Complex Principal Component and Correlation Structure of 16 Yeast Genomic Variables

Fabian J. Theis,[1] Nadia Latif,†,[2] Philip Wong,†,[1] and Dmitrij Frishman*,[1,2]

[1]Helmholtz Center Munich—German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, Neuherberg, Germany

[2]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Maximus-von-Imhof-Forum 3, Freising, Germany

†These authors contributed equally to this work.

***Corresponding author:** E-mail: d.frishman@wzw.tum.de.

Associate editor: Aoife McLysaght

Abstract

A quickly growing number of characteristics reflecting various aspects of gene function and evolution can be either measured experimentally or computed from DNA and protein sequences. The study of pairwise correlations between such quantitative genomic variables as well as collective analysis of their interrelations by multidimensional methods have delivered crucial insights into the processes of molecular evolution. Here, we present a principal component analysis (PCA) of 16 genomic variables from *Saccharomyces cerevisiae*, the largest data set analyzed so far. Because many missing values and potential outliers hinder the direct calculation of principal components, we introduce the application of Bayesian PCA. We confirm some of the previously established correlations, such as evolutionary rate versus protein expression, and reveal new correlations such as those between translational efficiency, phosphorylation density, and protein age. Although the first principal component primarily contrasts genomic change and protein expression, the second component separates variables related to gene existence and expressed protein functions. Enrichment analysis on genes affecting variable correlations unveils classes of influential genes. For example, although ribosomal and nuclear transport genes make important contributions to the correlation between protein isoelectric point and molecular weight, protein synthesis and amino acid metabolism genes help cause the lack of significant correlation between propensity for gene loss and protein age. We present the novel Quagmire database (Quantitative Genomics Resource) which allows exploring relationships between more genomic variables in three model organisms—*Escherichia coli*, *S. cerevisiae*, and *Homo sapiens* (http://webclu.bio.wzw.tum.de:18080/quagmire).

Key words: molecular evolution, genome analysis, proteomics, principal component, analysis.

Introduction

Over the past two decades, experimental high-throughput technology has been delivering increasingly accurate measurements describing the functioning of genes, their transcripts, and their encoded proteins in the cell. Genome- and proteome-wide measurements of functional genomic variables, such as the abundance of transcripts and proteins in the cell or gene essentiality, combined with the availability of completely sequenced genomes for a wide range of species, open unprecedented opportunities for studying collective behavior of genes, and their evolution. Multifaceted interdependencies between phenotypic and evolutionary properties of genes have been subject of active research, especially due to their implications for systems biology.

In particular, the evolutionary rate (ER) of proteins has been investigated in the context of a large variety of genome-scale data sets, and a number of correlates have been identified which include propensity of gene loss (PGL), protein length, designability, mRNA abundance, codon adaptation index (CAI), number of protein interaction partners, essentiality, and exon–intron structure, as well as functional annotation (Pál et al. 2006; Rocha 2006). Many of these observed variables are in fact measurable or computable surrogates of the underlying biological phenomena. Thus, for example, protein length is related to the cost of biosynthesis of proteins while mRNA abundance is a measure of gene expression level (EL).

The most comprehensive to date assessment of protein sequence ER in five closely related yeast species (Xia et al. 2009) found both protein and mRNA abundance, protein function, and the content of certain amino acids to be among its strongest genomic correlates. Other relevant features related to the evolution of proteins include structural disorder (SD), GC content, number of interactions, gene duplicability, and essentiality, as well as the number of transcriptional regulators.

Many observed pairwise correlations are very weak and may disappear when a confounding variable has been controlled for. For example, it was reported that in yeast, the correlation between the number of protein interactions and the ER is due to the bias of experimental interaction studies toward more abundant proteins (Bloom and Adami 2003). Interestingly, Lemos et al. (2005) reported that in

Drosophila melanogaster, such bias can be eliminated by using high-confidence interaction data.

More recently, it has been argued that the complex structure of interdependencies of varying strength between multiple quantitative genomics variables can be best captured by multidimensional analysis (Wolf 2006), and indeed, the first studies collectively analyzing multiple variables were published (Drummond et al. 2006; Wolf 2006; Wolf et al. 2006). Specifically, Wolf et al. (2006) applied principal component analysis (PCA) to a set of seven genomic variables to characterize phenotypic and evolutionary features of eukaryotic orthologous groups (KOGs, Koonin et al. 2004). The variables were EL, ER, knockout effect (KE), number of protein–protein (PPI) and genetic (GI) interactions a protein is involved in, number of paralogs (NP) as well as PGL. The first principal component in this analysis (PC1) received strong opposite contributions from ER and PGL on the one hand and PPI, EL, KE, and NP, on the other hand. PC1 could thus be interpreted as gene status, with high-status (important) genes being those that lead to lethal phenotypes when being knocked out, are highly expressed, strongly connected on the interaction network, have many paralogs, and are evolutionary conserved. The second principal component (PC2) was dominated by strong positive contributions of NP and GI and a negative contribution of KE. According to Wolf et al., PC2 reflects gene adaptability to changes in the cellular and extracellular environment. Many adaptable genes can be knocked out without major effects on fitness, and are functionally backed up by other genes.

One technical difficulty in analyzing multiple genomic variables is that the results of high-throughput experimentation often do not cover the entire gene complement of the organism studied. Many methods of multivariate statistics, such as PCA, cannot operate on incomplete data, in particular when some values are missing in almost all samples (genes). It is therefore crucial to either perform careful preprocessing of data or to adapt the estimation method appropriately.

The presence of outliers is another recurring problem in quantitative genomics analyses. In microarray studies, for example, up to 15% of the data involving extreme values is not a rarity. Commonly used second-order methods, based on the assumption that the underlying data follow a multivariate Gaussian distribution, are particularly prone to outliers. Less probable values far away from the mean cause the mean to shift toward these values. Difficulties caused by data outliers can be handled by either discarding the outliers altogether (provided that they can be robustly identified) or by using robust estimation methods. In the univariate case, the latter option can imply, for example, replacing the mean by the median, which is not affected by extreme values as long as there are not too many of them. In the multivariate setting robust estimation is a more challenging task because, for example, a definition of multivariate median cannot be formulated in closed form by arithmetic operations on the data. For this reason, sample removal is often considered sufficient for practical applications.

In Wolf et al. (2006), a three-step sample-reduction approach for data preprocessing was adopted: 1) remove all samples with more than one data value missing, 2) replace the remaining missing values by the mean of each respective variable, and 3) remove outliers with values deviating from the mean by more than one standard deviation. The number of samples (KOGs) for data analysis was reduced from 10,058 to 4,124 in step 1) and to 3,912 in step 3), still a sizeable data set. As we show below for bigger sets of incompletely defined variables, this simple sample-reduction approach can result in a prohibitively large fraction of the data excluded from consideration, justifying the need for more sophisticated data preprocessing strategies.

The goal of this work is to extend previously published analyses to a larger set of functional genomic variables and then to identify meaningful components/directions within this multivariate data set using an improved methodology capable of coping with large numbers of missing values and extreme outliers. Toward this end, we have constructed a database for genomic variables called Quagmire. We analyze a total of 16 variables from this database, adding protein abundance (ABU), CAI, SD, phosphorylation density (PHD), protein age (AGE), transfer RNA (tRNA) adaptation index (TAI), protein half-life (HL), pI (PI), and protein molecular weight (MW) to the initial set of variables used by Wolf et al. 2006. Correlation analysis shows that two-thirds of the variable pairs share significant correlation, suggesting evidence for constraint and adaptation. Functional classes of genes influencing these correlations are derived from enrichment analysis. Different correlates were found for the number of paralogs derived by sophisticated phylogenetic analyses (NP) and by a simplistic method involving counting significant Blastp (Altschul et al. 1997) hits (NPB). Instead of standard PCA, we apply Bayesian PCA, which allows for a probabilistic handling of missing values and can operate on the entire set of available samples.

Materials and Methods

Yeast Genome Data

Sequence Data

Whole genome sequence data (corresponding to the NCBI genome project ID 13838) as well as MW and pI values (pI) for 6,086 yeast gene products were downloaded from the PEDANT database (Walter et al. 2009). Database IDs for all other data types (described below) were mapped to standard gene names for *S. cerevisiae*. In the rare cases, when such mapping was not possible respective data items were excluded from consideration.

The percentage of structurally disordered residues (SD) was determined by DISOPRED (Ward et al. 2004). The number of disordered residues within a sequence was divided by protein length to calculate percentage of disordered residues for each protein. SD values range from 0 to 100.

Phosphorylation Density

Phosphoproteomics data was provided by the Matthias Mann group (Gnad et al. 2009). The obtained data set including unpublished data contained positions of phosphorylated TYR/THR/SER residues in 1,292 yeast proteins, with 4.8 phosphorylated residues per protein on average. For each protein, we calculated PHD by dividing the number of phosphorylated sites by protein length. All other yeast proteins were considered to be nonphosphorylated and were assigned zero PHD.

Protein Abundance

Newman et al. (2006) provided abundance data for a total of 4,130 proteins measured from growth in yeast extract peptone dextrose medium (supplementary table S1). Abundance values for 2,526 proteins were explicitly specified, whereas the remaining 1,604 proteins were described as having "very low abundance." This latter set of proteins was assigned zero abundance in our final data set.

Protein Half-Lives

Protein half-lives (HL) were measured by Belle et al. (2006) by monitoring the abundance of TAP-tagged proteins as a function of time upon inhibiting protein synthesis. The distribution of half-lives was found to be close to log-normal, with a mean and median half-life of approximately 43 min. The half-life values measured for 3,751 proteins range from 2 to 115,997 min. There are only seven proteins that have a half-life greater than 6,000 min.

Number of Protein–Protein and Genetic Interactions

Interaction data were downloaded from the BioGRID database (Breitkreutz et al. 2008) that provides a comprehensive literature-curated collection of protein and genetic interactions for major model organism species. The data are downloadable for each species in a separate text file (we used file version 2.0.45). The downloadable file provides the complete set of genetic and protein interactions along with respective literature sources and experimental details. Protein–protein and genetic interactions can be distinguished by the experimental system specified for each interaction. Proteins in BIOGRID involved only in self-interactions (PPI and GI) were assigned values of 0.

mRNA Expression Level

We used a reference set of mRNA expression values for 6,249 yeast genes constructed by Greenbaum et al. (2002) by merging and scaling the results of several previously published gene chip and serial analysis of gene expression experiments.

Codon Adaptation Index

CAI is a measure of synonymous codon usage bias and can take values from 0.0 (no bias) to 1.0 (maximum bias). Technically, it can be calculated based on the similarity of codon usage between a given gene and a trusted set of highly expressed genes from a given organism. It is thus a computational proxy for experimentally measured mRNA concentration (Sharp and Li 1986). We obtained CAI values for 5,980 yeast genes from the *Saccharomyces* Genome Database (Christie et al. 2004) (http://downloads.yeastgenome.org/protein_info/).

tRNA Adaptation Index

The TAI is a measure of the tRNA usage by coding sequences inspired by the CAI of Sharp and Li (1986). The TAI scoring scheme is described in detail in dos Reis et al. (2004). High TAI values correspond to high levels of translational efficiency and vice versa. The TAI data set was downloaded from Man and Pilpel (2007; supplementary table S2) and contained data for 3,338 genes with values ranging from −2 to 2.15.

Gene Essentiality

Gene essentiality (ESS) was defined based on disruption data downloaded from the FTP site of the CYGD database (Güldener et al. 2005) (ftp://ftpmips.gsf.de/yeast/catalogues/gene_disruption/gene_disruption_18052006). CYGD contains yeast essentiality data collected from over 800 sources. Within this data set, 940 genes have a lethal KE (assigned the value of 1) and 4,656 genes have a viable KE (assigned the value of 0), whereas data for 490 genes are not available.

Number of Paralogs

We downloaded the set of yeast paralogous families from the EnsemblCompara (Vilella et al. 2009) resource. These paralogy predictions were made by a sophisticated gene-tree based procedure that allows for reconstructing the evolutionary history of yeast gene families based on comparison with other species. Additionally, we used a straightforward approach to find paralogs based on Blastp similarity searches. The number of duplicates found by Blastp at an *E* value threshold of 10^{-10} is denoted NPB.

Propensity for Gene Loss

PGL is a measure of the evolutionary conservation of a gene as a whole introduced by Krylov et al. (2003). It was determined based on the phyletic patterns describing presence or absence of orthologous groups (KOGs) in seven eukaryotic species including yeast. Because the odds for a gene to be lost grow with time, the topology of the phylogenetic tree has to be taken into account. For a particular gene, PGL is calculated as the ratio between the sum of lengths of the tree branches associated with the loss of that gene and the sum of all branch lengths. If a given KOG occurs in all seven eukaryotic species, its PGL value is zero. As stated in the reference (Krylov et al. 2003), the value of one (the highest possible PGL) would theoretically be assigned to a gene present in the last common ancestor but lost in all lineages. Because such situations cannot be observed, PGL values can be in the range between zero and a maximum value of less than one. We downloaded PGL data from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl). The file contained values for 4,852 KOGs ranging from 0 to 0.49. We assigned the PGL value of each KOG to its member yeast genes. For example, if KOG0002 has a value 0.0675, then all yeast genes belonging to this KOG are assigned the value 0.0675. The final PGL data set contained values for 3,883 genes.

Protein Age Group (AGE)

Kim and Marcotte (2008) assigned yeast genes to different age groups based on the occurrence patterns of their constituent PFAM (Finn et al. 2010) domains in three kingdoms of life—archaea (A), bacteria (B), and eukaryotes (E)—as well as in fungi (F). The group "ABE" includes the oldest proteins containing domains found in all three kingdoms. The group "AE" lists proteins that are common to archaea and eukaryotes, whereas the group "BE" lists proteins that are common to bacteria and eukaryotes. Proteins in the group "E" are specific for eukaryotes excluding fungi. Finally, the group "F" contains the youngest proteins present only in fungi. We encoded age groups numerically such that the genes belonging to the ABE group were represented by the number 3, groups A and BE, presumed to be equally old, were both assigned the number 2, whereas E and F were encoded by 1 and 0, respectively.

Evolutionary Rate

ERs were estimated by Wall et al. (2005) by calculating the ratio between nonsynonymous (*d*N) and synonymous (*d*S) divergence based on DNA sequence alignments of *S. cerevisiae* genes with orthologs from three other species from the *Saccharomyces* genus: *S. bayanus*, *S. mikatae*, and *S. paradoxus*. The adjusted *d*N/*d*S values for 3,035 genes obtained from supplementary materials were in the range from 0 to 0.53.

Preprocessing of Data

The data set consists of 16 variables determined for 6,086 open reading frames and was stored as a matrix X. In a first preprocessing step, each variable was centered by subtracting its mean calculated over non-missing values. Then all variables were normalized to unit standard deviation in order to make them comparable on the same scale; again the standard deviation was calculated only over non-missing values.

Correlation Analysis

In this contribution, we decided to discuss data in the context of Pearson's correlation and corresponding principal components in order to describe the second-order moments of the data. Without going into the debate of nonparametric versus parametric models, we want to stress that of course also other nonparametric covariance estimators, such as Spearman statistics or Kendall's tau exist (see e.g., Visuri et al. 2000); however, they also do not measure higher order dependencies, and their use in PCA is not as straightforward. In particular, the inclusion into a Bayesian framework to efficiently deal with missing values is not yet clear. We decided to stay in the parametric world of Pearson's correlation in order to allow more direct comparison of findings with a previous study (Wolf et al. 2006) and to focus the reader's attention only on linear aspects of the data correlations.

PCA With Missing Values Using a Bayesian Framework (Bayesian PCA)

We seek to identify key components contributing mostly to the variance of the total data set *X*. The standard machine learning approach to this problem, the PCA (Jolliffe 2002), iteratively projects the centered data matrix *X* along its direction of maximal variance. The projections are known as principal components and are mutually orthogonal.

PCA can be interpreted as a compression of data, minimal in the sense of the mean square error. This is equivalent to the minimum of the cost function

$$C(\mathrm{U}, \mathrm{V}) = \sum_{\mathrm{X}_{ij} \,!=\, \mathrm{NA}} (\mathrm{X}_{ij} - (\mathrm{UV})_{ij})^2,$$

where X_{ij} is the value of variable *j* for gene *i*, and U is a matrix for mapping principal component matrix V to data matrix X, respectively.

This cost function ignores missing values (X_{ij} ! = NA). Methods for minimizing this cost function have been proposed (see Ilin and Raiko 2010; for a review). Here, we employ the probabilistic model for PCA (Bayesian PCA; Ilin and Raiko 2010), which has the benefit of a well-described regularization, important in our case of noisy large-scale data. The model is given by

$$\mathrm{X}_j = \mathrm{UV}_i + m + \text{error},$$

with Gaussian priors on V_j, mean *m* and error with variance s^2. Here, X_j and V_j denote the *j*-th columns of the matrices X and V, respectively. This results in a formulation of the likelihood $p(\mathrm{V}|\mathrm{X},\mathrm{U},s)$. Because without regularization such likelihood maximization is prone to overfitting, a Bayesian regularization of this likelihood was proposed (Ilin and Raiko 2010) by also imposing Gaussian priors on U and *m*. In our case, only the prior on U matters because we preprocess the data such that mean equals zero. We employ the variational Bayesian approach to approximate the posterior by expectation-maximization.

In our setting, we use Bayesian PCA with unrestricted Gaussian densities (which corresponds to full covariance matrices) for approximating the posterior distributions of U_i and V_j.

Influential Genes Affecting Pairwise Correlations

We employ a simple greedy algorithm to find top genes affecting pairwise correlations the most when not considered for the analysis. Genes were removed one-by-one from the set of yeast genes and the effect (magnitude and sign) on the Pearson correlation value each time was noted. The genes that changed the correlation value the most (in the direction of the opposite sign) when excluded from the correlation calculation were considered to be top influential genes. For the purposes of this paper, the number of top influential genes was chosen such that significant enrichment of annotated functions could be found.

Statistical Difference Between Distributions

The *P* value that two distributions are different is estimated by both the Mann–Whitney and Kolmogorov–Smirnov test. The maximal value from the two tests is reported.

FunCat Enrichment Analysis

Yeast genes are assigned to FunCat functional categories (Ruepp et al. 2004). *P* values for enrichment of certain categories in selected groups of genes compared with the rest

of the genome were estimated using the hypergeometric distribution. If k genes belonging to a particular functional category is found in s query genes and C genes have been assigned to this category in the entire genome of G genes, the enrichment P value is estimated by

$$P \approx 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{s-i}}{\binom{G}{s}}$$

The P value is subsequently Bonferroni corrected.

Results and Discussion

Quagmire: A Quantitative Genomics Resource

We created a novel database, QUAGMIRE, in which we manually collected data regarding various quantitative properties of genes and proteins, such as EL, protein abundance, protein half-life, PHD, etc. Information was extracted from online supplementary materials (Word documents, Excel tables) of original research papers as well as from Web pages of respective labs. The first version of QUAGMIRE focuses on three best-studied model organisms: the eubacterium *Escherichia coli*, the unicellular eukaryote *S. cerevisiae*, and *Homo sapiens*. We have set up a web interface that allows to explore relationships between genomic variables and to download the entire data set. Quagmire is freely available for download from http://webclu.bio.wzw.tum.de:18080/quagmire.

Overview of the Yeast Data Set and Correlation Structure

In this work, we consider 16 properties (genomic variables) of yeast genes or their gene products (table 1) from the Quagmire database. A dedicated webpage has been setup that shows pairwise correlations among all the genomic variables used in this paper (http://webclu.bio.wzw.tum.de:18080/quagmire/yeast-correlation.jsp). Some of these variables can be classified as evolutionary variables (PGL, ER, AGE, and NP) determined by genome comparison. Other variables are phenotypic variables describing various aspects of gene function and structure. MW, pI, SD, CAI, and TAI are intrinsic properties of proteins and genes that are directly computed from their amino acid or DNA sequences, respectively. Values for other phenotypic variables (PHD, ABU, HL, PPI, GI, EL, and ESS) were extracted from experimental papers.

A basic data structure required to understand the complex interplay between the genomic variables listed above is the matrix of their pairwise correlations shown in table 2. We will refer to this table in our analysis presented below.

Missing Values and Outliers

Although sequence derived variables such as MW or pI are complete in terms of having a value for each gene, experimental data often have values missing. For many variables, listed in table 1 values could only be determined for a subset of genes (fig. 1). For example, in our data, genetic interactions are known for only 3,665 (60.4%) of protein-coding genes. ER can only be calculated for 3,035 genes because other genes either do not have orthologs in one of the three yeast species (*S. bayanus*, *S. mikatae*, or *S. paradoxus*), contain introns or did not pass an alignment quality filter. The full set of 16 variables is available for only 640 (10.5%) protein-coding genes.

For our set of 16 variables, applying the three-step strategy utilized by Wolf et al. 2006 (removing genes with more than one missing value, replacing missing values left with the mean and removing outliers of more than one standard deviation) would result in removing 4,846 samples (genes), or 80% of the data.

Another problem with the data preprocessing approach mentioned above is its assumption of normality when removing outliers. As illustrated by the boxplots in figure 2 and by the quantile–quantile plots in supplementary figure S1, Supplementary Material online, the variables show substantial deviations from both normal and lognormal distributions. Moreover, when replacing a sizeable number of samples by their mean, excessively strong mass put on the mean can disturb the original distribution and further complicate outlier detection.

Application of the Bayesian PCA to the Extended Set of 16 Genomic Variables

Correlates to ER

Upon applying the Bayesian PCA method described in the Material and Methods section to the set of 16 genomic variables, we obtained the PCA biplot shown in figure 3 (Statistical significance of the found correlations is demonstrated in supplementary fig. S2, Supplementary Material online). As expected, the most pronounced effect observed in figure 3 is the strong opposite contributions made to PC1 by ER, on the one hand, and measures related to protein abundance (EL, CAI, and ABU), on the other hand. ER and protein abundance-related variables are significantly anticorrelated (fig. 4; table 2). It has been known for a number of years in a wide range of organisms that gene EL is a strong predictor of ER (Pal et al. 2001; Yang et al. 2009), and a number of explanations for this phenomenon have been introduced. Akashi (2001) proposed that for highly expressed genes, there is a selection against change of amino acids associated with less optimal codons in order to maintain high accuracy of protein synthesis (translational selection). Theoretical models have been proposed to explain the influence of codon bias on protein production rate (Gilchrist 2007). Later, it was argued that highly expressed genes evolve slowly in order to prevent mutation-induced and native sequence misfolding of their encoded proteins (translational robustness; Drummond et al. 2005; Wolf et al. 2010). The need for translational robustness can help explain the need for translation selection. The requirement for translational accuracy and robustness is related to organismal fitness. Rocha and Danchin (2004) postulated that selection against deleterious mutations

Table 1. Sixteen Yeast Variables Examined in This Study.

Name	Abbreviation	Type	Number of Genes	Value Range		References
				Min	Max	
Molecular weight	MW	Numeric	6,068	1,977	559,310	Walter et al. (2009)
Isoelectric point	pI	Numeric	6,086	3.195	13.460	Walter et al. (2009)
Percentage of structurally disordered residues	SD	Numeric	6,086	0 (24)	100	Ward et al. (2004)
Phosphorylation density	PHD	Numeric	1,292	0 (4794)	0.1059	Gnad et al. (2009)
Protein abundance	ABU	Numeric	4,130	0 (1616)	86,150	Newman et al. (2006)
Protein half-life	HL	Numeric	3,739	2	115,997	Belle et al. (2006)
Number of protein–protein interactions	PPI	Numeric	3,820	0 (78)	366	Breitkreutz et al. (2008)
Number of genetic interactions	GI	Numeric	3,665	1	922	Breitkreutz et al. (2008)
mRNA expression level	EL	Numeric	5,635	0.10	392	Greenbaum et al. (2002)
Codon adaptation index	CAI	Numeric	5,980	0.04	1	Christie et al. (2004)
tRNA adaptation index	TAI	Numeric	3,330	−2.09	2.15	Man and Pilpel (2007)
Gene essentiality	ESS	Nominal	5,596	0 (4656)	1	Güldener et al. (2005)
Propensity of gene loss	PGL	Numeric	3,883	0 (1418)	0.49	Krylov et al. (2003)
Protein age	AGE	Nominal	4,276	0 (347)	3	Kim and Marcotte (2008)
Evolutionary rate	ER	Numeric	3,035	0 (7)	0.53	Wall et al. (2005)
Number of Paralogs (Ensembl)	NP	Numeric	6,086	0 (4123)	26	Vilella et al. (2009)
Number of Paralogs (Blastp)	NPB[a]	Numeric	6,086	0 (3318)	96	Altschul et al. (1997)

NOTE.—Numbers in parentheses indicate how many zeros are in the corresponding data set. If no parentheses appear then no zeros are present.
[a] Two alternative definitions of the number of paralogs (NP and NPB) were used (see Materials and Methods).

must be stronger in proteins that have a large overall impact on the organismal fitness which, in turn, is expected to correlate with EL. Recently, a fitness cost–benefit model of gene expression has been related to ER (Gout et al. 2010).

A high magnitude negative correlation does not necessarily imply that variables are far apart along principal components. For example, the distance between AGE and ER along PC1 is not as great as between ER and EL, but we observe that AGE is more negatively correlated with ER as noted for mammalian proteins (Vinogradov 2010).

From table 2, we also observe a significant negative correlation between ER and PPI (from BIOGRID). There has been much debate as to whether proteins involved in many interactions evolve more slowly, with some studies confirming this trend (Fraser et al. 2002), whereas others rejecting it (Batada et al. 2006). To a large extent, the answer to this question depends on the protein interaction data set used. In a large-scale comparison of various interaction databases (Saeed and Deane 2006), it was established that the network connectivity of proteins derived from BioGRID data (at that time) as well as from DIP (Salwinski et al. 2004), MINT (Ceol et al. 2009), and two INTACT (Kerrien et al. 2007) subsets shows a weak but statistically significant association with the rate of protein evolution, whereas BIND (Bader et al. 2003) and small-scale manually annotated MIPS data sets (Güldener et al. 2006) are poorly correlated with it. Overall, it is probably fair to say that most of the studies conducted so far did find a rather weak negative correlation between interactivity and ER (Pál et al. 2006). From figure 3, we also observe that PPI and ER variables make an opposite contribution to the PC1.

Table 2. Pairwise Pearson Correlation of 16 genomic variables

	MW	PI	SD	PHD	ABU	HL	PPI	GI	EL	CAI	TAI	ESS	PGL	AGE	ER	NPB[a]
PI	−0.182*															
SD	0.098*	0.068*														
PHD	0.018	−0.117*	0.259*													
ABU	−0.049*	0.009	−0.099*	0.049*												
HL	−0.033*	0.026	−0.005	−0.002	0.018											
PPI	0.034*	−0.005	0.044*	0.030	0.000	0.039*										
GI	0.040*	−0.033	0.048*	−0.013	0.009	−0.017	0.232*									
EL	−0.112*	0.071*	−0.079*	0.089*	0.748*	0.004	0.033*	−0.029								
CAI	−0.101*	0.019	−0.124*	0.091*	0.529*	0.006	0.059*	−0.019	0.626*							
TAI	0.061*	−0.071*	0.050*	0.034*	0.121*	−0.011	0.019	0.030	0.144*	0.298*						
ESS	0.070*	−0.092*	−0.022	0.033*	0.012	0.006	0.265*	−0.055*	0.036*	0.048*	0.065*					
PGL	0.008	−0.045*	−0.071*	−0.030	−0.090*	−0.002	−0.177*	−0.046*	−0.120*	−0.198*	−0.097*	−0.236*				
AGE	−0.080*	−0.001	−0.374*	−0.151*	0.091*	0.028	−0.051*	−0.056*	0.077*	0.102*	−0.054*	−0.039*	0.014			
ER	0.019	0.066*	0.317*	−0.007	−0.165*	−0.004	−0.166*	−0.039	−0.170*	−0.369*	0.018	−0.198*	0.126*	−0.199*		
NPB	0.248*	0.032*	0.060*	0.215*	−0.004	−0.008	0.173*	0.017	−0.011	0.014	−0.011	−0.013	−0.054*	−0.092*	−0.087*	
NP	0.145*	0.000	−0.024	0.169*	0.057*	−0.009	−0.049*	−0.054*	0.022	0.074*	0.030	−0.133*	−0.032*	−0.096*	−0.025	0.464*

[a] Two alternative definitions of the number of paralogs (NP and NPB) were used (see Material and Methods). Cells are colored in the following manner: Light red—significant negative correlation $R \geq -0.2$; Red—significant negative correlation $R < -0.2$; Light green—significant positive correlation $R \leq 0.2$; Green—significant positive correlation $R > 0.2$; and White—nonsignificant correlation.
*R values are shown with significant correlations (P value < 0.05) marked with an asterisk.

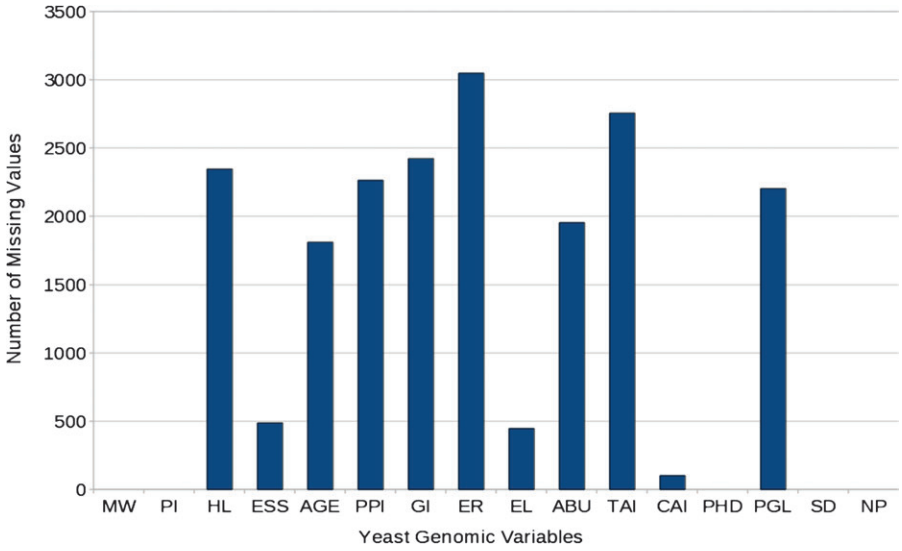

FIG. 1. Distribution of missing values over variables.

Propensity for Gene Loss and Protein Age

Besides sequence ER, we also include in our analysis two different measures pertinent to the conservation and loss of whole genes. The first measure, PGL, reflects the propensity of a certain KOG to be lost in the course of evolution. Because our analysis is yeast-centric, PGL can be considered an estimate of the fraction of eukaryotic genomes in which orthologs of a given yeast gene have been lost. It is computed as the sum of all branch lengths where a given ortholog group has been lost divided by the sum of all examined branch lengths (Krylov et al. 2003). High PGL values thus correspond to genes frequently lost in evolution. For the genomes, we examined the maximum PGL value is 0.49 (table 1). The second measure, AGE, attempts to capture the difference between ancient genes, present in all three kingdoms of life, and more novel genes, specific for all eukaryotes or even to yeast only. The AGE variable is encoded in such a way (see Materials and Methods) that its higher values correspond largely to older proteins (exceptions occur if certain genes are horizontally transferred across taxonomic domains). Importantly, PGL and AGE are not significantly correlated. Fungal-specific proteins (AGE = 0) have a higher than average PGL, but older proteins retained in yeast do not have a significantly higher or lower than average PGL. For example, large groups of relatively ancient proteins (AGE = 3; with homologs in Eukaryotes, Bacteria, and Archaea) can be retained in all examined organisms (PGL = 0) or lost in many such organisms (e.g., PGL > 0.3). There are 670 ancient yeast proteins (AGE = 3), which have a PGL of 0. A significantly enriched group of 139 proteins are involved in protein synthesis (Bonferroni corrected hypergeometric $P < 2.0 \times 10^{-27}$) according to FunCat annotation. There are also 670 proteins with homologs appearing in all three taxonomic domains, but these proteins are more frequently lost during eukaryotic evolution (PGL > ~0.16). One hundred and two of these proteins are involved in aspects of amino acid metabolism, and this group of

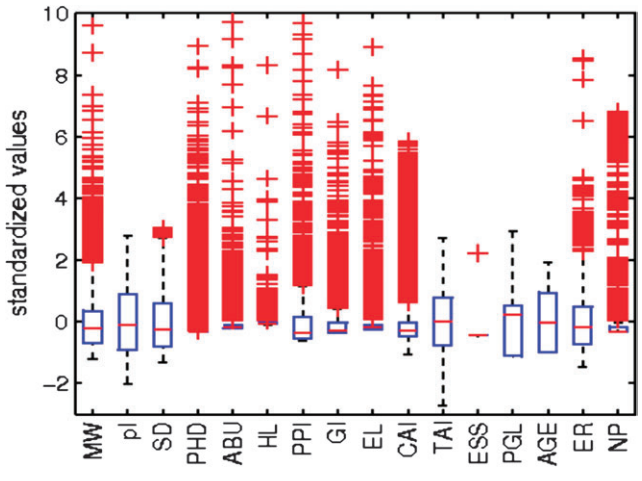

FIG. 2. Data statistics plotted for the full data set after mean removal and rescaling to unit variance. Only samples of magnitude smaller than 10 are shown for clarity; 25 outliers (0.02% of the data) were ignored.

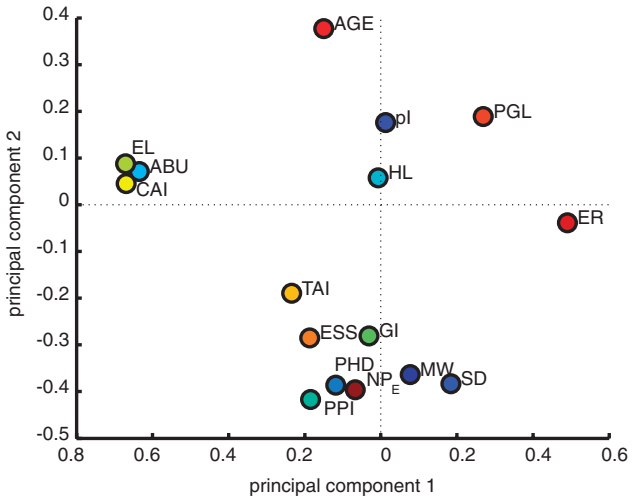

FIG. 3. PCA using the Bayesian PCA algorithm, which efficiently deals with missing values without sample removal.

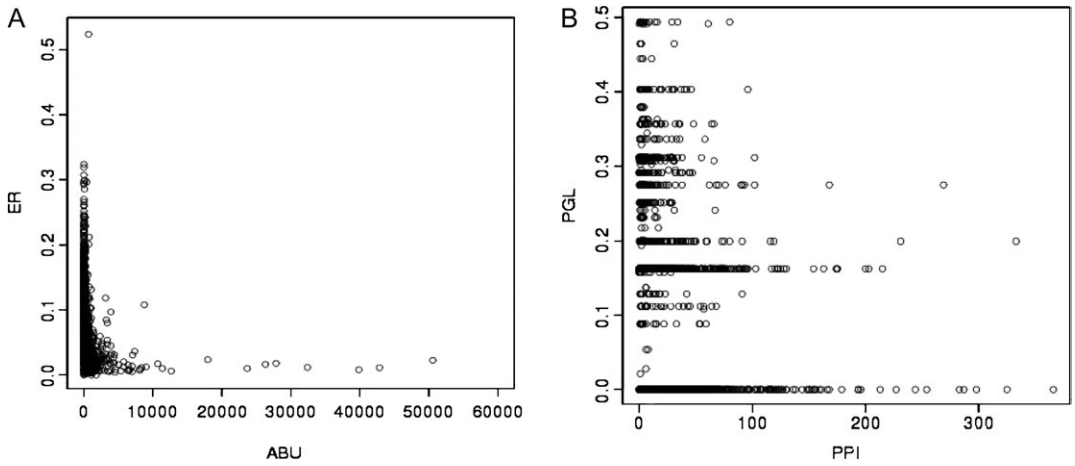

FIG. 4. (A) ER versus protein abundance (ABU). (B) PGL versus number of PPI.

proteins is significantly enriched compared with the fraction of proteins with such annotation in the genome (Bonferroni corrected hypergeometric $P < 1.4 \times 10^{-34}$). Amino acid metabolism related proteins are also significantly enriched amongst 240 proteins with PGL > 0.3. Amino acid metabolism is a category not enriched amongst those three taxonomic domain proteins with PGL $= 0$, and protein synthesis is not enriched amongst such proteins with PGL > 0.16. This example illustrates that certain classes of ancient proteins can be highly conserved (i.e., protein synthesis proteins), whereas other classes are often not retained in genomes (amino acid metabolism proteins). The presence of such classes of proteins contributes to the absence of significant correlation between AGE and PGL. Removing protein synthesis genes from analysis leads to a positive correlation ($r = 0.04$; $P < 0.04$), and removing amino acid metabolism genes leads to negative correlation ($r = -0.01$; $P < 0.04$) between AGE and PGL.

PGL makes a positive contribution to both PC1 and PC2. Along PC1, PGL makes a contribution of the same sign as ER. Genes frequently lost in evolution are often observed to evolve faster (Krylov et al. 2003), and we also see a significant positive correlation between ER and PGL (table 2). Highly divergent genes may have a high PGL because they have escaped ortholog detection. There is also a time dependence between ER and PGL. For many genes, it is likely that they have accumulated many mutations in the process of pseudogenization before they are lost. One fact that makes the two variables different in our data is that ER is computed within closely related yeast species, but PGL is a value computed over relatively distant eukaryotes. PGL is a strong negative correlate with essential gene status (Krylov et al. 2003) but is a weaker correlate to transcript/protein abundance related measures (EL, CAI, and ABU) compared with ER.

Like ER, PGL is clearly anticorrelated with PPI (table 2), as originally reported in (Krylov et al. 2003) and also discussed in (Saeed and Deane 2006). It is interesting that in the latter work, the authors define the age of a protein based on its propensity to have orthologs in other fully sequenced genomes, a measure which is essentially the inverse of PGL used in our work. Significant positive correlation between the genetic and physical interaction degree has been discussed before (Costanzo et al. 2010). Proteins that are highly connected, both physically and through genetic interactions, are observed to be less likely to be lost in evolution (table 2). Both PPI and GI are also distantly placed from PGL and ER along the PC2.

A plot of PGL and PPI is shown in figure 4. It is interesting to contrast this plot with that of ER and ABU (fig. 4), even though the correlation coefficients are similar (table 2). The ABU versus ER plot appears to be under strong constraint such that proteins present in more than 10,000 copies per cell do not have an ER above 0.05. One protein, YDR385W (Translation elongation factor eEF2) was measured at 8,764 copies per cell and has an ER above 0.1. The high ER might be due in part to its nonessentiality. Its abundance, nevertheless, suggests an important role for the protein, but its high ER helps make it a relatively species-specific target for inhibitors, such as sordarins (Shastry et al. 2001).

Translational Efficiency

There is a significant anticorrelation between TAI and PGL but no significant anticorrelation between TAI and ER. In other words, those yeast proteins predicted to have low translation efficiency have orthologs that are more readily lost amongst the examined eukaryotes. Lower translation efficiency might be a reflection of lower importance for the fitness of the cell. Indeed, we observe a significant positive correlation between essentiality and TAI. Ribosomal occupancy (OCC) is a variable which, like TAI, is related to protein production speed (Mittal et al. 2009). We also find significant correlations ($P < 0.05$) between OCC and PGL ($r = -0.09$) and OCC and ESS ($r = 0.06$).

Interestingly, a significant anticorrelation between TAI and AGE is observed. Relatively young fungal-specific proteins have significantly higher TAI's (AGE $= 0$, mean: 0.258 ± 0.650) than those of older proteins (AGE $= 1$, mean: 0.055 ± 0.766; AGE $= 2$: 0.073 ± 0.864; and AGE $= 3$: -0.004 ± 0.764).

Gene Dispensability, Duplication, and Interaction Degree

As initially reported by Jeong et al. (2001), more connected nodes of the yeast protein interaction network also tend to be more essential (less dispensable). A number of potential explanations has been proposed for this phenomenon, including dramatic disruption of network structure upon removal of network hubs (Jeong et al. 2001), more frequent involvement of hubs in essential protein interactions (He and Zhang 2006), as well as participation of prolific interactors in densely connected subnetworks corresponding to functional or structural modules (Zotenko et al. 2008). This effect can be observed if protein interaction data of sufficient quality is used (Batada et al. 2006); it is less prominent or may even be absent in networks derived from individual high-throughput two-hybrid interaction experiments (Coulomb et al. 2005). PPI makes a strong negative contribution to the PC2 together with gene essentiality, whereas PGL, another proxy of the overall gene importance for the cell, makes a strong positive contribution to the component. Entirely in line with the previously reported findings (Wolf et al. 2006), there is a significant positive correlation between PPI and ESS and a significant anticorrelation between these two variables with PGL (table 2). In other words, high-degree nodes tend to be essential and are less likely to be lost in evolution.

In this study, we find a significant anticorrelation between the number of paralogs (NP; Vilella et al. 2009) and PPI (table 2). Note that such anticorrelation is not observed if we use paralog counts derived from a simple Blastp search at an E value threshold of 10^{-10} (NPB; Altschul et al. 1997). Proteins have significantly higher number of paralogs according to NPB as opposed to NP (mean: 4.2 ± 13.4 vs. 1.1 ± 3.2; Mann–Whitney, Kolmogorov–Smirnov test [MW/KS-test]: $P < 6.5 \times 10^{-35}$). In particular, proteins having more than four interaction partners, that is, interacting with more proteins than the mean NP, have significantly more additional paralogs when NP is exchanged for NPB (mean difference: 5.5 ± 17.1 vs. 2.0 ± 8.5; MW/KS-test: $P < 2.1 \times 10^{-7}$). In other words, more hub-like proteins can align with more proteins when Blastp is used. It has been proposed that protein hubs experience more fitness constraints. Essential proteins have greater PPI in our data than nonessential proteins. From table 2, we also observe that there is a significant anticorrelation between NP and ESS. Thus, the anticorrelation between NP and PPI can be partially explained by assuming that there is more detrimental constraint on duplication if a protein is more hub like (Li et al. 2006). This would be the case if stoichiometric requirements exist for some of the genes. Presumably, certain genes from the yeast whole-genome duplication event (Kellis et al. 2004) have been lost due to fitness effects of dosage imbalance.

SD, PHD, and MW

A positive correlation between SD and PPI (table 2) supports the idea that prolific protein interactors tend to be more disordered (Haynes et al. 2006), although more subtle effects such as the differences in disorder between interactions of structurally different types (e.g., the number of binding interfaces) (Ekman et al. 2006; Kim et al. 2008) obviously cannot be captured in our analysis. The correlation between disorder and PPI may be in part due to the presence of signal integrators in the PPI network (Mittag et al. 2010). Known phosphorylation sites preferentially occur in disordered regions (Iakoucheva et al. 2004; Landry et al. 2009), and perhaps, this is why PHD is clustered together with SD and PPI in figure 3. It must be noted that in our work, we do not investigate individual phosphosites and their evolutionary conservation but rather use the number of phosphosites per protein. It has been shown that yeast phosphoproteins tend to have orthologs more frequently than the proteome average (Gnad et al. 2007), but we did not find a significant correlation between PHD and PGL. Many protein interaction hubs in yeast involve large multidomain proteins (Warringer and Blomberg 2006; Ekman et al. 2006) with a number of these proteins containing multiple interaction interfaces (Kim et al. 2006), and this may help explain why MW positively correlates (albeit quite weakly) with PPI. A weak but significant negative correlation between MW and expression related variables (CAI, EL, and ABU) presumably results in part from selective pressure to reduce overall costs associated with protein biosynthesis, homeostasis, and metabolism (Warringer and Blomberg 2006).

Disordered regions were observed to have elevated ERs for a variety of protein families (Brown et al. 2002; Landry et al. 2009), and we observe a significant positive correlation between the two variables in our data (table 2). SD is observed to be relatively close to ER along the first component. AGE is strongly anticorrelated with SD and PHD. Both SD (Xue et al. 2010) and phosphorylation-dependent signaling are much more prominent in eukaryotic organisms than in prokaryotes (Iakoucheva et al. 2004). PHD, however, is similarly low for proteins conserved in prokaryotes and eukaryotes (AGE = 1, SD = 0.002 ± 0.005; AGE = 2, SD = 0.002 ± 0.005; AGE = 3, SD = 0.001 ± 0.004) but is substantially higher for proteins found only in fungi (AGE = 0, PHD = 0.006 ± 0.014; MW/KS-test: $P < 0.05$). SD is similar for fungal-specific and eukaryote-specific proteins (AGE = 0, SD = 38.5 ± 22.7; AGE = 1, SD = 37.9 ± 23.1) but is significantly lower for more conserved proteins (AGE = 2, SD = 29.2 ± 20.9; AGE = 3, SD = 19.4 ± 17.5; MW/KS-test: $P < 0.05$). In fact, the anticorrelation between AGE and SD is the strongest in table 2.

Protein Half-Life and Isoelectric Point

The only point in figure 3 that does not make any noticeable contribution to both principal components is protein half-life (HL). The only two statistically significant correlations of HL are with PPI (positive) and with MW (negative) (Belle et al. 2006; Tompa et al. 2008), two variables which seem to offset each other along PC1. We were not able to confirm previous reports that unstructured proteins have shorter half-lives (Tompa et al. 2008) when DISOPRED was applied to our entire data set. We note, however, that in the data set of protein half-lives published by Belle et al. (2006), a cap of 300 min was introduced for many proteins

with very long half-lives, whereas many proteins were reported to have half-life values much higher than 300 min. Upon removing 511 entries with half-lives equal or greater than 300 min, we find a significant negative correlation between SD and half-life on our and disorder data from Gsponer et al. (2008); all modified correlations are given in the supplementary table S3*b*, Supplementary Material online. The resulting principal components do not differ much from the ones without removing the half-life entries (supplementary fig. S3*a*, Supplementary Material online), but PC1 does have a nontrivial contribution from half-life.

Finally, pI makes a strong contribution to PC2 with a strong anticorrelation with MW. Based on manual FunCat annotation, amongst the top 600 proteins (~10% of the yeast proteome) affecting this latter anticorrelation (see Materials and Methods), there is an enrichment of 107 ribosomal proteins (Bonferroni corrected hypergeometric $P < 2.5 \times 10^{-42}$) as well as an enrichment of 24 nuclear transport proteins (Bonferroni corrected hypergeometric $P < 0.002$), compared with the entire yeast proteome. The ribosomal proteins are significantly (MW/KS-test $P < 1.6 \times 10^{-29}$) smaller and have more basic amino acids (mean pI = 11.3 ± 1.1; mean MW = 20,462 ± 53,669) compared with the yeast proteome (mean pI = 7.5 ± 2.2; mean MW = 54,709 ± 43,418). The nuclear transport proteins, however, are larger and have more acidic amino acids (mean pI = 5.2 ± 1.4; mean MW = 142,734 ± 64,194) compared with the yeast proteome.

Conclusions

In this study, we have applied correlation analysis and Bayesian PCA to interpret 16 genomic variables from *S. cerevisiae*, gathered from the Quagmire database. Compared against standard PCA alone, Bayesian PCA allows one to treat input data with missing values under a probabilistic framework.

One can make hypotheses concerning what the principal components represent. Along the first component, the most striking feature is the opposite contribution between mutation (ER, PGL) and expression-related variables (CAI, EL, ABU, and TAI). Thus, PC1 can be interpreted largely as the direction where separation between genomic change and protein expression occurs. The second component consists of positive contributions concerning variables related to the presence or absence of genes or their protein products. AGE is related to the distribution of genes in genomes; PGL is proportional to the half-life of genes in genomes; EL, CAI, ABU, and HL are related to the concentration of the gene products in the cell. PI is related to localization (Schwartz et al. 2001) and thus a determinant of local gene product concentration. Negative contributions come from variables related to the interactions that they make (ESS, GI, PHD, NP, MW, and PPI). Thus, the second component might be interpreted as the direction where gene existence and their expressed protein functions, most importantly network status, are separated. These hypotheses are speculations for further research to pursue.

Like studies before, we found a negative correlation between pairs of variables ER–PPI, ER–AGE, PGL–ESS, PGL–PPI/GI, and ER–ABU and positive correlations between PPI–GI, PPI–ESS, PPI–SD, and SD–ER. New negative correlations were found between TAI–PGL, TAI–AGE, AGE–PHD, PI–MW, and AGE–SD; and a new positive correlation was found between TAI–ESS.

For pairs of genomic variables which have significant correlation, as illustrated by PI and MW, we applied a greedy method of gathering genes which affect the direction and magnitude of correlation the most. Subsequent enrichment analysis revealed functional categories of genes which changed the correlation the greatest when removed. In the illustration, ribosomal and nuclear transport proteins were demonstrated to be influential.

We also observed lack of correlation between certain genomic variables, such as PGL and AGE. By using enrichment analysis, we were able to determine that protein synthesis and amino acid metabolism genes obscure this correlation. Removal of these genes causes PGL and AGE to be significantly correlated.

It is evident that differences in data and definitions can result in different conclusions. For example, we observed a significant anticorrelation between the number of paralogs (NP) and both the number of interaction partners (PPI) and gene essentiality (ESS). Wolf et al. (2006) used ancient paralogs in that they analyzed the duplication of entire orthologous clusters present in seven eukaryotic species rather than individual genes. Neither the anticorrelation between NP and essentiality nor the anticorrelation between NP and PPI was observed in this case. For our data, proteins with more paralogs are longer (He and Zhang 2005) as found in (Wolf et al. 2006). NP is placed with ESS and situated opposite PGL along the second component, exactly as observed along the "gene status" component in (Wolf et al. 2006). Thus, analysis with different definitions of NP can result in similar but also different relations with other variables. Easily understanding what different data and definitions imply will help scientists navigate quagmires in the field. The methods of analysis used in this paper will help in this regard.

Supplementary Material

Supplementary table S3 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are indebted to Yuri Wolf for illuminating and fruitful discussions, to Mikhail Gelfand for suggesting several data sets to work on, and to Jürgen Cox for providing yeast phosphorylation data. We thank the systems biology group for help with PCA, Alexei Antonov for the idea of removing points from plots, and Martin Münsterkötter for the FUNCAT enrichment tool. Nadia Latif and Philip Wong acknowledge

the support of the German Academic Exchange Service and Helmholtz-center Munich, respectively. This research was partially supported by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (project CoReNe).

References

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*. 11:660–666.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.

Bader GD, Betel D, Hogue CW. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. 31:248–250.

Batada NN, Hurst LD, Tyers M. 2006. Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol*. 2:e88.

Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK. 2006. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci U S A*. 103:13004–13009.

Bloom JD, Adami C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol Biol*. 3:21.

Breitkreutz BJ, Stark C, Reguly T, et al. (12 co-authors). 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*. 36:D637–D640.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*. 55:104–110.

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. 2009. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*. 38:D532–D539.

Christie KR, Weng S, Balakrishnan R, et al. (23 co-authors). 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*. 32:D311–D314.

Costanzo M, Baryshnikova A, Bellay J, et al. (53 co-authors). 2010. The genetic landscape of a cell. *Science* 327:425–431.

Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC. 2005. Gene essentiality and the topology of protein interaction networks. *Proc R Soc B Biol Sci*. 272:1721–1725.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*. 32:5036–5044.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23:327–337.

Ekman D, Light S, Bjorklund AK, Elofsson A. 2006. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol*. 7:R45.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The Pfam protein families database. *Nucleic Acids Res*. 38:D211–D212.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296:750–752.

Gilchrist MA. 2007. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol*. 24:2362–2372.

Gnad F, de Godoy LM, Cox J, Neuhauser N, Ren S, Olsen JV, Mann M. 2009. High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* 9:4642–4652.

Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*. 8:R250.

Gout JF, Kahn D, Duret L. 2010. Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*. 6:e1000944.

Greenbaum D, Jansen R, Gerstein M. 2002. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* 18:585–596.

Gsponer J, Futschik ME, Teichmann SA, Babu MM. 2008. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 322:1365–1368.

Güldener U, Münsterkötter M, Kastenmüller G, et al. (20 co-authors). 2005. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res*. 33:D364–D368.

Güldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V. 2006. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*. 34:D436–D441.

Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*. 2:e100.

He X, Zhang J. 2005. Gene complexity and gene duplicability. *Curr Biol*. 15:1016–1021.

He X, Zhang J. 2006. Why do hubs tend to be essential in protein networks? *PLoS Genet*. 2:e88.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 32:1037–1049.

Ilin A, Raiko T. 2010. Practical approaches to principal component analysis in the presence of missing values. *J Mach Learn Res*. 11:1957–2000.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411:41–42.

Jolliffe IT. 2002. Principal component analysis, series: Springer series in statistics, 2nd ed. New York: Springer.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617–624.

Kerrien S, Alam-Faruque Y, Aranda B, et al. (24 co-authors). 2007. IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*. 35:D561–D565.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314:1938–1941.

Kim PM, Sboner A, Xia Y, Gerstein M. 2008. The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol*. 4:179.

Kim WK, Marcotte EM. 2008. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Comput Biol*. 4:e1000232.

Koonin EV, Fedorova ND, Jackson JD, et al. (18 co-authors). 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*. 5:R7.

Krylov D, Wolf Y, Rogozin I, Koonin E. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*. 13:2229.

Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet*. 25:193–197.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22:1345–1354.
Li L, Huang Y, Xia X, Sun Z. 2006. Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol*. 23:2467–2473.
Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet*. 39:415–421.
Mittag T, Kay LE, Forman-Kay JD. 2010. Protein dynamics and conformational disorder in molecular recognition. *J Mol Recognit*. 23:105–116.
Mittal N, Roy N, Babu MM, Janga SC. 2009. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc Natl Acad Sci U S A*. 106:20300–20305.
Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, Derisi JL, Weissman JS. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441:840–846.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.
Rocha EP. 2006. The quest for the universals of protein evolution. *Trends Genet*. 22:412–416.
Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol*. 21:108–116.
Ruepp A, Zollner A, Maier D, et al. (11 co-authors). 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*. 32(18):5539–5545.
Saeed R, Deane CM. 2006. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*. 7:128.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*. 32:D449–D451.
Schwartz R, Ting CS, King J. 2001. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res*. 11:703–709.
Sharp PM, Li WH. 1986. The codon adaptation index: a measure of directional synonymous codon usage, and its potential applications. *Nucleic Acids Res*. 15:1281–1295.
Shastry M, Nielsen J, Ku T, Hsu MJ, Liberator P, Anderson J, Schmatz D, Justice MC. 2001. Species-specific inhibition of fungal protein synthesis by sordarin: identification of a sordarin-specificity region in eukaryotic elongation factor 2. *Microbiology* 147(Pt 2):383–390.
Tompa P, Prilusky J, Silman I, Sussman JL. 2008. Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins* 71:903–909.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.
Vinogradov AE. 2010. Systemic factors dominate mammal protein evolution. *Proc R Soc B Biol Sci*. 277:1403–1408.
Visuri S, Koivunen V, Oja H. 2000. Sign and rank covariance matrices and rank covariance matrices. *J Stat Plan Inference*. 91:557–575.
Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*. 102:5483–5488.
Walter MC, Rattei T, Arnold R, et al. (14 co-authors). 2009. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res*. 37:D408–D411.
Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20:2138–2139.
Warringer J, Blomberg A. 2006. Evolutionary constraints on yeast protein size. *BMC Evol Biol*. 6:61.
Wolf YI. 2006. Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. *Trends Genet*. 22:354–357.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc R Soc B Biol Sci*. 273:1507–1515.
Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. *Genome Biol Evol*. 2:190–199.
Xia Y, Franzosa EA, Gerstein MB. 2009. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol*. 5:e1000413.
Xue B, Williams RW, Oldfield CJ, Dunker AK, Uversky VN. 2010. Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst Biol*. 4(Suppl 1):S1.
Yang D, Jiang Y, He F. 2009. An integrated view of the correlations between genomic and phenomic variables. *J Genet Genomics*. 36:645–651.
Zotenko E, Mestre J, O'Leary DP, Przytycka TM. 2008. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*. 4:e1000140.