[Fly 2:6, 291-299; November/December 2008]; ©2008 Landes Bioscience

Research Paper

Using whole genome presence/absence data to untangle function in 12 Drosophila genomes

Jeffrey A. Rosenfeld,[1] Rob DeSalle,[2,*] Ernest K. Lee[3] and Patrick O'Grady[4]

[1]Department of Biology; New York University; New York, New York USA; [2]Sackler Institute for Comparative Genomics; American Museum of Natural History; USA; [3]Center for Genomics and Systems Biology; Department of Biology; New York University; New York, New York USA; [4]Department of Environmental Science; Policy and Management; University of California; Berkeley, California USA

Key words: Drosophila, phylogeny, presence/absence, function, GO categories

The Drosophila 12 genome data set was used to construct whole genome, gene family presence/absence matrices using a broad range of E value cutoffs as criteria for gene family inclusion. The various matrices generated behave differently in phylogenetic analyses as a function of the e-value employed. Based on an optimality criterion that maximizes internal corroboration of information, we show that values of e^{-105} to e^{-125} extract the most internally consistent phylogenetic signal. Functional class of most genes and gene families can be accurately determined based on the *D. melanogaster* genome annotation. We used the gene ontology (GO) system to create partitions based on gene function. Several measures of phylogenetic congruence (diagnosis, consistency, partitioned support, hidden support) for different higher and lower level GO categories, were used to mine the data set for genes and gene families that show strong agreement or disagreement with the overall combined phylogenetic hypothesis. We propose that measures of phylogenetic congruence can be used as criteria to identify loci with related GO terms that have a significant impact on cladogenesis.

Introduction

Recently, high quality genome sequences have been produced and annotated for 12 species in the genus Drosophila.[1] A phylogenetic hypothesis for the 12 fully sequenced species was generated based on an analysis of sequences from over 180 thousand protein coding loci.[2] Questions concerning gene expression and function,[3] adaptation to host and habitat niches,[4] and genome evolution[5-9] can now be addressed in Drosophila with unprecedented levels of detail. The Drosophila 12 genome phylogenetic hypothesis[1,2] is highly robust and corroborates the accepted relationships based on earlier studies with more comprehensive taxon sampling.[10-16] While the phylogenetic results of these studies are not in question, more exploration of the data set may yield insight into genome evolution and function and provide a useful model with which to study the behavior of numerous analytical methods commonly used in phylogenetic analysis.[17] While the majority of whole genome analyses to date use DNA or amino acid sequence as character information (reviewed in ref. 2), we present an analysis of gene presence/absence characters to infer phylogenetic relationships and to explore gene function in a phylogenomic context.

In the current communication we (1) assess the role of similarity and homology assessment within gene families on phylogenetic reconstruction, (2) evaluate whether presence/absence matrices are consistent with phylogenetic hypotheses generated with DNA and amino acid sequence data and (3) place the functional characterization of proteins in an evolutionary context using a novel phylogenetic approach.

*Correspondence to: Rob DeSalle; American Museum of Natural History; Sackler Institute; 79th Street at Central Park West; New York, New York 10024 USA; Tel.: 212.769.5670; Email: desalle@amnh.org

Submitted: 07/15/08; Revised: 11/18/08; Accepted: 11/24/08

Previously published online as a *Fly* E-publication:
http://www.landesbioscience.com/journals/fly/article/7481

Results and Discussion

The impact of E value cutoff on phylogenetic hypothesis. Our approach does not attempt to establish orthology of individual members of gene families and instead simply counts the presence of a single member of a gene family as present for that gene family. In this context, the cutoff value for making a statement about inclusion of a gene in a gene family will affect the number of gene families found. Figure 2A shows the distribution of number of gene families as a function of E value cutoff, used to establish gene family membership. This pattern is consistent with other studies that have examined this problem.[18,19] Figure 2B shows the distribution of consensus fork indices (cfi) as a function of E value cutoff. This figure was produced by generating presence/absence matrices for a range of E values, generating a phylogenetic tree from the matrices using the consensus tree and the 85% bootstrap tree and comparing the tree topology to the topology obtained from previous phylogenetic analyses. It shows that for the Drosophila 12 genome data set, the phylogenetic hypothesis is very stable for the consensus trees because values from e^{-35} to e^{-235} give the accepted tree.[1,2,20]

The bootstrap and consensus patterns are different because the bootstrap trees are less resolved due to more stringent requirements for a node to be retained in a resultant tree. The booststrap tree cfi comparison does suggest that one E value may be more reliable than the others (e^{-105}) so consequently we use the matrix generated from this E value in all analyses where GO categories are used to further partition the data set. Figure 2C shows the distribution of

CCMR (Combined Corroboration Metric Retention Index[18]). The CCMR is simply the cfi scaled by the retention index of the tree. This figure shows the same general pattern as Figure 2B, although examination of values on the plateau at higher resolution indicates a maximum of the CCMR is observed for values e^{-105}, e^{-115} and e^{-125}. This result corroborates the choice of e^{-105} as an optimal value with which to conduct further functional analyses.

Trees derived from individual matrices generated by a range of E value cutoffs (e^{-5} to e^{-300}) yield five different topologies, all of which differ in the placement of the five *melanogaster* subgroup taxa (Suppl. Fig. 1). This analysis shows that at the extreme E value cutoffs, well-supported alternative hypotheses are generated.

The simple explanation for this pattern is that the matrices generated by BLAT using non-stringent E value cutoffs (e^{-5} to e^{-50}) are comprised of larger numbers of gene families relative to other E value cutoffs but the orthology assessment of these gene families is less reliable because of the relaxed cutoff. The matrices where E values are extremely stringent (e^{-200} to e^{-300}) are comprised of fewer gene families with very strong orthology assessment. Hence, the high E value matrices (e^{-5} to e^{-50}) will have more homoplasy in them than the lower E value matrices (e^{-200} to e^{-300}), but the low E value matrices will have fewer gene families and hence lower resolution. For all subsequent analyses we use the e^{-105} matrix which generates the tree in Figure 3. All nodes in this tree are extremely well supported (100% bootstrap at 90% replacement and posterior Bayesian probabilities at all nodes equal to 1.0) except for the node defining the species pair *yakuba* + *erecta*, which shows a bootstrap of 100% at 50% character replacement but shows a bootstrap of 75% with 90% character replacement.

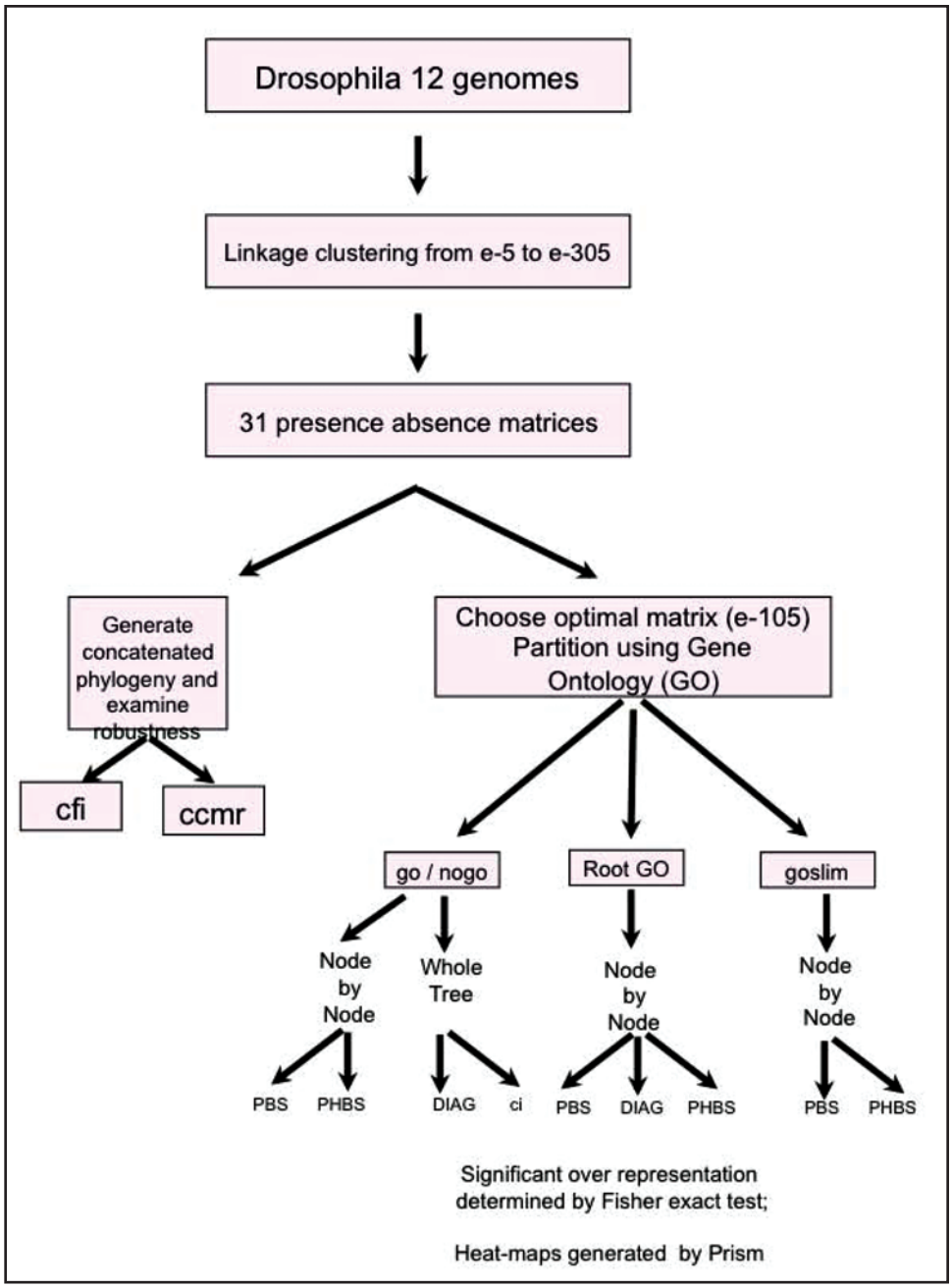

Figure 1. Flow chart showing the rationale and approaches used in the current study.

High level GO categories and tree stability. We explored the impact of gene ontology (GO) term on phylogenetic inference by dividing the gene presence/absence matrix into three partitions; those genes having identified GO terms in *D. melanogaster* ("go genes"), those genes lacking GO terms in *D. melanogaster* ("nogo1 genes"), and those genes absent in *D. melanogaster* but present in other species ("nogo2 genes"). Resultant phylogenies from each of these partitions are well supported but different from each other (Fig. 4). Surprisingly, none of these three partitions analyzed alone results in the whole matrix un-partitioned tree. This result suggests strong character interaction between partitions occurs in the concatenated matrix. The nogo2 partition generates a phylogeny where *D. ananassae* and *D. melanogaster* are each other's closest relatives, a result that is clearly at odds with all existing taxonomy. Interestingly, this partition is the only one that supports *erecta* and *yakuba* as sister taxa. Only slight differences exist between the go and nogo1 trees, all of which are due to placement of *D. melanogaster*, *D. sechelia* and *D. simulans*. This mirrors the difficulties previous phylogenetic analyses of single genes have had when trying to resolve closely related species in the *melanogaster* species subgroup.[16,21]

Further partitioning of the go genes into the three "root" GO categories (function, process and component) yields two different phylogenetic hypotheses each differing in the relationships among *D. melanogaster* (mel), *D. simlans* (sim) and *D. sechelia* (sec). Interestingly, CC (cellular component) and BP (biological processes)

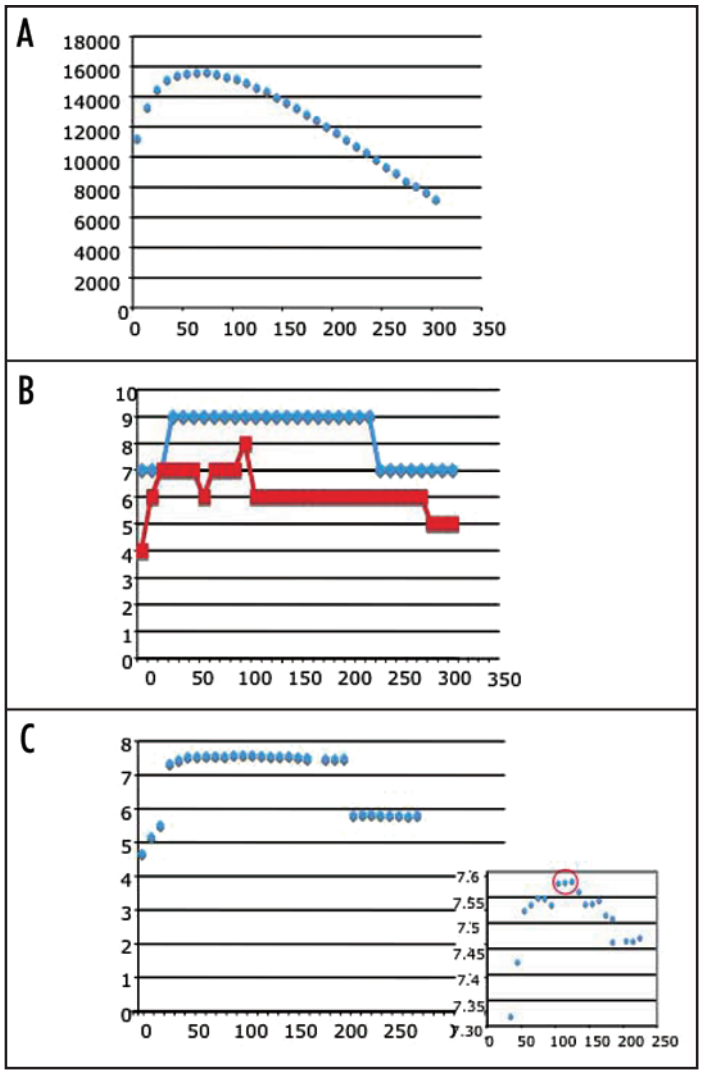

Figure 2. Analysis of E value space. Negative E value is graphed on the X axis in all three panels. (A) Graph showing the distribution of number of gene "families" (on Y axis) as a function of E value. (B) Graph showing the relationship of the consensus fork index (cfi) as a function of E value. Red line represents the cfi values computed for bootstrap trees and blue line represents cfi values computed for strict consensus trees. (C) Graph showing the relationship of the CCMR[18] as a function of E value. Inset shows a magnification of the plateau region of the larger graph between e^{-35} and e^{-225}. Red circle indicates maximal values for CCMR in this region.

partitions give the same topology ([[mel, sec], sim]), while MF (molecular function) gives a different topology ([mel, [sec, sim]]).

Correlating measures of phylogenetic consistency with function. For any partition or grouping of gene families we can estimate four important characteristics of the partition using phylogenetic analysis: (1) whether or not the presence or absence of a gene family is diagnostic for a node, (2) what the consistency of a gene family character or larger partition of gene categories is with respect to the overall phylogenetic hypothesis, (3) how much support each gene family character or groups of gene family characters contributes to each node (PBS) in the overall phylogenetic hypothesis and (4) how much support a gene family character or larger partition of gene families contributes to the overall phylogenetic hypothesis as a result of combining all partitions (PHBS). We assume that when we detect significant signal for any gene family character or larger partition of gene families for any of these four categories at a node, that the loss/gain of these gene families or partitions of genes are important in the evolution of the flies whose common ancestor is represented by the node. For more detail of these measures see the Materials and Methods section.

Consistency of characters and diagnosis of nodes in the concatenated tree. Consistency indices (a measure of how consistent the various characters are with the overall phylogenetic hypothesis) were calculated and characters were sorted into GO categories to examine hypotheses concerning function and phylogenetic events. Characters that are perfectly consistent with the combined phylogenetic topology will have a consistency index (CI) of 1.0. Non-optimal CI scores indicate incongruence at one (0.5), two (0.33), three (0.25) or four (0.20) points in the phylogeny. Figure 5A shows a heat map of the concentration of various goslim categories that are overrepresented for the four non-optimal consistency indices that are observed. Interestingly, several GO categories are statistically overrepresented in these convergent categories. For instance, genes involved in binding, metabolism, transporter activity and catalytic activity are overrepresented among gene family characters that are not entirely consistent with the phylogenetic hypothesis in Figure 3. All other GO categories appear to have fewer gene family characters that are inconsistent with the phylogenetic hypothesis (Fig. 3).

The tree in Figure 3 was used to identify gene families and higher GO categories that are diagnostic for all of the nodes in the tree. Such characters change unambiguously and can be considered unique indicators of genomic change at these nodes. In the context of gene family presence/absence, there are two kinds of unambiguous change that can occur at a given node. The first type of change is a loss of the gene family character at the node in question. Such losses are expected over evolutionary time through the action of natural selection, which can cause loci to diverge to the point of being unrecognizable as members of the same gene family. They can also be caused by drift, which can eliminate a gene from that genome via a stochastic event during a period of small effective population size. The second kind of change is the "gain" of a gene family. Gene family gain is thought to be less likely than loss in complex organisms like Drosophila because it requires rare events like horizontal transfer or neo-functionalization (the gain of new function in already existing genes). Indeed, most changes we observe that are diagnostic for nodes in the tree in Figure 3 are gene losses, in agreement with previous analyses of gene presence/absence studies on this data set.[2] Figure 5B shows that some of the higher level go/nogo (nogo1 + nogo2) partitions contribute the majority of diagnostic characters for tip nodes. For instance, Node 4 (*D. sechelia* + *D. simulans* + *D. melanogaster*) is supported almost entirely by gene families in the nogo1 partition, a set of gene families that includes loci present in *D. melanogaster* that lack GO annotation. Such gene families may have new functions since the divergence of these taxa in the common ancestor of *D. melanogaster*, *D. simulans* and *D. sechellia.* In contrast, Figure 5B shows that the nogo2 partition is the major source of diagnostic characters for the *D. persimilis—D. pseudoobscura* node. This partition includes loci that are absent in *D. melanogaster* and, as such, lack GO annotation, suggesting that these genes have originated in the *D. persimilis—D. pseudoobscura* lineage since the divergence of the *melanogaster* and *obscura* species groups roughly 15 million years ago.[11]

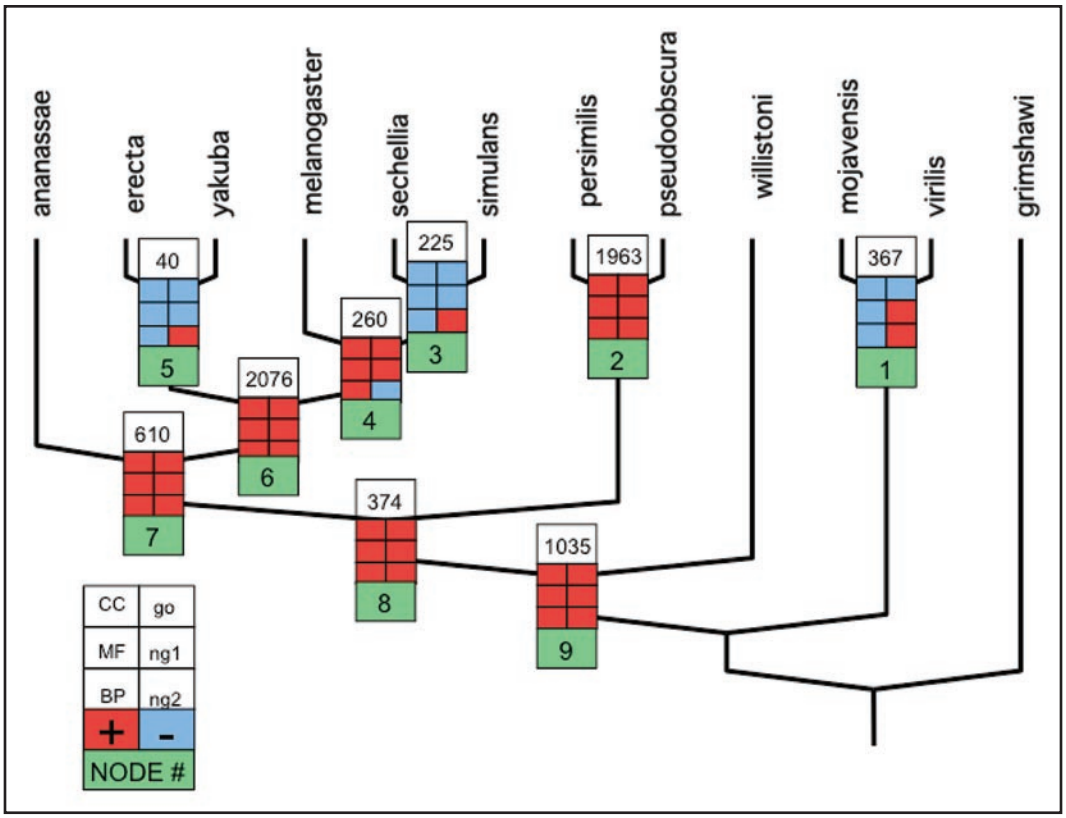

Figure 3. Phylogenetic hypothesis for the 12 genome species based on gene presence/absence for the e-^{105} matrix. All nodes are supported by 100% bootstrap and 100% jackknife proportions with 90 replacement and removal respectively and posterior Bayesian probabilities of 1.0, except for the node defining the sister pair *erecta* + *yakuba* (see text for details). The boxes on the nodes represent partitioned support measures. The number in the white box is the total Bremer or branch support measure. The six boxes below the white one tell whether the partitioned Bremer support is positive or negative for the partition indicated in the legend. The number in the green box at the bottom of each group of boxes is the node number. For example, for the node uniting *persimilis* and *pseudoobscura* (node 2 in the green box) the branch support is 1963, and all partitions are positive for partitioned Bremer supports.

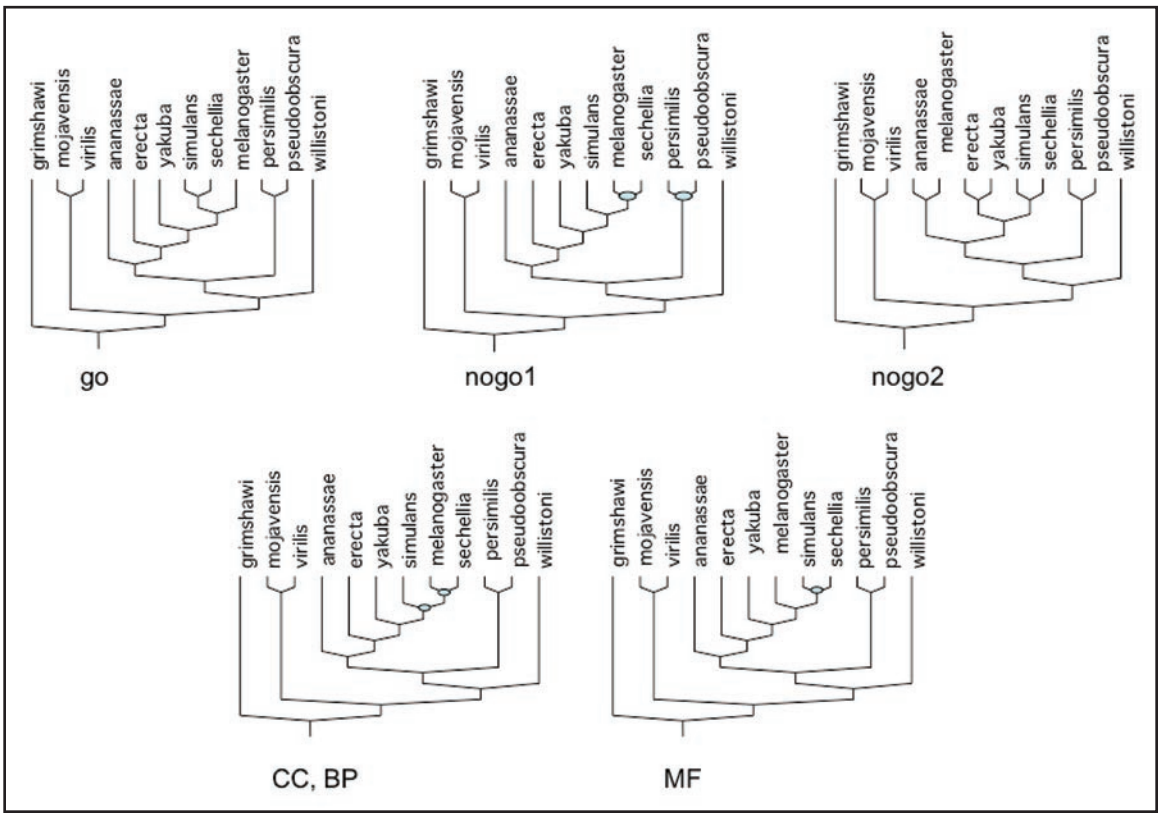

Figure 4. Trees generated by partitioning the overall data set into go categories or nogo categories as explained in the text are on the top. Trees generated by partitioning the go category genes into the three root categories Biological Process (BP), Molecular Function (MF) and Cell Component (CC). Blue dots indicate nodes that are not supported by bootstrap proportions over 80%.

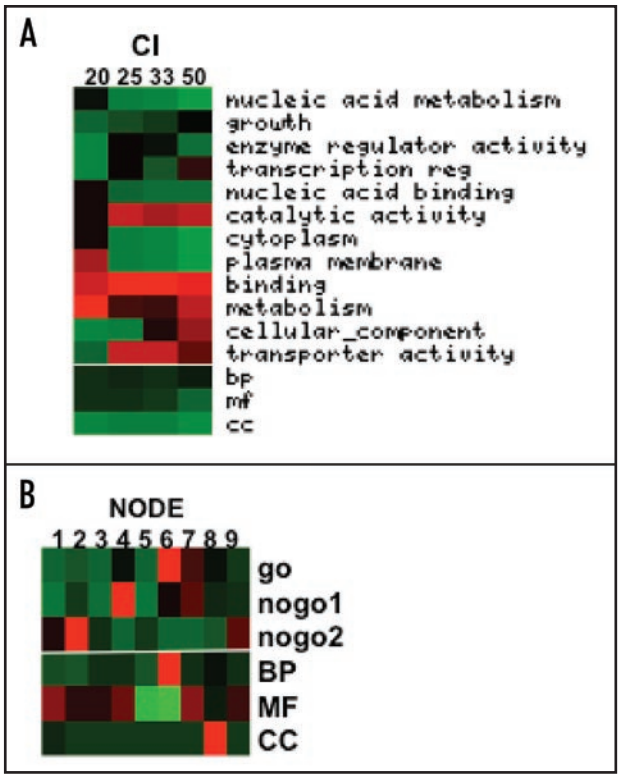

Figure 5. Heat maps showing the role of partitioning in understanding support to the Drosophila 12 genome tree. (A) Heat map constructed by grouping genes with consistency indices of 0.20 (20), 0.25 (25), 0.33 (33) and 0.5 (50) into separate partitions and then establishing GO root (below white line) and goslim (above white line) categories for these genes. Green = low number of genes, Red = high number of genes. (B) Heat map of diagnostic genes and gene families partitioned into highest category GO terms (go, nogo1 and nog2; see text for explanation) and for root categories (CC, MF and BP; see text for details) by node (1 thru 9; see Fig. 1 for description of node numbers). Green = low number of genes, Red = high number of genes.

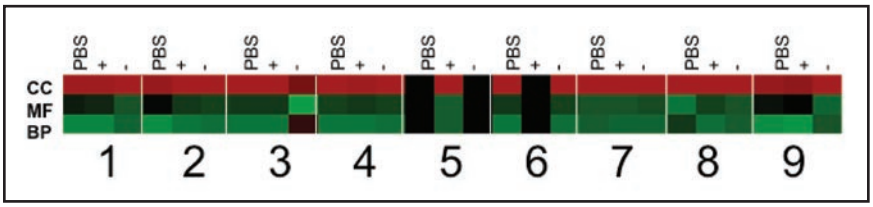

Figure 6. Heat map showing the number of statistically significant over represented genes rendering support at the nine nodes (1 thru 9) for the three GO root categories (BP, MF and CC; see text for description of root categories) in the Drosophila 12 genome tree. Three kinds of support are shown in the figure PBS, positive PHBS (+) and negative PHBS (-). Green = low number of genes, Red = high number of genes.

When root GO categories are examined for distribution of diagnostic characters, some categories are specific to major internal nodes. For example, diagnostic gene presence/absence events supporting the *melanogaster* species group (Node 6) are predominately derived from the root biological process (BP) GO category. For the node defining the sister group relationship between the *melanogaster* and *obscura* species groups (node 8), CC category genes are over-represented in the diagnostic category. Finally, we note the under-representation of genes in root category MF in diagnosis at all nodes, particularly those internal nodes supporting basal relationships within the phylogeny.

Hidden support and function. Partitioned branch support (PBS) and partitioned hidden branch support (PHBS) values for each gene family with GO annotation for each node were calculated.[22] PBS identifies the proportion or support (or conflict) provided by the presence/absence of genes and categories of genes at each node in the whole matrix tree (Fig. 3) and is positive in value or zero. PHBS determines the amount of support (or conflict) derived from data combination. We noted both positive and negative PHBS values for each of the 5,000 genes associated with GO terms. Consequently, each node can have positive or zero PBS and both positive and negative PHBS values, indicating gene content changes lending support at a given node, and PHBS values, identifying gene content changes that disagree with the relationship depicted at a specific node. We first examined root GO categories to assess what patterns might emerge upon partitioning (Fig. 6). We found that genes belonging to the BP root category were overrepresented in both positive PBS values at almost all nodes, suggesting that genes involved in biological processes are more likely to be impacted by or respond to cladogenetic events in species divergence and in divergence of major groups of Drosophila. In addition, it appears that CC category genes for the most part are least affected by cladogenesis or major divergence events.

The Division of gene families which are overrepresented for positive PBS and positive or negative PHBS values into narrower goslim categories is presented in Figure 7. Interestingly, only a few goslim categories are involved in presence/absence changes in this group (the rightmost six goslim categories in Fig. 7). Genes in two goslim categories, metabolism and binding, appear to be very active in contributing to PBS and PHBS at all nodes in the tree. These gene families fall into the biological process (BP) and molecular function (MF) root categories, both of which are expected to evolve rapidly in response to adaptation to new environments. As a consequence, presence absence data from these categories may be good markers of divergence over evolutionary time. Perhaps the most striking result of Figure 7 is the large impact of goslim categories on PBS at node 4, the ancestor of *D. sechelia, D. simulans* and *D. melanogaster*. A number of goslim categories are involved in supporting this relationship. This pattern may be the result of basing the assignment of GO terms on the *D. melanogaster* annotation and the fact that these three species are the most recently divergent in the data set. The largest contribution to support at this node is generated by reproductive genes, a class that includes some of the most rapidly evolving loci in the genome and may be involved in reproductive isolation of these closely related species.[23,24] In contrast, two goslim categories that show very little contribution to PBS values on the tree at this node are metabolism and binding.

Analysis of numbers of goslim categories supporting each node yields several interesting results. For example, tip nodes 1, 2, 3 and 5, all of which denote species pairs, are strongly supported by three goslim categories each. Internal node 6 in contrast, is supported by two goslim categories. PBS deviates from the overall pattern at node 7, which defines the *melanogaster* subgroup, and at nodes 8 and 9, the most basal in the tree. Node 7 is anomalous because the majority of support is derived from a single goslim category, the metabolism

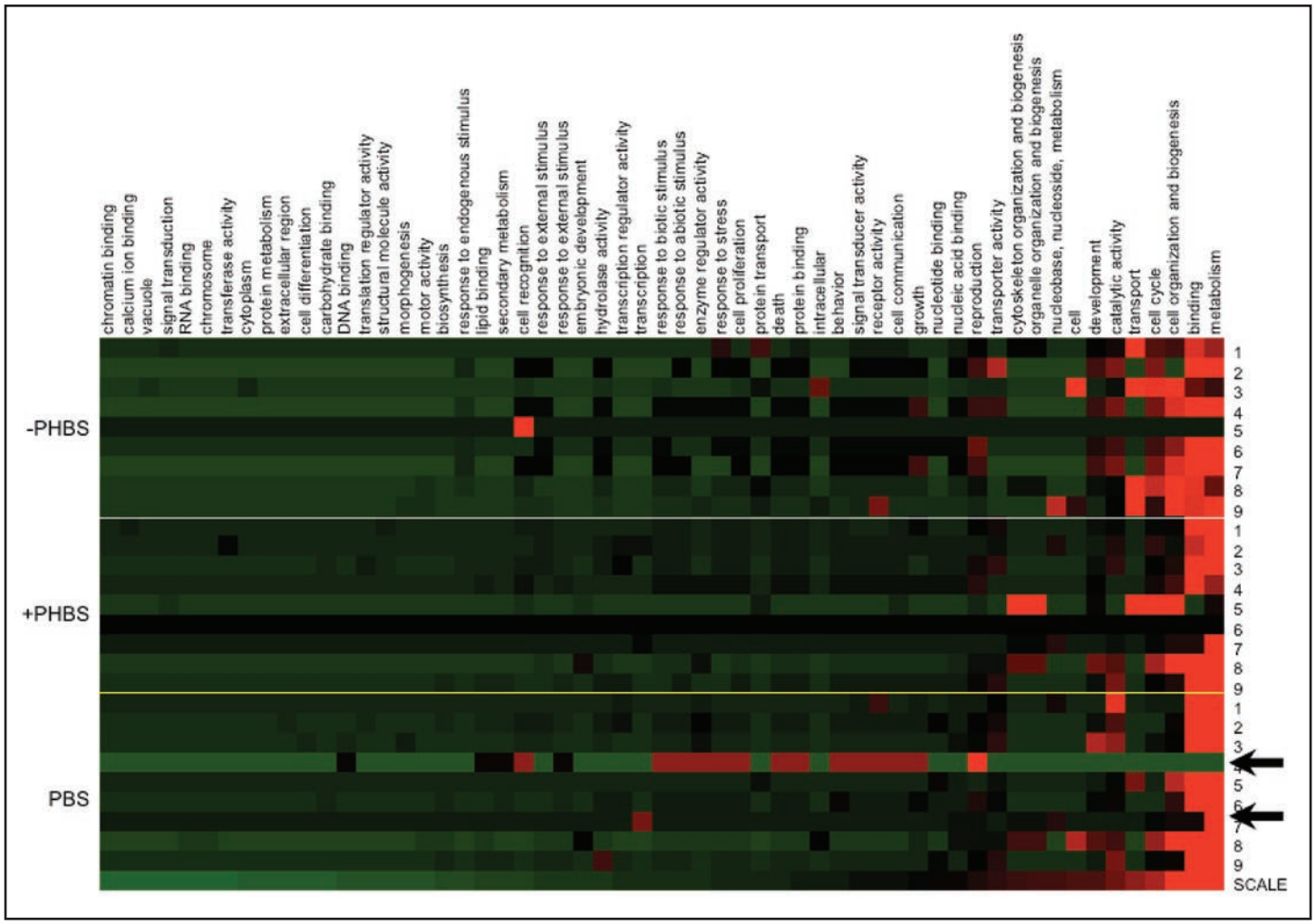

Figure 7. Heat map showing the number of statistically significantly over represented genes rendering support at the nine nodes (1 thru 9) in the Drosophila 12 genome tree for goslim categories. Three kinds of support are shown in the figure PBS (PBS), positive PHBS (+PHBS) and negative PHBS (-PHBS). Green = low number of genes, Red = high number of genes.

genes, perhaps reflecting adaptation to novel environments that the ancestor of the group occupied during its divergence. Nodes 8 and 9, which denote the *obscura—melanogaster* species group and the subgenus Sophophora respectively, are interesting because they have a broad range of goslim categories contributing to support at these nodes. This pattern may reflect the large amount of evolutionary time represented by these nodes. In other words, more goslim categories are involved because more time has elapsed since the common ancestor of these two nodes.

Gene family presence/absence matrices and phylogenomics. Constructing the tree of life is a major goal of modern biology. In order to produce a meaningful and robustly supported phylogeny, a large amount of character state information from diverse sources of data needs to be collected. The current study expands upon approaches that employ DNA or amino acid sequence data by examining characters based on gene family presence/absence and gene ontology. This approach has yielded several novel results not observed in previous analyses of DNA or amino acid sequence evolution in the 12 Drosophila genomes. For example, terminal nodes representing more recently divergent taxa (nodes 1, 3 and 5), while having reasonable branch support, bootstrap values and Bayes posterior probabilities, also possess a preponderance of negative support (*sensu*[23]). This negative branch support is often manifested in low bootstrap proportions or low Bayesian posterior probabilities. Alternately, basal and internal nodes all show strong positive branch support and correspondingly high posterior probabilities in Bayesian analyses, reflecting the deep divergence and long time to shared ancestry at these nodes.

We suggest that presence/absence data, based on genes or gene families, may provide a viable alternative to direct analyses of nucleotide and protein sequence characters for inferring the tree of life. We also demonstrate that such analyses may increase support at intermediate and basal nodes. This suggestion implies that whole genomes might become an indispensable aspect of phylogenetic reconstruction for deeper nodes in the tree of life.

Conclusion—patterns of divergence and support in presence/absence trees can reveal functional adaptation. An important aspect of using presence/absence data in phylogenetics concerns generating a character matrix containing all the relevant information. Genome-wide analyses often use E value scores to assign homology between loci or to determine which genes are included in a given analysis. The present study demonstrates that phylogenetic analysis is sensitive to

E value cutoff and that exploration of the E value space is critical to determining optimal E values based on combined congruence assessment (CCMR[18]).

Our analyses also indicate that different partitioning strategies allow for a more detailed view of support at nodes in the Drosophila 12 genome tree. We took two approaches to examine the role of describing function in supporting nodes. The first examined the gene family characters that are absolutely congruent with the phylogenetic hypothesis in Figure 3. We call these gene family characters diagnostic, and Figure 5B shows the patterns of distribution of diagnostic gene families in the three root categories by node. These diagnostic gene families are a small subset with GO annotation and their patterns are not representative of the entire data set. Nonetheless, diagnostic information appears to be dispersed in a complex pattern over the nodes in the tree. For instance, while the BP category can be designated as contributing the largest amount of support in the entire dataset, when diagnostic genes and gene families are examined this root category only contributes strongly to Node 6. The pattern for MF is even more complex with some nodes garnering strong support from genes in this root category and others getting very little support.

The second approach involved partitioning the data set into narrower goslim categories and examining the degree of support (PBS) and hidden support (PHBS) at each node from these narrower GO categories. Figure 6 demonstrates that we can, in general, rank the three major root categories in order of importance for contributing to nodes as BP > MF > CC. Some exceptions exist, but in general BP is the most important root category of genes lending support to nodes in the 12 genomes phylogeny. Figure 7 shows the results of this approach and demonstrates that in general only a few goslim categories are "active" in contributing support at nodes. These few important goslim categories are predominated by genes in the metabolism and binding categories for most of the nodes in the tree. At the broadest GO level (root categories) the ranking of these categories makes some biological sense as the cellular component category of genes might be expected to be less prone to gain/loss events as a result of strong purifying selection against changes affecting cell structure and membranes. On the other hand, genes in the BP and MF categories might be expected to tolerate more change than the CC category genes. The BP and MF categories include metabolism genes, genes involved in development and sexual reproduction and genes involved in regulating expression. All of these categories of genes have been suggested as important in Drosophila speciation and evolution. Metabolism genes have been singled out as potentially important for cladogenesis in Drosophila due to the adaptation of several of these species to unique ecological niches.

Admittedly, the taxon sampling in this data set is sparse. However, the potential for discovering genes and gene categories that have important functional roles in the evolution of the genus Drosophila is high, using this approach. We suggest that partitioning of data sets into finer and finer categories followed by a detailed analysis of the degree of support these categories lend to the overall phylogenetic hypothesis, can be a useful tool in functional genomics of these flies even with the sparse sampling that the Drosophila 12 genomes data set accomplishes. For instance, when we examine narrower GO categories, reproductive genes are shown to impact the node uniting *melanogaster*, sechelia and simulans. Another example is the strong association of support for node 7 with metabolism genes. This node defines the *melanogaster* group and the present analysis points to the importance in gain/loss of metabolism genes as a factor in the divergence of this species group, perhaps as a result of the adaptation of the ancestor of these flies to novel ecological challenges.

The inferences we make in this study are most likely strongest at the broadest GO categories, such as the three root categories. Inferences made at the lower GO categories like the goslim partitions are harder to interpret because of the sparse sampling. Denser sampling of species in this genus at the genome level, will yield stronger correlations of GO categories with specific divergence events. In other words, denser sampling will provide the researcher with more hypotheses to test with respect to GO categories.

In order to detect genes in specific GO categories that might be significant in the evolutionary process, Hahn et al., (2) defined several genes as lineage specific and then determined if any of these genes' GO terms were over represented. This approach works well with tips on the tree. We suggest that using phylogenetic approaches that determine genes that are diagnostic for internal nodes is another viable approach to identifying GO categories that have been important in the evolutionary process. Furthermore, the behavior of genes and gene families with respect to branch support (PBS and PHBS), may be yet another approach to classify genes with GO terms that are evolutionarily significant in cladogenesis.

Materials and Methods

Genomic information. All twelve full genomes with their predicted protein annotations were obtained from FlyBase (ftp://ftp.flybase.net/genomes/). All annotations were version 1.0 except for *melanogaster* which was version 5.5 and *pseudoobscura* which was 2.0.

Clustering. The 12 proteomes were compared to each other using BLAT (25). The matches were filtered based upon e-values from e^{-5} to e^{-305} at intervals of 5. All matches with an E value score less than the threshold were clustered using a single-linkage technique. If protein A matches to protein B and protein B matches to protein C, then all 3 proteins are put into the same cluster. Each species is then queried to determine if it contains at least one protein in each cluster. This query is iterated over each cluster and each species to produce a matrix indicating the presence/absence of each protein cluster (gene family) in each species. In this way, we determine if a species has at least one member of a particular gene family. We suggest that this approach is appropriately conservative since determining orthology of very closely related paralogs is very difficult and misleading orthology statements add more uncertainty to the analysis. Matrices were produced for 30 different E-values chosen arbitrarily in intervals of ten from e^{-5} to e^{-305}.

Character partitioning. We chose to examine in more detail a matrix of presence/absence information generated from e^{-105} (see results for justification of choice of this E value). The first step in the construction of the matrix was to exclude all genes that are present in a single taxon ("annotation artifacts" *sensu*[2]). These genes while interesting in many respects, have no impact on parsimony based phylogenetic analysis. The key to all of our subsequent analyses is the ability to partition the presence/absence information for gene families into smaller biologically meaningful partitions. To do this we used the Gene Ontology framework and nomenclature,[26] with Drosophila GO annotation from FlyBase version fb_2008_04.[27]

Our first level of partitioning was based on whether a gene family character had a GO number associated with it or not. Since the *melanogaster* annotation is the most accurate and extensive, we used it as a guide for assigning GO terms. In this context, we only assign GO terms to the genes that are present in the *melanogaster* genome. We call these genes "go genes". There will also be a number of genes and gene families in the *melanogaster* genome that do not currently have GO terms associated with them. We call this category of genes the "nogo1 genes". There will also be a number of genes in the matrix that are not present in *melanogaster*. For this study, we do not assign these genes GO terms and these genes we call "nogo2 genes". We examined several of the genes in the nogo1 and nogo2 categories to rule out the possibility that these are transposable elements. Our second level of partitioning is to use the root categories of the GO system—cellular component (CC), molecular function (MF) and biological process (BP). The next level was to use intermediate goslim categories as a guide for partitioning. In this scheme, Drosophila genes were organized into the over 40 different CC, MF and BP GO categories in the goslim list. Finally we created individual partitions for all 4,355 genes and gene families in the data set with GO terms associated with them.

Tree building and measures of support. We used PAUP*[28] to generate all parsimony trees and estimate tree consensus indices (using the "indices" command), bootstrap and jackknife trees, consensus trees and to process batch files for support indices generated by TreeRot. We used TreeRot[29] to determine Bremer supports by generating batch files in TreeRot, processing them in PAUP and parsing them in TreeRot. We used ASAP[30] cto generate the partitioned hidden support values for each partition for each node. MrBayes[31] was used to generate the Bayesian analysis. Given the presence/absence nature of the data in the matrix we used the "parsmodel" option in MrBayes and 20 million repetitions. The "burn in" for this data set is miniscule. We rooted all our trees with the classical break between subgenus Drosophila and subgenus Sophophora. This rooting approach allows us to polarize character change into the subgenus Sophophora and into the species pair (*mojavensis* + *virilis*) in the subgenus Drosophila. Any other statements about polarity are not possible without using a further removed outgroup.

Exploring E value space. For each of the E value partitions we generated, we estimated a phylogeny using parsimony both with a consensus tree and bootstrap constraints (character replacement set at 90%). The consensus tree will almost always be more resolved than the bootstrap tree because the bootstrap tree has more stringent requirements for the inclusion of a node in the final tree. For each of these trees we then used PAUP to estimate the tree fork indices. We used the consensus fork index (cfi) in subsequent comparisons. The cfi simply measures the number of nodes that are present in a query tree that are also present in the concatenated tree. The number of relevant nodes in the current concatenated tree (Fig. 3) is nine. Any query tree with all nodes identical to the concatenated tree has a cfi = 9; any tree that lacks one node of the concatenated tree regardless of where the node is in the tree has a cfi = 8 and so on. We then simply graph the E value versus the cfi for both consensus and bootstrap. We also used the Combined Corroboration Metric based on the retention index (CCMR[18]) that is simply the product of the retention index and the cfi for each tree. Supplemental Table 1 contains the 30 partitioned matrices generated using E values from -5 to -305.

Using the GO classification system to partition the overall dataset. In order to explore the behavior of gene presence/absence transitions with respect to the phylogenetic hypothesis, we used several approaches. Our first approach is to examine the characters that are purely diagnostic for the nine nodes in the tree (see also 2). These characters change only once at a single node and are either lost ($1 \rightarrow 0$) or gained ($0 \rightarrow 1$) at the node in question and nowhere else in the tree. Characters that have $1 \rightarrow 0$ transitions are most likely genes or gene families where a particular function is lost and hence the gene or gene family is either eliminated from the genome or diverges extremely from its previous orthologs and cannot be detected as an ortholog anymore. Characters that have $0 \rightarrow 1$ transitions are most likely genes that are the product of gene duplications. Alternatively an already existing gene might have diverged and acquired new function (neofunctionalized) at the particular node in the tree.

Our second approach deals with the characters in the matrix that are not purely diagnostic and is based on the simple idea that characters that are more consistent in the entire tree differ in significant ways from characters that are not consistent. We estimated consistency indices for each character in the matrix using PAUP and also estimated the partitioned Bremer support (PBS[32]) and partitioned hidden Bremer support (PHBS[22]) for each character and character partition in the matrix using the approaches described above.

To further explore the data set we used the various GO partitions described above to attempt to correlate function with phylogenetic pattern. In this approach we estimated support values for each of the GO category partitions (both PBS and PHBS) and used these estimates to further partition the GO categories. PBS and PHBS values for each character are calculated on a node by node basis. The values for PBS and PHBS can be positive, in which case the character contributes positive hidden support, negative in which case the character contributes negative hidden support or zero in which case all support from the character is found in the PBS. These support values can be used as indicators of how stringently a particular gene or class of genes agrees with the phylogeny. Our reasoning is that if there is strong positive or negative support or hidden support from a character partition at a node, that partition has undergone significant evolutionary change at the node.

Once diagnostic state, character consistency, PBS and PHBS values for each character are all estimated the gene columns are then partitioned according to their GO term. For instance, characters 532, 3696, 3735, 5650, 5754, 6404, 6610, 7075, 1307 and 10297 are all genes that are associated with the GO term GO:0006406. In this way, we can associate the measures of character consistency (consistency index or ci), diagnosis, PBS and PHBS mentioned above with the gene ontology system. Through this process we obtain a list of GO categories for each node that are contributing to four important categories in phylogenetic analysis: (1) whether or not the gene family character is diagnostic for a node, (2) what the consistency of a gene family character or larger partition of gene families is with respect to the overall phylogenetic hypothesis, (3) how much support each gene or groups of gene family characters contribute to each node (PBS) in the overall phylogenetic hypothesis and (4) how much support a gene family or larger partition of genes contributes to the overall phylogenetic hypothesis as a result of concatenation (PHBS). We assume that when we detect significant signal for any gene family or larger partition of genes for any of these four categories at a node,

that these genes or partitions of genes are important in the evolution of the flies whose common ancestor is represented by the node. Figure 1 shows a flow diagram of the analyses that were accomplished for this study. Supplemental Table 2 contains the entire data matrix with partitions into GO categories for e^{-105}.

Determining significance of over-representation of GO terms in diagnostics, ci, PBS and PHBS. Once GO category lists are obtained for each of the four categories above we determined if significant over-representation of GO terms occurs in the lists using the FatiGO[33] option in the Babelomics software suite (http://www.babelomics.org/).[34] This option gives a list of over represented terms relative to the entire genome of *melanogaster* by using a Fisher exact test using $p < 0.05$ as a cutoff for significance. The potential multiple comparisons problem of Fisher's exact test for over representation of GO terms is eliminated using FatiGO in the Babelomics software suite.[34] Once the GO lists are trimmed of non significant terms, we used the GO Category Classifier[35] to group terms on the basis of their root categories and their generic goslim categorization (www.spatial.maine.edu/~mdolan/MGI_GO_Slim.html). For each node we compiled lists of over represented terms for the PBS (in our study always positive), negative PHBS and positive PHBS. We then used the PRISM server http://noble.gs.washington.edu/prism/)[36] to construct heat maps to display the frequency of over represented GO terms at all nine nodes in the concatenated tree hypothesis.

Acknowledgements

The authors thank the Sackler Institute for Comparative Genomics at the American Museum of Natural History. Funding for this work was kindly provided by the Lewis B. and Dorothy Cullman Program in Molecular Systematics at the AMNH and the Korein Family Foundation at the AMNH. Ernest K. Lee and Jeffrey R. Rosenfeld are supported by NSF Grant DBI 0421604.

Note

Supplementary materials can be found at:
www.landesbioscience.com/supplement/RosenfeldFLY2-6-Sup.pdf
www.landesbioscience.com/supplement/Rosenfeld-data01.tds
www.landesbioscience.com/supplement/Rosenfeld-data02.tds

References

1. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007; 450:203-18.
2. Hahn MW, Han MV, Han SG. Gene family evolution across 12 Drosophila genomes. PLoS Genet 2007; 11:197.
3. Kopp A, Barmina O, Hamilton AM, Higgins L, McIntyre L and Jones C. Evolution of gene expression in the Drosophila olfactory system. Molecular Biology and Evolution 2008; 25:1081-92.
4. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences 2007; 104:4996-5001.
5. Clark AG. Genomics of the evolutionary process. TREE 2006; 21:316-21.
6. de Freitas Ortiz M, Loreto EL. Characterization of new hAT transposable elements in 12 Drosophila genomes. Genetica 2008; [Epub ahead of print].
7. Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. Evolution of protein-coding genes in Drosophila. Trends Genetics 2008; 24:114-23.
8. Estes P, Fulkerson E, Zhang Y. Identification of motifs that are conserved in 12 Drosophila species and regulate midline glia vs. neuron expression. Genetics 2008; 178:787-99.
9. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evol Biol 2007; 7:226.
10. Powell J and DeSalle R. Drosophila molecular phylogenies and their uses. In Evolutionary Biology (edited by M. Hecht, et al.,) 1995; 28:87-138.
11. Russo CA, Takezaki M, Nei M. Molecular phylogeny and divergence times of drosophilid species. Molecular Biology and Evolution 1995; 12:391-404.
12. Tatarenkov A, Ayala FJ. Phylogenetic relationships among species groups of the *virilis/repleta* radiation of Drosophila. Molecular Phylogenetics and Evolution 2001; 21:327-31.
13. Kwiatowski J, Ayala FJ. Phylogeny of Drosophila and related genera: conflict between molecular and anatomical analyses. Molecular Phylogenetics and Evolution 1999; 13:319-28.
14. Remsen J, DeSalle R. Character congruence of multiple data partitions and the origin of the Hawaiian Drosophilidae. Mol Phylogenet Evol 1998; 9:225-35.
15. Remsen J, O'Grady P. Phylogeny of Drosophilidae (Diptera), with comments on combined analysis and character support. Mol Phylogenet Evol 2002; 24:248-63.
16. Pollard D, Iyer VN, Moses AM, Eisen MB. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genetics 2006; 2:173.
17. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003; 425:798-804.
18. Lienau EK, DeSalle R, Rosenfeld JA, Planet PJ. Reciprocal illumination in the gene content tree of life. Syst Biol 2006; 55:441-53.
19. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics 2006; 22:699-707.
20. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J; Harvard FlyBase curators; Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 2007; 450:219-32.
21. Markow TA, O'Grady PM. Drosophila Biology in the Genomic Age. Genetics 2006; 177:1269-76.
22. Gatesy J, O'Grady P, Baker RH. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher-level artiodactyl taxa. Cladistics 1999; 15:271-313.
23. Holloway AK, Begun DJ. Molecular evolution and population genetics of duplicated accessory gland protein genes in Drosophila. Mol Biol Evol 2004; 21:1625-8.
24. Reed LK, Markow TA. Early events in speciation: polymorphism for hybrid male sterility in *Drosophila mojavensis*. Proceedings of the National Academy of Sciences 2004; 101:9009-12.
25. Kent WJ. BLAT-The BLAST-Like Alignment Tool. Genome Research 2002; 12:656-64.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25-9.
27. Wilson RJ, Goodman JL, Strelets VB. The FlyBase Consortium. FlyBase:integration and improvements to query tools. Nucleic Acids Research 2008; 36:588-93.
28. Swofford D. Phylogenetic Analysis Using Parsimony* (*and other methods). 4.0b10 ed. Sunderland: Sinauer 2005.
29. Sorenson MD and Franzosa EA. TreeRot, version 3. Boston University, Boston, MA 2007.
30. Sarkar IN, Egan MG, Coruzzi G, Lee EK, DeSalle R. Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phylogenomics. BMC Bioinf 2008; 9:103.
31. Huelsenbeck J, Ronquist F. MrBayes Version 3.0. *Uppsala, Sweden, Evolutionary Biology Centre, Uppsala University* 2003.
32. Baker R, DeSalle R. Multiple sources of molecular characters and the phylogeny of Hawaiian drosophilids. Syst Biol 1997; 46:654-73.
33. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004; 20:578-80.
34. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JMM, Conde L, Blaschke C, Vera J, Dopazo J. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Research 2006; 34:472-6.
35. Hu Z-L, Bao J, Reecy LM. A Gene Ontology (GO) Terms Classifications Counter. Plant & Animal Genome XV Conference, San Diego, CA 2007.
36. Wu W, Noble WS. Genomic data visualization on the Web. Bioinformatics 2004; 20:1804-5.