid sequences not starting with methionine. While the large number of duplicate genes generated by this process can be further reduced to only coding genes, the final analysis of genomic patterns does not focus on functional genes. This accounts for the large number of unique gene duplicates Conant and Wagner observed in the yeast genome (2088 in Conant and Wagner (2002) versus 397 in Lynch and Conery (2001)). The frequency of duplicated functional genes in yeast is generally thought to be smaller than many other organisms.

The most pressing concern is how to deal with genes that have been assigned to multiple functional categories. While it is reasonable to assign the most specific function to genes with multiple functions as an attempt to create statistical independence among the samples, it does not mitigate the problem of biological independence. It is without question that the underlying reason genes have been assigned to functional categories is that there exists some biological reason for them to be so defined. Many functional classifications are assigned based on a sequence pattern. In some genes there may be more than one functional motif, and thus in a very real sense those genes cannot be treated as independent data. It also stands to reason that selection has caused hitchhiking among of functions when genes with single functions become linked with divergent functions.

These duplicated genes with multiple functions represent a potentially interesting class of duplicate genes, and they should definitely be investigated. However, as a first approximation it was best to remain conservative in our approach and eliminate genes with two or greater functional annotations from the analysis. When genes with multiple functions were included in the analysis, ten categories showed a significant deviation from the expected number of duplicated genes (unpublished results). The number of categories with significant deviations decreased to six and the significance of each remaining category increased when genes with multiple functional annotations were removed from the analysis.

Genes in different functional categories may either not be equally likely to undergo duplication or not equally likely to be preserved after having undergone duplication. This distinction is important because the distribution of gene duplicates among functional categories depends on 1) the mechanism of duplication and 2) the selection regimes that fix duplicates in the population. The distribution of duplicate genes observed in the *Saccharomyces cerevisiae* genome unfortunately sheds no light on this distinction. Instead, the non-random distribution that has been observed indicates only that certain classes of duplicated genes persist in the genome. Thus, we can speculate that those classes that do show a disproportionate number of duplicates do so because they have had a historically beneficial effect on the fitness of the organism and that not all genes are created equal.

Literature Cited

- Anderson, R. P. and Roth, J. R. 1977. Tandem genetic duplications in phage and bacteria. *Annual Review of Microbiology* 31:473.
- Ares, M., Jr., Grate, L., Pauling, M. H. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* 5:1138-1139.
- Brookfield, J. F. Y. 2001. Genome Evolution. In *Handbook of Statistical Genetics*. D. J. Balding et al., eds. John Wiley and Sons, Ltd., Chichester, England.
- Brown, C. J., Todd, K. M., and Rosenzweig, R. F. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution* 15(8):931-942.
- Conant, G. C. and Wagner, A. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Research* 30(15):3378-3386.
- Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *PNAS* 96:9721-9726.
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. 2002. Selection in the evolution of gene duplications. *Genome Biology* 3:1-9.
- Lynch, M. and Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151-1155.
- Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maiertl, A., Oliver, S. G., Pfeiffer, F., and Zollner, A. 1997. Overview of the yeast genome. *Nature* 387:Supp. 7-8.
- Riehle, M. M., Bennett, A. F., and Long, A. D. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *PNAS* 98:525-530.
- Rison, S. C. G., Hodgman, T. C., and Thornton, J. M. 2000. Comparison of functional annotation schemes for genomes. *Functional and Integrative Genomics* 1:56-69.
- Vision, T. J., Brown, D. G., and Tanksley, S. D. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290:2114-2117.
- Wagner, A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Molecular Biology and Evolution* 19(10):1760-1768.
- Wall, L., Christiansen, T., Schwartz, R. L. 1996. *Programming Perl*, 2nd edn. O'Reilly, Sebastopol, Calif.
- Wolfe, K. H., and Shields, D. C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708-713.
- Zar, J. H. 1996. *Biostatistical Analysis*. Prentice Hall, Inc., Upper Saddle River, New Jersey, USA.

Fig. 1 The predicted and observed numbers of gene duplicates among 18 MIPS (Munich Information Center for Protein Sequences) functional categories in the *Saccharomyces cerevisiae* genome. Chi-squared analyses revealed a significant difference between the predicted and observed values for six categories ($p < 0.0027$, $df = 1$) and the overall distribution ($p \ll 0.001$, $df = 17$). Asterisks represent those categories that had a significantly larger deviation from the predicted distribution. The leftmost bar for each category is the predicted number of duplicated genes if redundancies between categories are left in the analysis. The middlemost and rightmost bars represent the predicted and observed numbers, respectively, of duplicated genes when all genes that reside in more than one functional category are removed from the analysis.

