CHAPTER

Correlations between Quantitative Measures of Genome Evolution, Expression and Function

Yuri I. Wolf, Liran Carmel and Eugene V. Koonin*

Abstract

In addition to multiple, complete genome sequences, genome-wide data on biological prop properties of genes, such as knockout effect, expression levels, protein-protein interactions, and others, are rapidly accumulating. Numerous attempts were made by many groups to examine connections between these properties and quantitative measures of gene evolution. The questions addressed pertain to the most fundamental aspects of biology: what determines the effect of the knockout of a given gene on the phenotype (in particular, is it essential or not) and the rate of a gene's evolution and how are the phenotypic properties and evolution connected? Many significant correlations were detected, e.g., positive correlation between the tendency of a gene to be lost during evolution and sequence evolution rate, and negative correlations between each of the above measures of evolutionary variability and expression level or the phenotypic effect of gene knockout. However, most of these correlations are relatively weak and explain a small fraction of the variation present in the data. We propose that the majority of the relationships between the phenotypic ("input") and evolutionary ("output") variables can be described with a single, composite variable, the gene's "social status in the genomic community", which reflects the biological role of the gene and its mode of evolution. "High-status" genes, involved in house-keeping processes, are more likely to be higher and broader expressed, to have more interaction partners, and to produce lethal or severely impaired knockout mutants. These genes also tend to evolve slower and are less prone to gene loss across various taxonomic groups. "Low-status" genes are expected to be weakly expressed, have fewer interaction partners, and exhibit narrower (and less coherent) phyletic distribution. On average, these genes evolve faster and are more often lost during evolution than high-status genes. The "gene status" notion may serve as a generator of null hypotheses regarding the connections between phenotypic and evolutionary parameters associated with genes. Any deviation from the expected pattern calls for attention—to the quality of the data, the nature of the analyzed relationship, or both.

Quantitative genomics involves numerous measures reflecting different aspects of the evolutionary history and the physiological role of a given gene (protein). One can estimate the evolution rate of a gene, measured in different organisms; its expression level in different tissues and in different taxonomic groups; the tendency of a gene to be lost during evolution of different lineages of organisms or its tendency to produce paralogous copies via duplication; its

*Eugene V. Koonin—National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, U.S.A.
Email: koonin@ncbi.nlm.nih.gov.

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber. ©2005 Eurekah.com.

Table 1. Connections between various measures of sequence evolution rate, gene loss, expression, and fitness effect[a]

	K_{aa}	K_N	K_S	K_5	K_3	PGL	E_H	B_H	E_C	E_Y	E_Y
1 protein evolution rate (K_{aa})	x										
2 CDS non-synonymous evolution rate (K_N)	+	x									
3 CDS synonymous evolution rate (K_S)		+	X								
4 5'-UTR evolution rate (K_5)		+	+	x							
5 3'-UTR evolution rate (K_3)		+	+	+	x						
6 propensity for gene loss (PGL)	+					x					
7 expression level in human (E_H)	-	-	-	0	-	-	x				
8 expression breadth in human (B_H)	-	-	-	0	-		+	x			
9 expression level in *C. elegans* (E_C)	-					-	+		x		
10 expression level in *S. cereviseae* (E_Y)	-					-	+		+	x	
11 viability of gene disruption in *S. cereviseae* (E_Y)	+					+	-	-	-	-	x

[a] The data was from [15, 16].

position in the metabolic, signaling and protein interaction networks; and a variety of other quantities (e.g., refs. 1-4). Not unexpectedly, many of such measures are not independent. The literature on the subject (see specific references below) reports numerous positive and negative correlations: between the synonymous and nonsynonymous evolution rates within a gene; between evolution rate and expression level; between propensity of gene loss and fitness effect; and many more (Table 1). Some of these correlations are very strong for quite obvious reasons, such as evolution rates in different lineages or expression levels of orthologous genes; others are less trivial, e.g., the correlation between the degree of conservation of a gene's presence in different lineages and the degree of conservation of its sequence; yet others are remarkably low or absent, sometimes running contrary to expectations (evolution rate *vs*. number of protein-protein interactions or conservation of gene sequence and that of expression profiles).

Diverse as they are, all these purported correlations, except for the most obvious ones, share one somewhat disturbing feature: although they may be highly statistically significant due to the large number of data points, they typically explain only a small fraction of the variance of the analyzed quantities. Hence considerable debate around many of these observations, which is further compounded by problems with the completeness and quality of much of the data involved, particularly that coming from genome-scale analyses of gene expression, protein-protein interactions, and other aspects of gene functioning. For example, the argument about the link—or lack thereof—between the connectivity of a protein in protein-protein interaction networks and its evolutionary rate has already gone through at least three cycles of opposing claims, and there is still no definitive solution in sight.[5-10] Even when the existence of a link is not seriously questioned, as is, e.g., the case with the negative correlation between a gene's expression level and sequence evolution rate, the nagging question remains as to the ultimate importance of these observations. Given that the nontrivial correlations, however statistically significant they might be, are all relatively weak, it is quite legitimate and, probably, prudent to ask whether one should emphasize the existence of a particular link or the fact that the effect of one of the analyzed variables on the other(s) is only modest. Answering these questions is not easy, and yet, they are pressing because the higher-level problems addressed in this area of research are, arguably, among the most fundamental ones in biology, e.g., what determines the fitness effect of a gene's knockout or the rate of its evolution.

Quantitative genomics is a very young discipline which started in earnest only at the brink of the 21st century, when genome-wide data beyond the sequences themselves (gene expression, protein-protein interaction etc) began to accumulate. Nevertheless, in these few years, a fairly complex maze of observations on connections—or lack thereof—between all kinds of quantities has emerged. We believe that the field is in rather urgent need of a coherent conceptual framework that would allow one, simply put, to make sense of these diverse and often contradictory bits and pieces of information. Here, we present a brief overview of the available results on genomic correlations and discuss some preliminary glimpses of a would-be synthesis.

Evolution Rate, Expression Level and Expression Breadth

Numerous reports, including our own research, point to a significant correlation between the measures of evolutionary conservation of a protein and measures of its expression[11-16] (Table 1). Several notable conclusions emerge from these analyses. Firstly, there is a strong cohesion between measures of the same quantity obtained for different, in many cases, phylogenetically distant species. Despite obvious biological differences between, e.g., mammals, nematodes, and yeasts, expression levels of orthologous proteins from these species display positive correlations with *r*-values of 0.3-0.5 (with many hundreds of proteins in the dataset, the correlations are significant at *p*-values $<<10^{-10}$).[15] Likewise, evolution rates estimated for different lineages and across different ranges of distances tend to show even greater concordance (*r*-values of 0.7-0.9 between distantly related bacterial lineages).[17] Secondly, expression breadth, defined as the number of different tissues where a gene is significantly expressed, and the connectivity (node degree) in the gene coexpression network behave in essentially the same way as the expression level in expression *vs.* evolution rate comparisons. Specifically, there is a highly significant negative correlation between each of these parameters of gene expression and sequence evolution rate; in other words, highly and widely expressed genes, which have numerous coexpression partners, tend to evolve slowly (Fig. 1). Finally, while this negative correlation between evolution rate and expression parameters holds for the great majority of the relevant data, a notable exception breaks this nearly universal pattern. Analysis of mammalian microarray data shows that, while the synonymous and non-synonymous nucleotide evolution rates within the coding sequence and the nucleotide sequence evolution rate in the 3'-UTR behave as expected, the evolution rate of the 5'-UTR shows no correlation with expression.[16] This apparent discrepancy probably points to a distinct mode of evolution and/or a specific and still not understood connection between the sequence and its expression for the 5'-UTRs of (at least) mammalian genes (Fig. 1).

Evolution Rate, Gene Loss and Fitness Effect

The statement that sequence evolution rate and the tendency of a gene to be lost during evolution are correlated at first glance seems almost trivial—after all, it should be expected that evolutionarily conserved (i.e., slowly evolving) genes are also phylogenetically conserved, i.e., their orthologs more densely populate the tree of life than those of fast-evolving genes. In support of this straightforward line of reasoning, a highly significant correlation between these parameters has been observed[15] (Table 1, Fig. 2). However, these two faces of evolutionary conservation are not linked directly via a cause and effect relationship. Most likely,

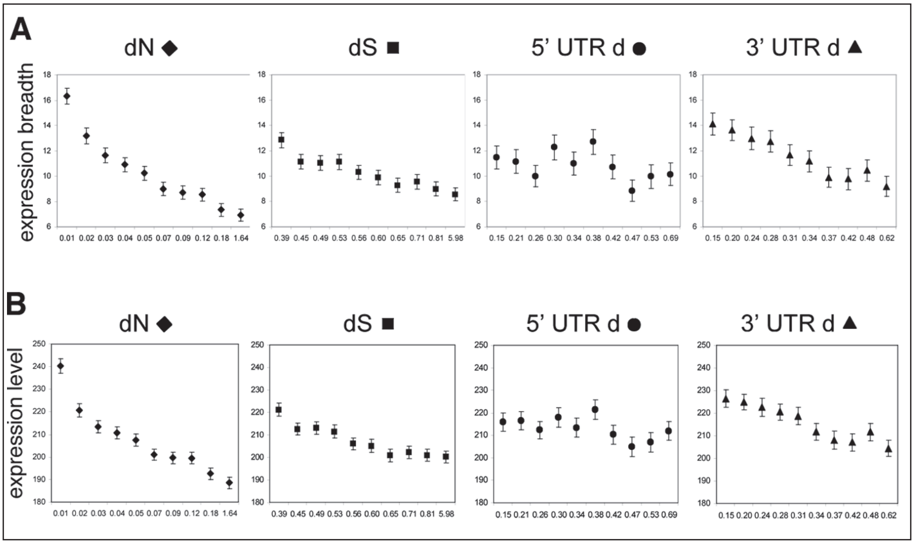

Figure 1. Correlation between evolution rate, expression breadth and expression level. Human microarray data from GEO database (GDS181.soft) were analyzed to determine expression breadth (a; number of tissues with expression level ≥200 AD) and expression level (b; the sum of the $\log_2$ normalized AD values over all tissues) as described previously.[16] Evolutionary distance between the human gene and its mouse ortholog was determined in nonsynonymous (dN) and synonymous (dS) sites of the coding region, 5'-UTRs (5'UTR d) and 3'-UTRs (3'UTR d). Genes were grouped according to the evolutionary distances (which, for orthologs, can be used as proxy for rates) in bins of approximately equal size; mean and variance of expression level and expression breadth were calculated for each bin.

the strongest factor affecting the connection is the local (i.e., species-specific) fitness effect of the gene, usually measured as gene dispensability in knock-out experiments. It has been pointed out that genes experimentally shown to be essential tend to evolve slower than non-essential ones[18,19] although, again, the causal relationship between these parameters has been questioned.[20] Obviously, the gene dispensability over short evolutionary intervals entirely depends on the fitness effect of the gene loss (genes with a lethal knock-out phenotype in a particular species, by definition, cannot be lost in that species), while long-range loss propensity is more subtly determined by the evolution of the whole genome (changes in availability of a complementing gene, availability of an alternative pathway or acquisition and loss of entire modules of the molecular machinery). It is noteworthy that, despite the high significance of the observed correlations between the long-term (as captured in a gene's phyletic patterns) and short-term (determined in actual knockout experiments) propensities for gene loss, the actual dependence is relatively weak. The strength of correlation between nominal variables can be represented in terms of mutual entropy, i.e., the amount of information that can be gained when the data on gene phyletic pattern is added to the data on gene knockout effect (Appendix 1). For *C. elegans* and *S. cerevisiae*, the relative information gain, i.e., the improvement in the prediction of the outcome of the genome-wide gene knockout experiment, from using phyletic patterns was calculated to be ~10.5% and ~15%, respectively. It seems notable that the link between the phyletic pattern (which represents the history of gene losses across the eukaryotic crown group) and the knockout effect are 1.5 times stronger for yeast than for the nematode; this probably reflects the greater complexity and the associated partial redundancy of the metazoan cellular machinery.

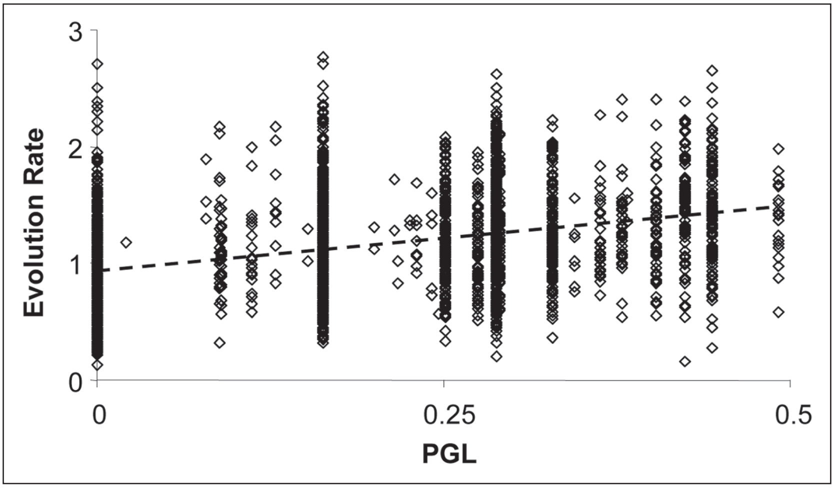

Figure 2. Correlation between sequence evolution rate and gene loss. The data was taken from ref. 15. Horizontal axis—propensity for gene loss; vertical axis—evolution rate (distance from Arabidopsis protein to fungal or metazoan ortholog). Spearman rank correlation coefficient $R = 0.40$ (significant at $p \ll 10^{-10}$).

Gene Duplications and Evolution Rate

The crucial role gene duplications play in evolution had been recognized since the early days of modern evolutionary biology[21,22] and was molded into a coherent concept by Susumu Ohno in the classical 1970 book "Evolution by Gene Duplication."[23] A largely underappreciated aspect in the relationships between gene duplication and evolution of function is that several evolutionary forces, acting on a freshly duplicated pair of genes, seem to work in opposite directions. Duplication creates functional redundancy, which results in an immediate decrease of the purifying selection pressure. However, with the frequency of deleterious mutations being much higher than that of advantageous mutations, the loss of selective pressure leads to rapid "pseudogenization", making the Ohno-style neofunctionalization[23] an unlikely event. Several theoretical explanations have been proposed to resolve this apparent paradox, in particular, the subfunctionalization model, whereby young duplicate genes undergo partial loss of function, leading first to retention of both copies necessary for genetic complementation between them and, later, to functional divergence;[24] dosage effect, postulating direct selective advantage of the increase of gene (product) dosage brought about by duplication,[25] and tissue- or development stage-specific epigenetic silencing of one of the duplicates, which exposes both copies to purifying selection.[26] The observed reduction of species-wide sequence polymorphism in recently duplicated genes in Arabidopsis suggests a role of selection sweeps in initial fixation of duplications.[27] Interestingly, as a counter-point to the common notion of the creative role of gene duplication, a gene loss that "compressed" functions of two paralogs into a single copy has been suggested as the main event that "unlocked" the evolutionary path to flowering plants.[28]

Regardless of the exact nature of the relationships between gene duplication and evolution mode and rate, complex dependencies are seen in quantitative comparisons. The initial increase of evolution rate (but apparently not to the level of the neutral expectation) has been widely observed[25,29-31] although reports differ on whether the two duplicated copies typically evolve at similar[25] or significantly different[32,33] rates. Apparently, the asymmetry in the evolutionary fates of the duplicated copies extends to the patterns of expression and protein-protein interactions, and the response to environmental stress and gene disruption.[34] Large-scale studies indicate, however, that genes which have close paralogs, on average, evolve slower than singletons;[31,35] this probably reflects the stronger tendency of slower-evolving essential genes to retain a duplication for an extended period of time. Interestingly, duplication itself tends to diminish experimentally detectable fitness effect of gene disruption due to the very reason of introducing redundancy into the genetic makeup of the organism.[20,36]

Interactions between Three and More Parameters: More Than the Sum of the Parts?

Considering more than two parameters gives an additional insight into the quantitative-genomic relationships. For example, there appears to be a weak but detectable negative correlation between the evolution rate and experimentally determined number of protein-protein interactions.[5,6] Both of these parameters are correlated with expression level—highly expressed proteins tend to evolve slower and have more interactions. Accounting for the expression level brings the (already weak) correlation between the evolutionary rate and protein interactions below the significance level.[9] There is a convincing argument that the experimental detection of protein-protein interactions is strongly affected by the protein abundance; thus, the interaction data set is biased towards having an artificially high number of interaction partners for highly expressed proteins. This suggests that there might be no direct connection between the position of the protein in the interaction network and its rate of evolution. The debate that followed[7-10,37] failed, so far, to provide a definitive answer beyond the general agreement that "the large-scale data sets remain woefully noisy and incomplete."[8]

Rocha and Danchin applied multiple regression and partial correlation analysis to the data on evolution rate, expression level, functional category, essentiality and metabolic cost of genes in two model bacteria, *Bacillus subtilis* and *Escherichia coli.*[38] They showed that an indirect measure of expression level, the Codon Adaptation Index (CAI), is responsible for the major part (91-94%) of the variance in the evolution rate of bacterial genes, which is explained by multiple linear regression. Rocha and Danchin argue that, when controlled for CAI contribution, the other factors play "minor (if any) role" in determining the evolution rate of bacterial genes and explain the correlations reported by other researchers[18,19] by indirect influence of differences in expression level. While their analysis is very similar in spirit to that of Bloom and Adami,[9,10] there seems to be an important distinction: Bloom and Adami invoke an experimental bias as an explanation of the observed connection between the number of protein-protein interactions and protein abundance which indirectly explains the apparent correlation between the number of interactions and evolution rate; by contrast, Rocha and Danchin consider real correlations between three (more or less) independent variables. The former case, if solid, seems to warrant the dismissal of the observed correlations as artificial; the latter calls for development of a conceptual model taking into account the full complexity of multi-dimensional, inter-correlated data.

The "Social Status" Model

We would like to propose an idealized model that might help in developing biologically relevant null hypotheses for observed connections between quantitative measures of genome evolution and function. Firstly, let us note that it seems useful to make a distinction between "phenotypic" and "evolutionary" variables. The former, e.g., gene expression level or viability of a knock-out mutant, reflect the biology of extant organisms; by contrast, the latter, e.g., sequence evolution rate or propensity for gene loss, reflect various aspects of genome conservation and change over the course of evolution. The relationship between the phenotypic and evolutionary variables appears to have a distinct polarity: the former affect the latter because natural selection constraints or drives the evolutionary change by "testing" the organism's phenotype for fitness, but not vice versa. Phenotypic parameters directly interact with each other (e.g., codon bias of a gene affects its expression level) whereas evolutionary parameters are indirectly correlated (e.g., in accord with the neutral theory of evolution,[39] the evolutionary rates of orthologs in different lineages tend to be similar inasmuch as they perform similar functions in the respective organisms). In a sense, the phenotypic parameters provide the "in-

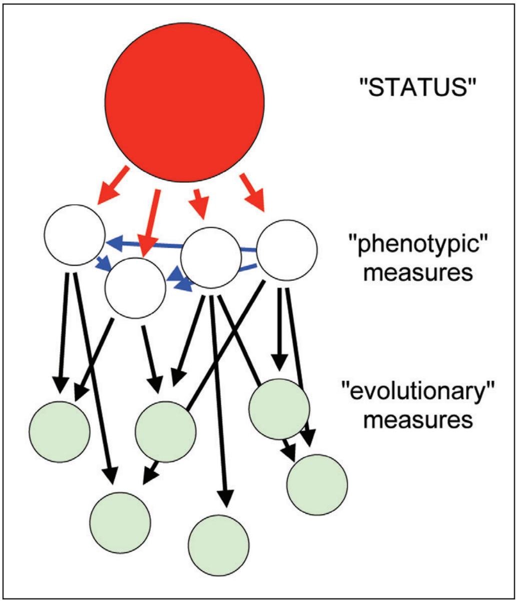

Figure 3. The "gene social status" model. Blue arrows—direct interactions between "phenotypic" (input) variables; black arrows—influence of "phenotypic" (input) variables on "evolutionary" (output) variables; red arrows—manifestation of "social status" in the "phenotypic" variables. A color version of this figure is available online at http://www.Eurekah.com.

put" and the evolutionary parameters represent the "output" of a biological system. Evolutionary parameters are readily produced by comparative genomic techniques (although systematic error may creep in, e.g., in calculations of evolutionary rates over long time spans) whereas most phenotypic parameters can be obtained, on genome scale, only through costly and, at this stage, highly error-prone large-scale experiments.

We suggest that the majority of the relationships between the input parameters (and, indirectly, between the output parameters) can be described with a single, composite variable which reflects the role of the gene in the cell physiology and its mode of evolution. This variable is akin to the gene's "social status in the genomic community" and relates to the importance of its functions in the overall scheme of things. "High-status" genes, which are involved in key house-keeping processes, are more likely to be higher and broader expressed, to have more interaction partners, and to produce lethal or severely impaired knockout mutants. These genes also tend to evolve slower and are less prone to gene loss across various phylogenetic lineages. "Low-status" genes are expected to be weakly expressed, have fewer interaction partners, and exhibit narrower (and less coherent) phyletic distribution. They also, on average, evolve faster and are more often lost during evolution than high-status genes.

Parameters that contribute to the status with the same sign are expected to show positive correlation between each other, whereas those that contribute in the opposite direction are expected to be negatively correlated. Thus, input parameters, which all make a positive contribution to the status (high-status genes are, generally, highly expressed, their products interact with many other proteins, their knockouts have severe fitness effects etc), are positively correlated with each other but negatively correlated with output parameters (the fast-evolving genes typically have a low status). The notion of gene status may provide a useful generator of null hypotheses regarding the connections between variables associated with functioning and evolution of genes (Fig. 3). Any deviation from the expected pattern calls for attention – to the quality of the data, the nature of the analyzed relationship, or both.

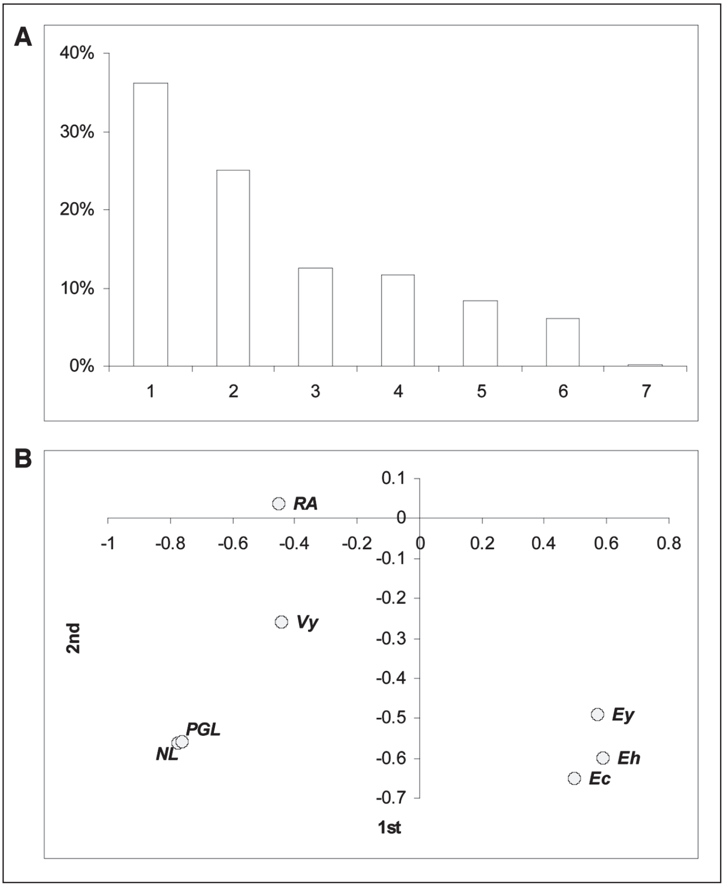

Figure 4. Principal Component Analysis of phenotypic (input) and evolutionary (output) variables associated with eukaryotic KOGs. The values of 7 variables were obtained as previously described:[15] (i) evolution rate (measured as distance from an Arabidopsis protein to fungal or metazoan ortholog), (ii) number of losses in the KOG history (reconstructed using Dollo parsimony), (iii) propensity for gene loss (essentially, number of losses, normalized for the lengths of the corresponding branches), (iv) expression level in yeast, (v) expression level in *C. elegans*, (vi) expression level in humans, and (vii) viability of yeast knock-out mutant, A) Distribution of variance among principal components. Horizontal axis - principal components; vertical axis—fraction of total variance. B) Loadings plot for the original variables in the plane of the first two principal components. Note the positive contribution of the expression level (associated with high status) and the negative contribution of evolution rate, gene loss, and viability of gene disruption mutants (associated with low status) to the first principal component. The data was from refs. 15,16.

Multi-Dimensional Structure of Expression, Evolution Rate, and Gene Loss Data

We investigated the multi-dimensional structure of expression, evolution rate, and gene loss data for a set of orthologous gene families and the correlations of these parameters with the viability of yeast knockout mutants. The principal component analysis (PCA) shows (Fig. 4) that the major direction of the data scatter, which accounts for nearly 40% of the entire variance, is formed by positive contribution from various measures of expression level (EST data for human genes and microarray data for yeast and worm orthologs) and the negative contribution from different measures of evolutionary rate (evolutionary distances between Arabidopsis proteins and their fungal or animal orthologs) and gene loss (propensity for gene loss calculated as previously described[15] or, simply, the number of losses). A coordinate on such an axis can be directly interpreted as a measure of the gene's "social status"; the fact that it is the most significant direction in terms of the data variance indicates that the "status" defined in this fashion is, indeed, important in determining the place of a gene in the data space.

Importantly, despite the fact that all pairwise correlations between the parameters are highly significant and follow the predictions of the "status" model in their signs, the overall level of interdependence between the parameters is quite low. We used the above data on the gene expression, evolutionary rate,and gene loss to predict an outcome of a gene knockout experiment in yeast (Appendix 2). As expected, slowly evolving, evolutionarily stable,and highly expressed genes are more likely to produce a nonviable phenotype compared to the genes from the opposite side of the "status" spectrum. However, the contribution of all these factors is remarkably low - using this information, the Bayesian Linear Discriminator removed only 0.5-5% from the original entropy of the gene knockout data.

Conclusions

The opportunity to analyze, systematically and quantitatively, the connections between numerous measures of genome evolution and function is one of the most alluring avenues of study opened up by the development of genomics and systems biology. Under an optimistic scenario, this might be the key to the main point of entire systems biology enterprise, transforming biology "from stamp collection to physics". Yet, it seems that any researcher who attempts to examine and evaluate the wealth of literature that has accumulated in this area in the last few years hardly can avoid a feeling of uneasiness. There seem to be too many contradicting reports on the same issue and too many high claims based on rather weak (even if statistically significant) evidence. One can easily think of at least four, certainly not exclusive, causes of this situation: i) lack of a general conceptual framework for analysis of connections between genomic variables, ii) the low and nonuniform quality of many types of data, iii) inadequacy of the presently analyzed variables for understanding the connections between evolution and phenotype (we are barking on a wrong tree), iv) the current parameters are, more or less, the best that can be measured, but they are intrinsically of limited importance for understanding those connections, which simply cannot be adequately captured by quantitative analysis (we are barking on the right tree but it is a small one).

Here we made a preliminary attempt to address problem (i) by introducing the notion of the "social status" of a gene and the distinction between "input" and "output" parameters. These are simplistic attempts on a synthesis of the information on genome-evolution-phenotype connections but they seem to work in the sense that the status concept gives unequivocal predictions on the nature of the connection (negative or positive correlation) between any two variables, and these predictions hold for the great majority of the available trials. Thus, any deviations can be construed as a signal of alarm and/or interest.

The apparent utility of the "status" concept is the flip side of the coin. The flop side comes up when we determine how much all the available information on the values of input and output parameters can improve the prediction of the outcome of a genome-scale gene knockout experiment. The improvement that could be achieved with the best possible combination of these parameters was almost shockingly small. This suggests that some combination of factors (ii)-(iv) defines the situation. The problem with the data (ii) surely is transient; there is no doubt that, within the next few years, we will witness a dramatic improvement in the completeness and accuracy of genome-wide measurements of expression, protein interactivity, and other input parameters. There is a chance that this dramatically improves the predictive power of the "gene's social status". If not, the choice will be between (iii) and (iv). The latter possibility, while perhaps discouraging, is not at all unimaginable: the principal determinants of the output values (e.g., evolutionary rate) may well lie in the features of gene and protein structure and function that cannot be captured in simple, numerical values.

***Table A1. Connection between gene knockout data and phyletic patterns for* C. elegans *and* S. cerevisiae**

	C. elegans	*S. cerevisiae*
Total entropy, bit	0.5554	0.8437
Mutual entropy, bit	0.0585	0.1252
Relative gain	10.52%	14.84%

Appendix 1. Mutual Entropy of Gene Knockout Data and Phyletic Patterns

Let p_L be the fraction of genes that produce lethal knockout mutants (obviously, there is a fraction of 1-p_L genes producing a viable mutant phenotype). Taking P_L and 1-P_L as estimates of the probability of a gene to be lethal or nonlethal, respectively; then, the total entropy that can be associated with gene knockout data is

$$H_0 = -p_L\log_2(p_L)-(1-p_L)\log_2(1-p_L)$$

Now, let us group the genes according to their phyletic patterns, and let f_i be the frequency of the i-th pattern. Let us denote the fraction of genes with lethal knockouts in the i-th pattern by p^i_L. If we think of the knockout lethality of a gene as one random variable and of its phyletic pattern as a second random variable, we can compute the conditional entropy of knockout lethality given the phyletic pattern from

$$H_1 = \Sigma f_i[-p^i_L\log_2(p^i_L)-(1-p^i_L)\log_2(1-p^i_L)]$$

The mutual entropy between these two random variables is defined as H_0-H_1; this is an accepted measure for the amount of information that each random variable carries about the other.[40] Here, we shall use the relative gain, which is a normalized version of the mutual entropy, defined as $(H_0-H_1)/H_0$.

The data on viability of gene knockout mutants were obtained from[41] for *C. elegans* and from[42] for *S. cerevisiae*. Phyletic patterns for KOGs were taken from the eukaryotic KOG database.[43]

Appendix 2. Expression Level, Evolution Rate, and Gene Loss As Predictors of Viability of Gene Knockout Mutants

We attempt to predict the viability of gene knockout mutants[41,42] using the data on expression level, evolution rate and gene loss.[15] We employed Bayesian Linear Discriminant Analysis[44] to find an optimal linear discriminant function. In brief, we compute a linear function $g(X)$, where X is a vector of variables (namely, expression level in yeast, nematode, and human, minimum and average evolutionary distance from Arabidopsis to fungi and metazoan, PGL, and number of gene losses in a KOG). For a given X, the gene knockout is predicted to be lethal if $g(X) > 0$ and nonlethal if $g(X) < 0$. The function $g(X)$ is the linear function that guarantees minimum classification error on the training dataset.

As with associating the mutant phenotype with phyletic patterns, we define the entropy of the gene knockout data as

$$H_0 = -p_L\log_2(p_L)-(1-p_L)\log_2(1-p_L)$$

where p_L is the total fraction of lethal mutants. With the prediction, obtained using Bayesian Linear Discriminator, let us define the fraction of predicted lethals as f^L, fraction of lethal phenotypes observed among predicted lethals as p^L_L (true positives), and fraction of lethal

***Table A2. Prediction of gene knockout phenotype from expression level, evolution rate and gene loss for* C. elegans *and* S. cerevisiae**

	C. elegans	*S. cerevisiae*
Total initial entropy, bit	0.7635	0.9125
Mutual entropy, bit	0.0034	0.0484
Relative information gain	0.0044	0.0530

phenotypes observed among predicted nonlethals as $p^{N}{}_{L}$ (false positives). The entropy, given the prediction, is

$$H_1 = f^L[-p^L{}_L\log_2(p^L{}_L)-(1-p^L{}_L)\log_2(1-p^L{}_L)]+(1-f^L)[-p^N{}_L\log_2(p^N{}_L)-(1-p^N{}_L)\log_2(1-p^N{}_L)]$$

Again, the mutual entropy is defined as (H_0-H_1) and the relative gain is (H_0-H_1)/$H_{0.}$[40]

References

1. Steinmetz LM, Davis RW. High-density arrays and insights into genome function. Biotechnol Genet Eng Rev 2000; 17:109-146.
2. Steinmetz LM, Davis RW. Maximizing the potential of functional genomics. Nat Rev Genet 2004; 5(3):190-201.
3. Hurst LD, Pal C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 2004; 5(4):299-310.
4. Wolfe KH, Li WH. Molecular evolution meets the genomics revolution. Nat Genet 2003; 33(Suppl):255-265.
5. Fraser HB, Hirsh AE, Steinmetz LM et al. Evolutionary rate in the protein interaction network. Science 2002; 296(5568):750-752.
6. Jordan IK, Wolf YI, Koonin EV. No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly. BMC Evol Biol 2003; 3(1):1.
7. Fraser HB, Wall DP, Hirsh AE. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 2003; 3(1):11.
8. Fraser HB, Hirsh AE. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol Biol 2004; 4(1):13.
9. Bloom JD, Adami C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol 2003; 3(1):21.
10. Bloom JD, Adami C. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. BMC Evol Biol 2004; 4(1):14.
11. Duret L, Mouchiroud D. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 2000; 17(1):68-74.
12. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics 2001; 158(2):927-931.
13. Zhang P, Gu Z, Li WH. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 2003; 4(9):R56.
14. Zhang L, Li WH. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 2004; 21(2):236-239.
15. Krylov DM, Wolf YI, Rogozin IB et al. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 2003; 13(10):2229-2235.
16. Jordan IK, Marino-Ramirez L, Wolf YI et al. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol 2004; 21(11):2058-2070.
17. Novichkov PS, Omelchenko MV, Gelfand MS et al. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 2004; 186(19):6575-6585.
18. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature 2001; 411(6841):1046-1049.
19. Jordan IK, Rogozin IB, Wolf YI et al. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 2002; 12(6):962-968.

20. Yang J, Gu Z, Li WH. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol 2003; 20(5):772-774.
21. Fisher RA. The possible modification of the response of the wild type to recurrent mutations. Am Nat 1928; 62:115-126.
22. Haldane JBS. The part played by recurrent mutation in evolution. Am Nat 1933; 67:5-19.
23. Ohno S. Evolution by gene duplication. Berlin-Heidelberg-New York: Springer-Verlag, 1970.
24. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics 2000; 154(1):459-473.
25. Kondrashov FA, Rogozin IB, Wolf YI et al. Selection in the evolution of gene duplications. Genome Biol 2002; 3(2):RESEARCH0008.
26. Rodin SN, Riggs AD. Epigenetic silencing may aid evolution by gene duplication. J Mol Evol 2003; 56(6):718-729.
27. Moore RC, Purugganan MD. The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 2003; 100(26):15682-15687.
28. Albert VA, Oppenheimer DG, Lindqvist C. Pleiotropy, redundancy and the evolution of flowers. Trends Plant Sci 2002; 7(7):297-301.
29. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science 2000; 290(5494):1151-1155.
30. Nembaware V, Crum K, Kelso J et al. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res 2002; 12(9):1370-1376.
31. Jordan IK, Wolf YI, Koonin EV. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol 2004; 4(1):22.
32. Conant GC, Wagner A. Asymmetric sequence divergence of duplicate genes. Genome Res 2003; 13(9):2052-2058.
33. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004; 428(6983):617-624.
34. Wagner A. Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol 2002; 19(10):1760-1768.
35. Davis JC, Petrov DA. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol 2004; 2(3):E55.
36. Gu Z, Steinmetz LM, Gu X et al. Role of duplicate genes in genetic robustness against null mutations. Nature 2003; 421(6918):63-66.
37. Hahn MW, Conant GC, Wagner A. Molecular evolution in large genetic networks: Does connectivity equal constraint? J Mol Evol 2004; 58(2):203-211.
38. Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 2004; 21(1):108-116.
39. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press, 1983.
40. Cover TM, Thomas JA. Elements of information theory. Boston: Wiley-Interscience, 1991.
41. Kamath RS, Fraser AG, Dong Y et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 2003; 421(6920):231-237.
42. Giaever G, Chu AM, Ni L et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature 2002; 418(6896):387-391.
43. Koonin EV, Fedorova ND, Jackson JD et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 2004; 5(2):R7.
44. Duda RO, Hart PE, Stork DG. Pattern classification. Boston: Wiley-Interscience, 2000.