A Novel Method to Detect Proteins Evolving at Correlated Rates: Identifying New Functional Relationships between Coevolving Proteins

Nathaniel L. Clark*,[1] and Charles F. Aquadro[1]

[1]Department of Molecular Biology and Genetics, Cornell University

***Corresponding author:** E-mail: nlc47@cornell.edu.

Associate editor: Rasmus Nielsen

Abstract

Interacting proteins evolve at correlated rates, possibly as the result of evolutionary pressures shared by functional groups and/or coevolution between interacting proteins. This evolutionary signature can be exploited to learn more about protein networks and to infer functional relationships between proteins on a genome-wide scale. Multiple methods have been introduced that detect correlated evolution using amino acid distances. One assumption made by these methods is that the neutral rate of nucleotide substitution is uniform over time; however, this is unlikely and such rate heterogeneity would adversely affect amino acid distance methods. We explored alternative methods that detect correlated rates using protein-coding nucleotide sequences in order to better estimate the rate of nonsynonymous substitution at each branch (d_N) normalized by the underlying synonymous substitution rate (d_S). Our novel likelihood method, which was robust to realistic simulation parameters, was tested on Drosophila nuclear pore proteins, which form a complex with well-documented physical interactions. The method revealed significantly correlated evolution between nuclear pore proteins, where members of a stable subcomplex showed stronger correlations compared with those proteins that interact transiently. Furthermore, our likelihood approach was better able to detect correlated evolution among closely related species than previous methods. Hence, these sequence-based methods are a complementary approach for detecting correlated evolution and could be applied genome-wide to provide candidate protein–protein interactions and functional group assignments using just coding sequences.

Key words: coevolution, correlated evolution, rate correlation, protein interactions, nuclear pore.

Introduction

The sequencing of complete genomes has greatly expanded the catalog of annotated genes in many organisms, and evolutionary information has been key to these annotations by revealing features that are conserved and presumed to be functionally important (Brent 2008). Our next challenge is to understand how these genes interact and which functions they influence. Biochemical and genetic screens can provide many predictions of genetic interactions; however, they are not guaranteed to reveal all interactions and can suffer from practical limitations. There is a need for alternative approaches that can prioritize potential interactions for experimental verification. The study of divergence can provide functional predictions by searching for those genes that evolve at correlated rates and hence are likely to be functionally linked via protein–protein interactions or by participating in a common function. Such an evolutionary approach is relatively fast and can take advantage of the expanding wealth of sequenced genomes.

Several lines of reasoning lead to the hypothesis that proteins participating together in a biological function will evolve at correlated rates. First of all, physically interacting proteins can influence each other's rates of divergence (Pazos et al. 1997). Although sequence conservation is a major mechanism maintaining protein interactions (Mintseris and Weng 2005), interaction interfaces diverge over time due to genetic drift and/or adaptation. In the face of such change, compensatory amino acid substitutions are required to maintain the interaction, a process we will refer to as intermolecular coevolution. Intermolecular coevolution could also be driving change in biological systems experiencing adaptive evolution. For instance, coevolution has been proposed to contribute to the rapid adaptive divergence of proteins mediating host–pathogen interactions and of reproductive proteins (Clark et al. 2006, 2009; Sawyer and Malik 2006). In such molecular "arms races," adaptive pressures would drive the initial divergence of the interface, whereas coevolution would drive subsequent compensatory changes, leaving a signature of correlated rates. Another reason to expect correlated evolution is that functionally related proteins are affected by shared evolutionary pressures, which either constrain or drive their divergence. Changes in the intensity of these shared pressures would result in parallel changes of evolutionary rate. Functionally related proteins could also evolve at correlated rates if their expression levels covary over time. One of the major determinants of protein evolutionary rate is expression level (Duret and Mouchiroud 2000; Pal et al. 2001; Subramanian and Kumar 2004; Drummond et al. 2005, 2006), and cofunctional proteins tend to have similar expression levels, especially in protein complexes (Eisen et al. 1998; Ge et al. 2001; Grigoriev 2001; Veitia et al.

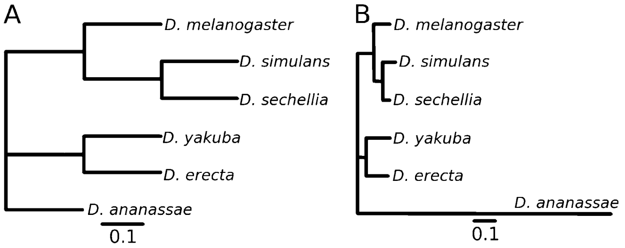

Fig. 1. Phylogenetic trees under which the "simple" data set (*A*) and "realistic" data set (*B*) were simulated. The simple tree was constructed with the same node-to-node distance for all branches; however, the total tree length matches the realistic tree. Branch lengths on the realistic tree were based on estimates from real alignments. Note the difference in scale between the two trees.

2008). Hence, shared changes in expression could lead to correlated evolutionary rates as previously proposed by Fraser et al. (2004) and Hakes et al. (2007).

In practice, rate correlations have been found between interacting protein pairs by measuring the correlation between all pairwise amino acid distances between species and selecting those protein pairs with the highest correlation coefficients (Goh et al. 2000; Pazos and Valencia 2001; Goh and Cohen 2002; Ramani and Marcotte 2003; Tan et al. 2004). The power of this "mirror tree" approach was later improved by factoring out the inherent similarity of distances due to the underlying species phylogeny (Fraser et al. 2004; Pazos et al. 2005; Sato et al. 2005; Shapiro and Alm 2008). More recently, further refinements including partial correlations have promised even more insight into the interrelatedness of proteins within pathways and complexes (Juan et al. 2008). All the above studies, except for Fraser et al., used amino acid substitution distances for detecting correlated rates. Although amino acid distances correspond directly to protein divergence, rates of amino acid substitution also depend on the underlying nucleotide mutation rate, especially for those amino acid substitutions that have minimal impact on evolutionary fitness (Kimura 1983; Ohta 1992).

Nucleotide mutation rates vary between genomic regions and can change over phylogenetic lineages (Ellegren et al. 2003; Singh et al. 2007, 2009), which introduces troublesome rate variation in tests for correlated amino acid distances. However, by instead comparing coding sequences, we can normalize the nonsynonymous (amino acid replacement) rate (d_N) using the roughly clocklike synonymous rate (d_S), which reflects the local nucleotide mutation rate. Studying correlation using the d_N/d_S ratio as such has three theoretical benefits: 1) Values on phylogenetic branches do not need to be adjusted for the underlying species phylogeny because d_N/d_S is a rate and not a distance. 2) The effect of variation in the local nucleotide mutation rate is controlled. 3) Multiple amino acid changes at the same codon are better estimated in a coding sequence model. In order to explore these potential benefits, we developed two new approaches to detect correlated evolution by comparing branch-specific values of d_N/d_S between protein-coding genes. We evaluated their power side by side with alternative mirror tree approaches to characterize their strengths and weaknesses on diverse data sets. Considering the power these approaches demonstrated, we applied them to a network of known interactions between *Drosophila* nuclear pore proteins (Nups), and the results supported the potential of correlated evolution methods to detect functional relationships between uncharacterized proteins.

Methods

Simulating Correlated Sets of Genes

We simulated coding sequence alignments using the program evolverNSbranches in the PAML4.0 package (Yang 2007). We used simulation parameters based on multiple alignments of six closely related *Drosophila* species (*Drosophila melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*). The "simple" simulations used the median gene length from 60 randomly chosen multiple alignments of *Drosophila* genes (850 codons). Equilibrium codon frequencies and the transition/transversion ratio were estimated by model M0 of codeml from a concatenation of the same 60 alignments. Simple phylogenetic branch lengths were scaled so that total tree length was equal to that estimated for the 60 alignments, although all branches were given the same length (fig. 1*A*). In the "realistic" simulations, gene lengths and equilibrium codon frequencies were randomly sampled from the 60 control alignments. Estimated gene trees were also sampled from these 60 alignments to simulate mutation rate variation in the realistic data set.

We created sets of branch d_N/d_S values at high and low levels of correlation. We first drew random values from a gamma distribution fit to genome-wide estimates of d_N/d_S in Drosophila (Larracuente et al. 2008). These random draws were then correlated to each other in groups of ten genes. For each data set, we generated five sets of ten genes, where each set had a different target level of correlation (ρs = 0, 0.25, 0.5, 0.75, or 0.95). When comparisons are made within sets, this yields a total of 225 comparisons over a wide range of simulated correlation coefficients. Correlation was performed as follows: Each set of random draws was placed in a vector, *R*, of length equal to the number of genes (10). There were nine such vectors, each corresponding to a branch of the species phylogeny. We then correlated these d_N/d_S vectors using Cholesky decomposition in which a matrix of desired correlation coefficients, *C*, was decomposed into matrix *U*. Then, each vector, *R*, of uncorrelated values was multiplied by *U* to yield correlated values. These branch d_N/d_S values were then assigned to each of the ten genes, and simulations were performed. Although this procedure produces correlations close to the target value, the "true" simulated correlation coefficient was determined for each gene pair by least squares regression. The same sets of correlated branch values were used for all simulated data sets.

Detecting Correlated Evolution

Point estimates of branch-specific d_N/d_S, d_N, and amino acid divergence were made by the branch model of codeml (Yang 2007). Pairwise amino acid distances were computed by protdist from the PHYLIP package (Felsenstein 1989). Point estimates were analyzed for correlation using Perl scripts written by N.L.C. Likelihood models were composed and optimized using custom scripts for HyPhy version 1.0: hypothesis testing using phylogenies (Pond et al. 2005). Public versions of the HyPhy scripts are available at http://mbg.cornell.edu/cals/mbg/research/aquadro-lab/software.cfm. Each likelihood model considers one gene pair at a time by using two data partitions, one for each of the two coding sequence alignments, to which were assigned the assumed species tree topology and the Goldman and Yang codon model (GY94) (Goldman and Yang 1994). Values of d_N and d_S were freely estimated in the free model but were constrained in the correlated model as follows for each phylogenetic branch "*i*":

$$d_N(\text{Gene1}, i) = \text{slope} \times d_N/d_S(\text{Gene2}, i) \times d_S(\text{Gene1}, i) + y \text{ intercept} \times d_S(\text{Gene1}, i),$$

where slope and *y* intercept are global parameters that define the correlation line. The correlated model was optimized from three different initial values of the slope parameter. To improve convergence of the correlated model, the following constraint employing the tangent function was used because we found that it was better able to explore the slope parameter space. In this case, the slope limits were set at positive and negative π/2. Maximum likelihoods and parameter estimates were not changed by this addition.

$$d_N(\text{Gene1}, i) = \text{tangent}(\text{slope}) \times d_N/d_S(\text{Gene2}, i) \times d_S(\text{Gene1}, i) + y \text{ intercept} \times d_S(\text{Gene1}, i).$$

In the null model, the slope parameter was set to zero. We set the HyPhy precision parameter to 1×10^{-5} to achieve better convergence. In practice, these models were optimized for Gene1 versus Gene2 and for Gene2 versus Gene1, as their outcomes differed. The more conservative outcome of the two comparisons (the lower test statistic) was used to represent the gene pair. Power analysis of the various methods was performed using the R package "ROCR" (Sing et al. 2005).

Drosophila Nuclear Pore Proteins

Coding sequence alignments were retrieved from the 12 Genomes project alignments (Drosophila 12 Genomes Consortium et al. 2007). We studied Nup 75, 96, 98, 107, 133, and 153. We drew 30 random control proteins from the 12 genomes data, and because these six Nup proteins show evidence of positive selection (Presgraves and Stephan 2007), we drew an additional 30 control proteins from a list of proteins inferred under positive selection at a false discovery rate of 5% by Larracuente et al. (2008). These positively selected genes were chosen to guard against an unexpected correlated signal simply because of genes under positive selection. Finally, if any control was annotated with a function in the nuclear pore or in the nucleus or to participate in nuclear transport, it was removed and replaced with another random draw because these could potentially be interacting with the nuclear pore.

In addition to our methods, we also analyzed the Nups data set using the pairwise amino acid distance methods of Pazos (tol_mirrortree) and Sato (ρ^{AVE}) (Pazos et al. 2005; Sato et al. 2005). For tol_mirrortree, we used the *rosy* gene to correct for the underlying species tree (Ko et al. 2003; Wong et al. 2007). The ρ^{AVE} method corrects for the underlying species tree using the average distances for all the proteins.

Results

Simulating Correlated Evolution

We conducted simulations to evaluate methods of detecting correlated evolution. We wanted to simulate conditions that would resemble six closely related *Drosophila* species that have sequenced genomes (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*). When estimating evolutionary rates from coding sequences, it is important that divergence at synonymous sites is not saturated to ensure accurate estimates of d_S. These six species were selected for this reason. The simulated species tree had an unrooted topology and total tree length as estimated for these species (Wong et al. 2007).

We controlled simulation parameters to create progressively more difficult and realistic data sets. The parameters we altered were the tree branch lengths, gene lengths, and the equilibrium codon frequencies. The simplest simulation, called simple, used tree branches of uniform length (fig. 1*A*) because such branches should allow good estimates of d_N and d_S. All genes in the simple data set had the same gene length (850 codons) and the same set of equilibrium codon frequencies. Because of its uniform simulation parameters, the simple data set should pose the least challenge to detect correlated evolution. A more difficult data set, realistic, was made to be more *Drosophila*-like by using parameter estimates from real *Drosophila* genes. These realistic genes were simulated over trees sampled from 60 randomly chosen *Drosophila* genes. These gene trees have differing branch lengths at corresponding branches and simulate rate variation between genes, which is crucial to test whether methods are robust to mutation rate variation. The consensus tree of these gene trees is shown in figure 1*B*. To introduce more *Drosophila*-like variation, genes in the realistic data set were also assigned lengths and equilibrium codon frequencies sampled from the 60 control genes. Variation of these parameters should pose a greater challenge to the methods. Finally, to test specific aspects of the methods, additional data sets were simulated by adding or removing sources of variation in

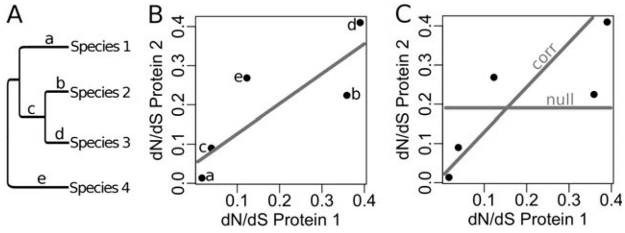

FIG. 2. Detecting correlated evolutionary rates using d_N/d_S ratios. (*A*) Example species tree over which evolutionary rates are estimated independently for each branch (labeled a–e). (*B*) d_N/d_S estimates for one protein are plotted versus another protein for each branch of the species phylogeny. A line of regression shows the relationship between them. (*C*) Similar logic is employed in the joint likelihood models except that the line of correlation ("corr") is optimized within the evolutionary model rather than using d_N/d_S estimates.

a stepwise fashion. The simple and realistic data sets represent the extremes of these simulations between which all others are nested.

We simulated correlated evolution between sets of genes by assigning them correlated branch d_N/d_S values. This was done using a correlation matrix and Cholesky decomposition on random gamma-distributed d_N/d_S values and assigning those values to branches (see Methods). Each data set contained five correlated groups of ten genes that were simulated at different degrees of correlated evolution. Comparisons within these groups provided 225 tests of correlated evolution over simulated correlation coefficients ranging from −0.6 to 0.99.

Detecting Correlated Evolution with d_N/d_S Point Estimates

Two proteins evolving in a correlated manner are predicted to have evolutionary rates that covary along the species phylogeny. This relationship can be visualized by plotting rates along each branch for one protein versus rates for the other protein (fig. 2*A* and *B*). In a similar manner, our first approach tested for correlated evolution by performing a linear regression on d_N/d_S point estimates. The strength of correlation between any pair of proteins was judged by the correlation coefficient (r) of their d_N/d_S branch values. When applied to the simple simulated data set, the point estimate method distinguished between strong and weak correlated evolution reasonably well (fig. 3*A*), and the agreement between simulated and estimated coefficients was good ($r = 0.76$). However, on the realistic data set, the point estimates method was much less able to infer the simulated degree of correlated evolution (fig. 3*B*) ($r = 0.13$). We also analyzed the power of this method by testing its ability to find true cases of correlated evolution. We designated those gene pairs with simulated correlation coefficients greater than 0.7 to be true positives and those below 0.3 to be true negatives. The d_N/d_S point estimate method identified true positives in the simple data set with very infrequent inclusion of false positives, as seen in its receiver operator characteristic (ROC) curve (fig. 4*A*, bold dotted line). By taking the area under the ROC curve (AUC), we obtain a measure of performance ranging from 0 to 1, for which 1 represents perfect discernment and 0.5 represents the expectation if the method has no predictive value. The d_N/d_S point estimate method did well for the simple data set (AUC = 0.98) but had more difficulty with the realistic data set (AUC = 0.75) (fig. 4*B* and table 1). After testing the point estimate method on additional simulations, we saw it that suffers when there are short

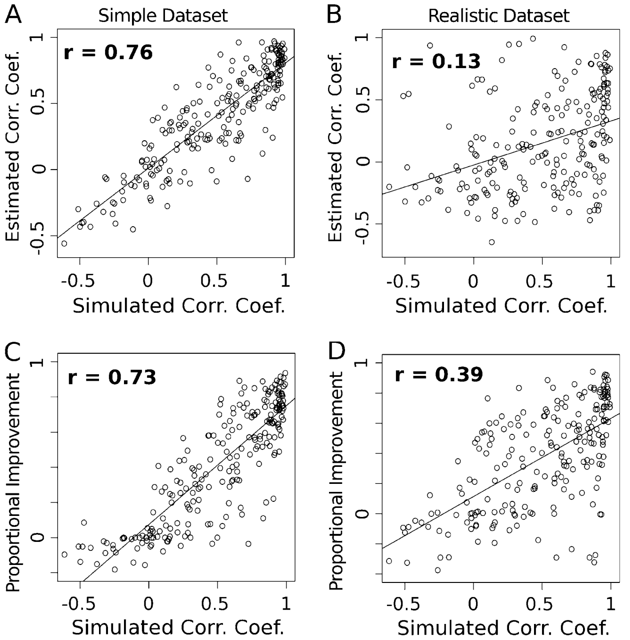

FIG. 3. Performance on simulated data sets. Each plot compares the correlation coefficient under which the data were simulated (simulated corr. coef.) to the output of a test for correlated evolution. (*A*) d_N/d_S point estimate method on "simple" simulated data set. (*B*) d_N/d_S point estimate on "realistic" simulated data set. (*C*) Joint likelihood method as "proportional improvement" on simple data set. (*D*) Joint likelihood method on realistic data set. Although both methods performed well on the simple data set, the joint likelihood method was less affected by the more realistic simulation parameters.

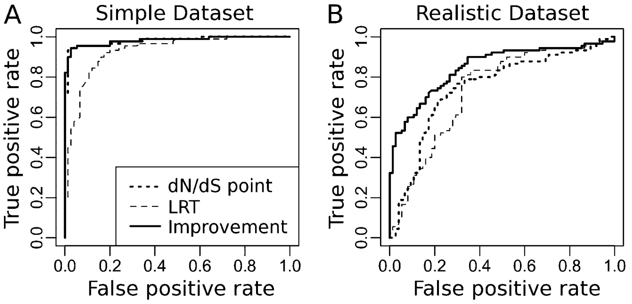

FIG. 4. Power analysis of d_N/d_S-based methods. Plots of the true-positive versus the false-positive rates (ROC curves) show the sensitivity and specificity of three different tests for correlated evolution on the "simple" (*A*) and "realistic" (*B*) simulated data sets. In (*A*), the d_N/d_S point estimates method and the proportional improvement statistic performed in a similar way so that their curves overlap.

Table 1. Performance of Various Methods to Detect Correlated Evolution.

Distances[a]	Statistic[b]	Simple Simulation[c]	Realistic Simulation[d]
Pairwise a.a.	*r*	0.84	0.76
Branch a.a.	*r*	0.99	0.74
Pairwise d_N	*r*	0.84	0.76
Branch d_N	*r*	0.99	0.76
Branch d_N/d_S	*r*	0.98	0.75
Branch d_N/d_S	LRT	0.93	0.71
Branch d_N/d_S	LRT per codon	NA[e]	0.73
Branch d_N/d_S	Improvement	0.98	0.85

Note.—a.a., amino acid distance; NA, not applicable.

[a] Measure of divergence used in method.

[b] Test statistic: *r*, correlation coefficient of corresponding distances; LRT, likelihood ratio test between "null" and "correlated" models; and improvement, proportional improvement.

[c] AUC curve of methods on "simple" simulated data set.

[d] AUC curve of methods on "realistic" simulated data set.

[e] Simple genes were of uniform size, so per codon correction was not applied.

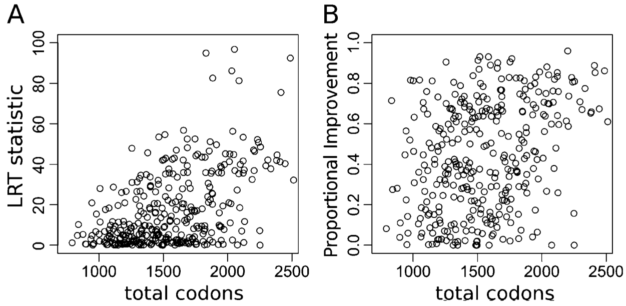

Fig. 5. The dependence of test statistics on gene size. (*A*) The LRT statistic is more dependent on the total length of the two genes being compared (*x* axis) than the "proportional improvement" statistic (*B*). In (*B*), the gene size is plotted against the absolute value of proportional improvement. These comparisons are between *Drosophila* nuclear pore proteins and control proteins that are not expected to interact.

branches in the species phylogeny; the area under the curve declined from 0.98 to 0.87 when *Drosophila*-like branch lengths were used (supplementary table S1, Supplementary Material online, simulation 1 vs. 4). Yet introducing variable equilibrium codon frequencies or gene length did not affect the area under the curve, so the d_N/d_S point estimate method is robust to these sources of variation (supplementary table S1, Supplementary Material online, simulation 1 vs. 2 and 3). We suspect that the method is sensitive to short branch lengths because their d_N/d_S estimates are uncertain, which can lead to spurious correlations or degrade a true correlated signature. Generally, realistic branch lengths present problems with uncertainty and variable information content. To overcome these limitations, we next created a method that could deal with uncertainty in a probabilistic way and which could evaluate correlated evolution within an evolutionary model.

Detecting Correlated Evolution with Joint Likelihood Models

Previous researchers have developed powerful likelihood codon models upon which we built a method to detect correlated evolution (Goldman and Yang 1994; Muse and Gaut 1994). We created a novel set of models to evaluate correlated evolution jointly between the two genes, which we predicted would better handle realistic data sets. We used them to infer correlation entirely within a model of coding sequence evolution. We describe the three models starting with the most general. The parameter-rich free model simultaneously estimates a d_N/d_S ratio for each branch of both gene trees, so that no relationship is modeled. Next, the correlated model evaluates the strength of a correlated relationship between the two coding sequences. It constrains their branch d_N/d_S values to a linear relationship using two global parameters, representing the slope and intercept of the correlation line (fig. 2C, "corr"). It is important to note that the estimated correlation line can differ from that of the point estimate method because the correlated model weighs branches according to an evolutionary model (as seen in our didactic example, fig. 2B and C). The simplest model, null, is nested within the correlated model, so that the hypothesis of a correlation can be tested. The null model is also a linear model; however, the slope parameter is fixed at zero. In effect, the null model is used to test whether the slope parameter of the correlated model is different from zero. In other words, their comparison tests whether there is value in knowing the d_N/d_S ratios of one protein to predict the ratios of the other.

Because the null model is nested within the correlated model, an option to test for correlation is a likelihood ratio test (LRT). In practice, we calculated the LRT for gene1 versus gene2 and vice versa using the lower more conservative value to represent the gene pair. An LRT is useful for a single pair of genes; however, it is not ideal for comparing the strength of correlation between various pairs of genes. This is because the LRT statistic depends on the length of the genes considered (fig. 5A), so that two small highly correlated genes could score less than two large weakly correlated genes. This bias is confounding because we want to compare correlations between genes of varying sizes, and we are more interested in the degree of correlation than in statistical power to detect it. We tested two methods to correct for this gene size effect. The first was to normalize the LRT for gene size by dividing by the total number of codons in both genes (LRT per codon). LRT per codon effectively removed the relationship of LRT with gene size. The second method was a new test statistic that uses the relationship between all three models and is based on the following observation. The correlated model can never fit the data better than the free model because the latter is more general and parameter rich. However, under the case of perfect correlation ($r = 1$), they fit the data equally well. That is, their likelihoods are equal. For example, the free and correlated models optimize to the same likelihood value when a gene is tested for correlation with itself. Because of this relationship, the free model serves as an upper bound for the potential improvement of the correlated model over the null model. Our new statistic reflects the proportional improvement of the correlated model over the null relative to the maximum possible improvement, which is the free over the null. We call this statistic the "proportional improvement," and it is calculated as the difference in log likelihood between the correlated and null models divided by the difference between the free and null models. Calculated as such proportional improvement can range from 0 to 1. However, when the estimated correlation line has a negative slope, "improvement" is assigned a negative value, so that it reflects the sign of the inferred correlation. As before, we compared gene1 and gene2 and vice versa using the value of lower magnitude. When we applied proportional improvement to the data, there was no longer a relationship with gene size (fig. 5*B*). Because both "LRT per codon" and proportional improvement corrected for gene size, we proceeded with both of these statistics to compare correlated evolution between different gene pairs.

When applied to the simple simulated data set, the joint likelihood models distinguished well between correlated and uncorrelated simulations using the proportional improvement statistic (fig. 3*C*) and agreed well with simulated values in general ($r = 0.73$). It also inferred a negative slope for most of the gene pairs with a negative simulated correlation coefficient (fig. 3*C*). In power analysis, improvement scored very well on the simple data set (AUC = 0.98) (fig. 4*A*), whereas the LRT statistic had less power (AUC = 0.93). It is interesting that improvement scored better than that the LRT on the simple data set because those genes were of uniform size. Therefore, there may be a general benefit to comparing all three models instead of just the correlated and null models. In simulations with variable gene size, the LRT per codon statistic performed better than LRT alone (table 1, Realistic Simulation); however, it never performed better than the improvement statistic over seven different simulation conditions (table 1 and supplementary table S1, Supplementary Material online). Hence, we chose improvement as the statistic for the joint likelihood method. However, due to the unconventional nature of the improvement statistic, we also applied the LRT per codon correction to all analyses for comparison.

A difference between the joint likelihood and point estimate d_N/d_S methods became apparent with the realistic data set, in which the likelihood method was more robust to variable, *Drosophila*-like parameters (fig. 3*D*). The likelihood method maintained a moderate correlation with the realistic data set ($r = 0.39$), whereas the point estimate method was more strongly affected ($r = 0.13$). Similarly, the joint likelihood method could identify true positives better than the point estimate method (AUC = 0.85 vs. 0.75) (fig. 4*B*). According to these tests, the likelihood method deals well with biologically realistic parameters and can detect correlated evolution using as few as six species.

Comparing Several Methods on a Common Set of Simulations

We analyzed additional simulated data sets to learn how various evolutionary parameters affected our d_N/d_S methods and other simpler methods (supplementary table S1, Supplementary Material online). We evaluated the performance of the d_N/d_S point estimate method, the d_N/d_S likelihood method, and multiple mirror tree distance methods. The d_N/d_S point estimate method used the correlation coefficient (r) as above. The mirror tree methods are distance methods that measure the correlation coefficient between genes using amino acid distances (Goh et al. 2000; Pazos and Valencia 2001; Goh and Cohen 2002; Ramani and Marcotte 2003; Tan et al. 2004; Pazos et al. 2005; Sato et al. 2005). We also created and evaluated novel mirror tree methods that use the nonsynonymous distance, d_N, so that finally, we employed four different distance measures: pairwise amino acid distances, branch amino acid distances, pairwise d_N, and branch d_N. For the likelihood method, we evaluated three different statistics: LRT, LRT per codon, and proportional improvement. After comparing all these methods, we discerned three main points: 1) It is often better to compare branch lengths instead of pairwise distances between species. This benefit was strongest for simulation 1 in supplementary tableS1, Supplementary Material online, for which branch values outperformed pairwise comparisons (AUC = 0.99 vs. 0.84). The benefit was the same whether d_N or amino acid distances were used. 2) For mirror tree methods (i.e., those using point estimates of distances) using d_N was only a slight improvement over amino acid distances, as seen in simulations 1, 3, 4, 6, and 7 (supplementary table S1, Supplementary Material online). Similarly, point estimates of d_N/d_S were not an improvement over using nonnormalized distances. 3) Considering all methods, the d_N/d_S likelihood method employing proportional improvement was the most robust and powerful under the most realistic parameters tested, including variation in mutation rate (simulation 7, the realistic data set). The realistic data set genes were simulated over an empirical sample of estimated gene trees, which were meant to mimic local mutation rate variation. Because the sampled gene trees were from actual *Drosophila* genes, they should not have exaggerated the intended effect. Such rate variation adversely affected all methods, but the proportional improvement statistic performed best under these conditions; its AUC was 0.85 compared with the next highest score of 0.76 (supplementary table S1, Supplementary Material online).

Demonstration that Drosophila Nuclear Pore Proteins Evolve at Correlated Rates

The nuclear pore is a large structure that controls traffic between the cytoplasm and the nucleus. It spans the nuclear envelope and is composed of large complexes of proteins many of which are Nucleoporins (Nup). In *Drosophila*, divergence at *Nup* genes is known to contribute to genetic incompatibility between sister species, and this divergence was attributed to positive selection (Presgraves et al. 2003; Presgraves and Stephan 2007; Tang and Presgraves 2009). Because of these observations—as well as their documented physical interactions—several Nup proteins were

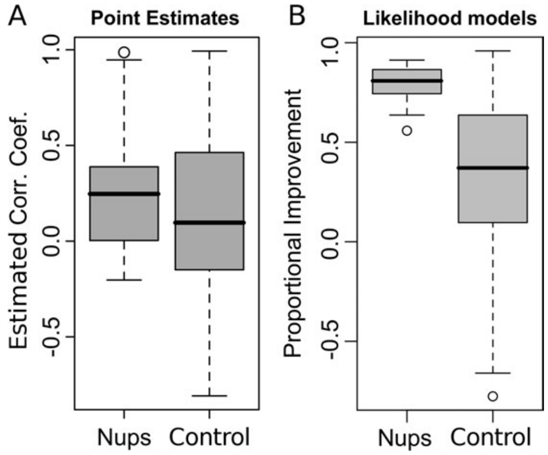

Fig. 6. Boxplots show the distribution of test statistics for Nup–Nup comparisons (Nups) versus Nup control comparisons (control). For (*A*), the test statistic is the correlation coefficient "*r*" on d_N/d_S point estimates. For (*B*), the statistic is "proportional improvement" inferred from the joint likelihood models. The bold line represents the median value and box limits are the upper and lower quartiles. The whiskers extend to the most extreme data point outside the box that is no more than 1.5 times the interquartile range. Any data points more extreme are plotted as a circle.

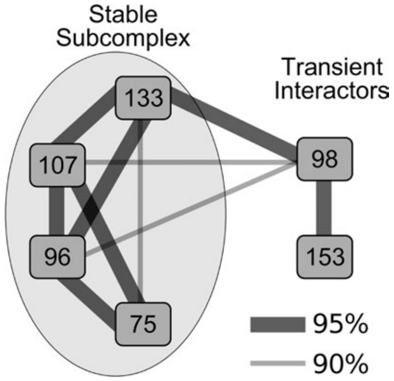

Fig. 7. Correlations between specific nuclear pore protein pairs. Each node is a Nup protein identified by its Nup number. Each edge represents the rate correlation between those two proteins with its width reflecting the empirical *P* value (either 95% or 90%). Nups 75, 96, 107, and 133 participate in the stable Nup 107 subcomplex, whereas Nups 98 and 153 are transient interactors.

hypothesized to coevolve and hence may have diverged at correlated rates (Presgraves and Stephan 2007). There are many proteins in the nuclear pore from which we chose six physically interacting and positively selected proteins as a test case for our methods (Nup 75, 96, 98, 107, 133, and 153). Their physical interactions have been demonstrated by pull down, coexpression, and affinity chromatography experiments (Vasu et al. 2001; Lutzmann et al. 2002). We then randomly chose 60 negative control genes that were not expected to coevolve with the nuclear pore. Because the six chosen Nups have evolved under positive selection, half of the control genes were chosen because they also showed significant positive selection. This measure was to guard against a potential general correlation between positively selected genes. All 15 pairwise comparisons were made between Nups, and each Nup was tested against all 60 control genes.

The two d_N/d_S methods performed differently on the Nup data set. The point estimate method found mostly positive correlation coefficients for the Nup versus Nup comparisons (Nup–Nup) as one would predict, whereas the noninteracting comparisons (Nup control) ranged from approximately −0.8 to 1 (fig. 6*A*). Although two Nup–Nup comparisons were highly correlated (r = 0.986 and 0.947), point estimates did not demonstrate a general correlation between Nup proteins because the Nup–Nup distribution was not significantly different from the Nup control distribution (Wilcoxon rank sum test, P = 0.2897). In contrast, the joint likelihood approach revealed global evidence for correlated evolution between Nup proteins (fig. 6*B*). There was a highly significant difference between Nup–Nup and Nup control comparisons using the proportional improvement and LRT per codon statistics (Wilcoxon rank sum test, $P = 1.1 \times 10^{-7}$ and $P = 5.8 \times 10^{-7}$, respectively). Because these six Nup proteins have evolved under positive selection, it was important to verify that the correlated signal was not an unexpected result of comparisons between positively selected genes. When the global test was performed using just the 30 positively selected controls, the Nup–Nup correlation remained highly significant (proportional improvement, $P = 1.6 \times 10^{-6}$; LRT per codon, $P = 1.8 \times 10^{-5}$).

Given the global evidence for correlation between Nup proteins, we next looked for biological insight among the individual Nup–Nup comparisons. We set significance thresholds for individual tests using the empirical Nup control distribution of proportional improvement. Of the 15 Nup–Nup comparisons, 7 were significant at a 95% significance threshold, and 10 were significant at a 90% threshold (fig. 7). The most striking observation was that the strongest correlations were within a stable subcomplex of the nuclear pore, the Nup 107 subcomplex (fig. 7) (Tran and Wente 2006). Within the subcomplex, five of six comparisons were significant at a 95% threshold. In contrast, correlations between the transiently interacting proteins and the subcomplex were weaker or absent; only one of eight comparisons was significant at a 95% threshold.

Discussion

Interacting protein pairs and cofunctional proteins are predicted to evolve in a correlated manner, and there is need for rigorous methods to quantitatively test such hypotheses. We explored two methods employing d_N/d_S ratios to detect correlated evolution. The first simple method used conventional correlation statistics on d_N/d_S point estimates. We found that this approach performed reasonably well when the underlying species tree had uniform branch lengths; however, it suffered when realistic branch lengths were used. Perhaps this particular set of *Drosophila* species was not ideal for the point estimate method because there were many short branches, which produce uncertain d_N/d_S estimates. We expect that if species were chosen so that the phylogeny had few short branches, the point estimates method would be useful because it is computationally fast. However, our joint likelihood method performed better under more realistic simulation conditions. Because the evolutionary model is part of the inference process, this method can better handle uncertainty in d_N/d_S estimates.

One complicating factor when comparing correlations among gene pairs is that genes will be of different sizes. If we simply compare the correlated and null models, this variation introduces a bias toward larger genes because the greater number of substitutions provides more power. To correct for this bias, we introduced a relative statistic, proportional improvement, which should be more useful for making comparisons between different gene pairs. This statistic could be thought of as simply analogous to a correlation coefficient because it reflects the degree and sign of correlation without respect to its statistical significance. Proportional improvement cannot be used for model-based hypothesis testing as is often done with likelihood models; however, by employing an empirical null distribution, it is useful for making comparisons between genes of varying size. In future applications, it will be important to generate an empirical distribution for each new set of species because it will likely be different. Also, the results from our simulated data sets suggest that more power would be gained by choosing species that produce a well-resolved tree with few short branches.

One must also be aware of the method's assumption that changes at synonymous sites are neutral. Given the evidence for negative selection on synonymous sites, this is likely to be violated (e.g., Chamary et al. 2006). However, such selection would only affect a correlation test if it were to fluctuate over time and if it did so variably between the two genes. There is evidence of positive selection at synonymous sites leading to drastic changes in codon usage (Singh et al. 2007; Bauer DuMont et al. 2009). Because such episodes could mislead a correlated evolution study, one should rule them out using an appropriate method such as those developed by DuMont et al. (2004) or Nielsen et al. (2007). In our study, none of the six *Nup* genes showed evidence of positive selection at synonymous sites on the *D. melanogaster* or *D. sechellia* lineages (Singh et al. 2007). Finally, the likelihood method using proportional improvement was the most robust to mutation rate variation (table 1), which is crucial for application to real sequence alignments. Yet, because proportional improvement is novel, we recommend parallel application of the LRT per codon correction.

When we applied our likelihood method to data from six *Drosophila* species, it revealed a signature of correlation between nuclear pore proteins, supporting the hypothesis of Presgraves and Stephan that these protein have been evolving adaptively partly due to coevolution (Presgraves and Stephan 2007). There are multiple lines of evidence that suggest this signature reflects biological reality. Our likelihood models showed that the proportional improvement statistic was significantly higher among the Nups as a class when compared with randomly chosen control proteins. In addition, each of these Nup proteins shared a signature of positive selection that could result from a shared selective pressure and/or coevolution among them (Presgraves and Stephan 2007). Furthermore, the signature of correlated evolution was more pronounced between members of the Nup 107 subcomplex, which is expected if consistent physical interaction and/or stoichiometric balance (i.e., co-varying expression levels) are important factors (Pazos et al. 1997; Fraser et al. 2004; Hakes et al. 2007). Our evidence of coevolution between nuclear pore proteins is of evolutionary interest because divergence at two *Nup* genes has been shown to contribute to hybrid incompatibility between *D. melanogaster* and *D. simulans* (Presgraves et al. 2003; Tang and Presgraves 2009). Such incompatibilities could have arisen over the millions of generations that the nuclear pore has been evolving independently in each species. Hence, coevolution could be a major force creating incompatibilities between closely related species and populations (Coyne and Orr 2004; Dobzhansky 1937). In general, detecting correlated rates is interesting for evolutionary studies because it provides a rigorous test for hypotheses of coevolution, which are often evoked to explain cases of natural selection (Clark et al. 2006; Sawyer and Malik 2006).

For comparison with our d_N/d_S methods, we also applied distance methods to the simulated data sets. One important observation was that using branch distances rather than pairwise distances can be more powerful to detect correlated evolution. Also, for distance methods, d_N is not necessarily an improvement over amino acid distances, as we saw only a small and inconsistent advantage to d_N alone. Finally, point estimates of d_N/d_S were not able to correct for mutation rate variation in a satisfactory way. Rather, the joint likelihood method was required to maintain power in the face of the rate variation modeled here.

We also tested two published mirror tree methods on the nuclear pore data set, tol_mirrortree and ρ^{AVE} (Pazos et al. 2005; Sato et al. 2005). These amino acid distance methods correct for the underlying species tree by using a reference protein or set of reference proteins to represent general species divergence, whereas our d_N/d_S methods use the gene-specific synonymous distance on each branch. For power analysis, we made the assumption that all Nup–Nup comparisons were true positives and all Nup controls were true negatives, although this is not necessarily true. Both mirror tree methods also detected significantly higher correlation between Nup–Nup comparisons compared with controls (Wilcoxon rank sum test, tol_mirrortree $P = 0.0095$ and ρ^{AVE} $P = 1.5 \times 10^{-5}$). However, the d_N/d_S joint likelihood method demonstrated the highest power on the Nup data set (AUC: 0.91, 0.58, 0.70, and 0.83 for the d_N/d_S likelihood, d_N/d_S point estimate, tol_mirrortree, and ρ^{AVE} methods, respectively). Hence, there seems to be a benefit to using nucleotide sequences and d_N/d_S ratios to follow correlated evolution.

To make a comprehensive assessment of the d_N/d_S and distance methods, it helps to consider the types of data being analyzed. The closely related *Drosophila* species used here were chosen so that synonymous substitutions were not saturated, as required for d_N/d_S methods. In contrast, the mirror tree distance methods were designed to work on large sets of species that cover much longer evolutionary distances. We suggest that studies concerned with correlated evolution among closely related species would benefit from the additional power provided by the d_N/d_S joint likelihood method. This would be the case when examining a genetic pathway specific to a taxonomic group or to exploit a cluster of closely related genome sequences. On the other hand, when distances are long enough to saturate synonymous sites, the amino acid distance methods are the only suitable approach. This might be the case if one wants to find correlations among pathways that are conserved over long time periods and for which the addition of divergent species would be beneficial. Overall, we see the d_N/d_S and mirror tree distance methods as complementary and choosing between them depends on which biological question is being asked.

We have demonstrated correlated evolutionary rates using a novel model-based approach that can now be applied to make functional predictions for less characterized proteins. These predictions could do much to prioritize and direct experimental efforts hence making them more efficient. The potential for such an approach is great as the number of closely related genome sequences continues to grow. Historically, the factors that influence evolutionary rate have been of much interest to biologists, and now the discovery of rate correlations promises even more insight into the processes that influence the evolution of entire genetic networks. In turn, these insights could provide molecular biologists with a wealth of functional information.

Supplementary Material

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

Acknowledgments

We thank Rasmus Nielsen and two anonymous reviewers for helpful comments and suggestions. This work was supported by a National Institutes of Health postdoctoral fellowship (GM084592) to N.L.C. and a National Institutes of Health grant (GM36431) to C.F.A.

References

Bauer DuMont VL, Singh ND, Wright MH, Aquadro CF. 2009. Locus-specific decoupling of base composition evolution at synonymous sites and introns along the *Drosophila melanogaster* and *Drosophila sechellia* lineages. *Genome Biol Evol.* 1:67.

Brent MR. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet.* 9:62–73.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7:98–108.

Clark NL, Aagaard JE, Swanson WJ. 2006. Evolution of reproductive proteins from animals and plants. *Reproduction* 131:11–22.

Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ. 2009. Coevolution of interacting fertilization proteins. *PLoS Genet.* 5:e1000570.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Dobzhansky TG. 1937. Genetics and the origin of species. New York: Columbia University Press.

Drosophila 12 Genomes Consortium. Clark AG, Eisen MB, et al. (417 co-authors). 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA.* 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.

DuMont VB, Fay JC, Calabrese PP, Aquadro CF. 2004. DNA variability and divergence at the notch locus in Drosophila melanogaster and D. simulans: a case of accelerated synonymous site divergence. *Genetics* 167:171–185.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol.* 17:68–74.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA.* 95:14863–14868.

Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev.* 13: 562–568.

Felsenstein J. 1989. PHYLIP–Phylogeny Inference Package Version 3.6. *Cladistics.* 5:164–166.

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA.* 101:9033–9038.

Ge H, Liu Z, Church GM, Vidal M. 2001. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet.* 29:482–486.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol.* 299:283–293.

Goh CS, Cohen FE. 2002. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol.* 324:177–192.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.

Grigoriev A. 2001. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res.* 29:3513–3519.

Hakes L, Lovell SC, Oliver SG, Robertson DL. 2007. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci USA.* 104: 7999–8004.

Juan D, Pazos F, Valencia A. 2008. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA.* 105:934–939.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, Cambridgeshire: Cambridge University Press.

Ko WY, David RM, Akashi H. 2003. Molecular phylogeny of the Drosophila melanogaster species subgroup. *J Mol Evol.* 57: 562–573.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet.* 24:114–123.

Lutzmann M, Kunze R, Buerer A, Aebi U, Hurt E. 2002. Modular self-assembly of a Y-shaped multiprotein complex from seven nucleoporins. *EMBO J.* 21:387–397.

Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci USA.* 102:10930–10935.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11:715–724.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Evol.* 24:228–235.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst.* 23:263.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. *J Mol Biol.* 271:511–523.

Pazos F, Ranea JA, Juan D, Sternberg MJ. 2005. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol.* 352:1002–1015.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14:609–614.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. *Nature* 423:715–719.

Presgraves DC, Stephan W. 2007. Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. *Mol Biol Evol.* 24:306–314.

Ramani AK, Marcotte EM. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol.* 327:273–284.

Sato T, Yamanishi Y, Kanehisa M, Toh H. 2005. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* 21:3482–3489.

Sawyer SL, Malik HS. 2006. Positive selection of yeast non-homologous end-joining genes and a retrotransposon conflict hypothesis. *Proc Natl Acad Sci USA.* 103:17614–17619.

Shapiro BJ, Alm EJ. 2008. Comparing patterns of natural selection across species using selective signatures. *PLoS Genet.* 4:e23.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21:3940–3941.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in Drosophila. *Mol Biol Evol.* 26:1591–1605.

Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. 2007. Patterns of mutation and selection at synonymous sites in Drosophila. *Mol Biol Evol.* 24:2687–2697.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168:373–381.

Tan SH, Zhang Z, Ng SK. 2004. ADVICE: automated detection and validation of interaction by co-evolution. *Nucleic Acids Res.* 32:W69–W72.

Tang S, Presgraves DC. 2009. Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323:779–782.

Tran EJ, Wente SR. 2006. Dynamic nuclear pore complexes: life on the edge. *Cell* 125:1041–1053.

Vasu S, Shah S, Orjalo A, Park M, Fischer WH, Forbes DJ. 2001. Novel vertebrate nucleoporins Nup133 and Nup160 play a role in mRNA export. *J. Cell Biol.* 155:339–354.

Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* 24:390–397.

Wong A, Jensen JD, Pool JE, Aquadro CF. 2007. Phylogenetic incongruence in the Drosophila melanogaster species group. *Mol Phylogenet Evol.* 43:1138–1150.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.