Parallel evolution of chimeric fusion genes

Corbin D. Jones*† and David J. Begun‡

*Department of Biology and Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599; ‡Center for Population Biology, University of California, Davis, CA 95616

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved June 27, 2005 (received for review April 29, 2005)

To understand how novel functions arise, we must identify common patterns and mechanisms shaping the evolution of new genes. Here, we take advantage of data from three Drosophila genes, jingwei, Adh-Finnegan, and Adh-Twain, to find evolutionary patterns and mechanisms governing the evolution of new genes. All three of these genes are independently derived from Adh, which enabled us to use the extensive literature on Adh in Drosophila to guide our analyses. We discovered a fundamental similarity in the temporal, spatial, and types of amino acid changes that occurred. All three genes underwent rapid adaptive amino acid evolution shortly after they were formed, followed by later quiescence and functional constraint. These genes also show striking parallels in which amino acids change in the Adh region. We showed that these early changes tend to occur at amino acid residues that seldom, if ever, evolve in Drosophila Adh. Changes at these slowly evolving sites are usually associated with loss of function or hypomorphic mutations in Drosophila melanogaster. Our data indicate that shifting away from ancestral functions may be a critical step early in the evolution of chimeric fusion genes. We suggest that the patterns we observed are both general and predictive.

*Department of Biology and Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, CA 27599; ‡Center for Population Biology, University of California, Davis, CA 95616

To understand how novel functions arise, we must identify common patterns and mechanisms shaping the evolution of new genes. Here, we take advantage of data from three Drosophila genes, jingwei, Adh-Finnegan, and Adh-Twain, to find evolutionary patterns and mechanisms governing the evolution of new genes. All three of these genes are independently derived from Adh, which enabled us to use the extensive literature on Adh in Drosophila to guide our analyses. We discovered a fundamental similarity in the temporal, spatial, and types of amino acid changes that occurred. All three genes underwent rapid adaptive amino acid evolution shortly after they were formed, followed by later quiescence and functional constraint. These genes also show striking parallels in which amino acids change in the Adh region. We showed that these early changes tend to occur at amino acid residues that seldom, if ever, evolve in Drosophila Adh. Changes at these slowly evolving sites are usually associated with loss of function or hypomorphic mutations in Drosophila melanogaster. Our data indicate that shifting away from ancestral functions may be a critical step early in the evolution of chimeric fusion genes. We suggest that the patterns we observed are both general and predictive.
mimica, and Drosophila nigra. For Adh-Twain, we used Drosophila ambigua, D. guanche, D. madeirensis, D. melanogaster, Drosophila pseudoobscura, and D. subobscura.

There are three major steps to using PAML: (i) choice of appropriate model, (ii) parameterization of that model, and (iii) stepwise comparison simpler to more complex hypotheses by using log-likelihood ratio tests using these models to see which is superior to fit the data. We used the codon model (codeml; refs. 20 and 25) for all analyses because we limited ourselves to the ORFs of the genes of interest and had DNA sequence data. For the analyses discussed in Results, our a priori hypothesis is that the lineage created by the formation of the CFG will experience more rapid evolution than the Adh lineages. We then analyzed the evolution of the Adh-derived sequences of each CFG independently. Twice the difference in the log-likelihood (2Δl), which is $\approx \chi^2$ distributed for the relevant degrees of freedom, was used to estimate a P value. Most significant results had P < 0.001. In general, the estimated codon table fit the data best, except in the repleta group analysis, where the estimated codon table was only marginally better than the F3 × 4 model ($P = 0.0331$). For consistency across analyses, we use the estimated approach, anyway. When appropriate for the analysis, $\kappa$ and $\omega$ were estimated (see ref. 19; as necessary, convergence to maximum-likelihood estimates was verified by changing the small-difference parameter (see ref. 20, p. 19)). Reconstruction of ancestral sequences was done by using both joint and marginal reconstruction (26). These sequences were typically identical regardless of method. Testing for selection early in the evolution of each fusion gene is described in Results.

Stepwise statistical methods such as PAML typically involve making many tests of a hypothesis. This approach can lead to a multiple testing problem when determining the significance of P values. For the analyses detailed in Results, most P values are robust to a Bonferroni correction. Any exceptions are noted.

We obtained 61 complete Adh coding sequences from GenBank (accession numbers available upon request) for the following species: Drosophila adiastola (Adh1 and Adh2), Drosophila affinitidisjuncta, D. ambigua, D. americana, Drosophila arizonae, D. borealis, D. buzzatii (Adh1 and Adh2), Drosophila crassifemur, Drosophila differens, Drosophila equinoxialis, Drosophila erecta, Drosophila flavomontana, Drosophila funebris (Adh1 and Adh2), D. guanche, Drosophila hawaiiensis, Drosophila heteroneura, D. hydei (Adh1 and Adh2), Drosophila immigrans, Drosophila insularis, Drosophila kunteii, Drosophila lacciola, Drosophila lummei, D. madeirensis, Drosophila mauritiana, Drosophila mayaguan, D. melanogaster, D. metteri, Drosophila mimica, D. miranda, Drosophila mojavensis (Adh1 and Adh2), Drosophila montana, D. mulleri (Adh1 and Adh2), D. navajoa, Drosophila nebulosa, D. nigra, D. orena, Drosophila paulistorum, Drosophila picticornis, Drosophila planitibia, D. pseudoobscura, Drosophila sechellia, Drosophila silvestris, Drosophila simulans, D. subobscura, D. teissieri, Drosophila texana, Drosophila tropicalis, Drosophila virilis, Drosophila wheeleri, Drosophila willistoni, D. yakuba, Idiomyia grimmshawi, and Scaptomyza pallida. These sequences span >60 million years of Adh evolution. We used CLUSTAL software (27) to create a multiple alignment of these sequences. We identified invariant sites by simply scanning these sequences for sites at which the amino acid did not vary in any species. We then contrasted these sites to the 52 D. melanogaster Adh mutations for which the molecular basis of the mutation is known that we identified in FlyBase (www.flybase.org).

There are two additional amino acids coded for in Adh in some of the taxa listed above. We ignored these sites in our analysis because they cannot be aligned in all taxa. Amino acid residue positions described below are in terms of the smaller Adh.

Within species, DNA polymorphism at Adh and CFGs can influence both the PAML analysis and the alcohol dehydrogenase (ADH) protein analysis. In the PAML analysis, polymorphism may slightly inflate the rate of evolution (both synonymous and nonsynonymous) for all tip branches (Adh and CFG). The reconstructed branches, which are derived from multiple sequences, should be less susceptible to this bias. In the Adh analysis, polymorphism will reduce the number of sites that we define as invariant.

Results

Early Adaptive Evolution of CFGs. By comparing the ratio of nonsynonymous to synonymous substitutions (Dn/Ds), we can determine whether rapid amino acid evolution along a branch of a phylogeny is consistent with adaptive evolution. Ratios less than 1 imply functional constraint on the amino acid sequence of a protein. Ratios around 1 are consistent with low functional constraint or neutral evolution of a protein. Ratios greater than 1 suggest directional selection on the protein. Jones et al. (19) reported that the Adh region of Adh-Twain in D. subobscura, D. madeirensis, and D. guanche shows strong evidence for adaptive evolution shortly after the formation of this CFG (Dn/Ds $\gg$ 1). After the speciation of D. subobscura and D. guanche, the Dn/Ds ratio drops well below 1 (0.399), consistent with increasing functional constraint, although this ratio is still nearly an order of magnitude greater than the ratio observed for Adh in these lineages (0.0411).

Long and Langley (14) and Begun (16) both suggested that there was an elevated rate of amino acid substitution after the formation of jingwei and Adh-Finnegan, respectively. Codon-based maximum-likelihood models of sequence evolution were not available at the time of these earlier papers. Jones et al. (19), however, did use these models for their analysis of Adh-Twain in D. subobscura and its relatives. For consistency, we reanalyzed the data from jingwei and Adh-Finnegan by using these modern tools.

To determine whether the pattern observed for Adh-Twain was mirrored by jingwei and Adh-Finnegan, we used maximum likelihood to compare models of sequence evolution. Adh-Twain analysis by Jones et al. (19) suggested at least three distinct Dn/Ds ratios along the branches of our gene trees of the CFGs and their Adh paralogs. Most branches of the tree, including the CFG ancestor, should have one ratio. The branch immediately after the formation of the CFG should have a different ratio. The extant CFG branches should have a third ratio (three-ratio model).

Using Adh and jingwei sequences from D. melanogaster, D. orena, D. tessieri, Drosophila tsacasi, and D. yakuba, we showed that a three Dn/Ds ratio model was best. We compared this three-ratio model to a one-ratio model (all branches are constrained to a single Dn/Ds ratio), to a free-ratio model (all branches have independent Dn/Ds ratios), and to a two-ratio model (jingwei branches have one Dn/Ds ratio, all other branches have a different ratio). The three Dn/Ds ratio model fit the data better than the one-ratio model (three-ratio model ln l = −1800.72, in which ln l is likelihood; one-ratio model ln l = −1815.24; three vs. one ratio, 2Δln l = 29.04, df = 2, P < 0.0001), and better than the two-ratio model (two-ratio model ln l = −1804.43; three vs. two ratio, 2Δln l = 7.42, df = 1, P = 0.0065). The more parameter-rich free-ratio model did not fit the data significantly better than the three-ratio model (free-ratio model ln l = −1792.31; three vs. free ratio, 2Δln l = 16.82, df = 9, P = 0.052). We also explored a few other variations of these models, none of which fit the data significantly better than the three-ratio model (data not shown).

Although the Dn/Ds ratio is $>1$ for the branch immediately after the formation of jingwei, the signal of adaptive amino acid substitutions is not as strong as that observed for Adh-Twain (early jingwei branch, Dn/Ds = 1.27; later jingwei branches, Dn/Ds = 0.123; and Adh branches, Dn/Ds = 0.036). Neverthe-
less, the Dn/Ds ratios of the two sets of jingwei branches are 35 times and 3.4 times greater than that of the Adh lineages, respectively. This pattern also is consistent with that observed for Adh-Twain. It is worth noting that if the distant outgroup D. tatai is removed from the analysis, the three-ratio model is still highly supported (three- vs. two-ratio, 2Δln l = 11.4, df = 1, P = 0.0008). In this case, the Dn/Ds ratio is substantially >1 (we estimated 12 nonsynonymous substitutions to 0 synonymous substitutions).

Adh-Finnegan is older than jingwei and Adh-Twain and is present in a large number of repleta group species (15, 16). Comparison of Adh-Finnegan to Adh in this group is complicated by the existence of several Adh duplications in these lineages. Despite these differences from the earlier analyses, the overall pattern remains the same. Again, the three Dn/Ds ratio model fits the data better than the one-ratio model (three-ratio model ln l = −4886.19; one-ratio model ln l = −4904.90; three vs. one ratio, 2Δln l = 37.42, df = 2, P < 0.0001) and better than the two-ratio model (two-ratio model ln l = −4902.08; three vs. two ratio, 2Δln l = 31.78, df = 1, P < 0.0001). The more parameter-rich free-ratio model did not fit the data significantly better than the three-ratio model (free-ratio model ln l = −4,865.22; three vs. free ratio, 2Δln l = 22, df = 2, P = 0.059).

Dn/Ds is 2.67 for the branch immediately after the formation of Adh-Finnegan, which again suggests early adaptive evolution in the Adh region of Adh-Finnegan. As observed in Adh-Twain and jingwei, the amino acid substitution rate at Adh-Finnegan dramatically slows after this initial round of adaptive evolution (later Adh-Finnegan branches, Dn/Ds = 0.096; Adh branches, Dn/Ds = 0.080).

These three analyses suggest that the proteins of all three CFGs evolved adaptively shortly after they were formed (Fig. 1). Subsequent evolution was more constrained although typically more rapid than what is normally observed at Drosophila Adh.

Convergence of Early Amino Acid Substitutions. Four of the seven shared sites of early amino acid changes in Adh-Twain and Adh-Finnegan are transitions to the same amino acid. Likewise, the jingwei to Adh-Twain comparison suggests that four of the five shared amino acid substitutions were to the same amino acid. This pattern suggests a surprising degree of biochemical convergence. However, only one of the seven changes shared between jingwei and Adh-Finnegan was to the same amino acid.

One explanation for the high degree of biochemical convergence is that it is an artifact arising from some bias in our reconstruction of ancestral states. This explanation is unlikely. All reconstructed shared changes had a 95% or greater probability, except for one in Adh-Finnegan, which was only 76%. This site, however, is not one at which convergence of amino acid was observed. Second, for the sites with the same amino acid changes, we looked at the parsimony reconstruction of the site (29). In all cases, it was generally congruent with our marginal and joint reconstructions. Third, we assessed how often these sites vary across all extant sequences in our data set. Only one of the shared changes ever varies across the Adh sequences (site 245). Three of the sites vary within the CFG lineages, two in Adh-Finnegan (sites 68 and 219), and one in Adh-Twain (site 127), but each of these changes is limited to a single species. These data suggest that the convergence observed at these sites is not a byproduct of the methods used to reconstruct these ancestral sequences.

Formally, it is possible that the observed biochemical convergence results from a strong substitution bias toward a particular base or bases. However, we observed no clear patterns of such a bias. The most common substitution was unique to each gene: in Adh-Twain, it was G to T; in jingwei, it was C to G and G to A; and in Adh-Finnegan, it was C to A. In both jingwei and Adh-Finnegan, more As substituted than any other base, although the proportions were different (43% and 58%, respectively). Ts, at 37%, were most frequent base substituted in Adh-Twain. As a further test, we compared the transition/transversion rate ratios (κ) among all three CFG data sets. If a shared substitution bias exists, one may expect that κ would be similar across these genes. We tested this idea by fixing the κ of one CFG–Adh data set with the estimated κ from the other two data sets and seeing whether this provided a better fit to the data than a κ estimated from that particular CFG–Adh data set. This test is conservative, because the fixed κ has fewer degrees of freedom than does the estimated κ. The estimated κ generally fit significantly better than the κ from the other
Fig. 1. The protein sequence of new Adh-derived CFGs evolves rapidly shortly after these genes are formed. A comparison of Dn (Left) and Ds (Right) for Adh-Twain, jingwei, and Adh-Finnegan as estimated by our three-ratio model (see Results). Red highlights the Dn branches immediately after the new genes are formed. Red arrows indicate where the equivalent Ds branches are or would be if they existed. Green indicates the subsequent CFG branches. Blue indicates Adh lineages.
Table 1. Frequency of shared amino acid changes

<table>
<thead>
<tr>
<th>Gene comparison</th>
<th>% of sites shared*</th>
<th>P value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>jingwei to Adh-Twain</td>
<td>45 (5 of 11)</td>
<td>0.005</td>
</tr>
<tr>
<td>jingwei to Adh-Finnegan</td>
<td>63 (7 of 11)</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>Adh-Finnegan to jingwei</td>
<td>30 (7 of 23)</td>
<td>0.004</td>
</tr>
<tr>
<td>Adh-Finnegan to Adh-Twain</td>
<td>30 (7 of 23)</td>
<td>0.004</td>
</tr>
<tr>
<td>Adh-Twain to Adh-Finnegan</td>
<td>50 (7 of 14)</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>Adh-Twain to jingwei</td>
<td>36 (5 of 14)</td>
<td>0.021</td>
</tr>
</tbody>
</table>

*Percent and number of the early amino acid changes in the gene listed first found in the early amino acid changes found in the second gene.
†A binomial test was used to determine the probability that this overlap is due to chance (see text).

CFG–Adh data sets, with one exception (analysis not shown). jingwei using the Adh-Finnegan κ was borderline (P = 0.074, although the Adh-Finnegan with jingwei κ was a significantly worse fit to the data (P = 0.0002). Interestingly, despite the exchangability of their transition/transversion rate ratios, jingwei and Adh-Finnegan had the least similar amino acid changes. These analyses make it seem unlikely that the shared amino acid changes observed evolved by chance. This observation leaves natural selection as the more likely explanation for the rampant parallelism observed in the data.

Discussion

Our analysis of three Adh-derived CFGs in Drosophila suggests a fundamental similarity in the temporal, spatial, and types of amino acid changes in these proteins. These proteins experienced a burst of adaptive substitution shortly after they were formed, followed by a slowing of evolution, which is consistent with increased evolutionary constraint or a slowing of adaptive evolution. The early amino acid changes often occurred at the same residues across these three proteins and at sites that do not normally evolve in Adh. Finally, our analysis suggests a surprising degree of convergence across all three proteins as to which amino acids arise at these shared, evolving residues.

The early burst of protein evolution in each of the three novel genes is the easiest of these three patterns to explain. CFGs that survive are likely to be under strong directional selection, because the initial chimeric protein is probably suboptimal. The level of parallelism in the residues that evolve in these novel proteins, however, is more difficult to explain.

The parallel changes often occurred at conserved residues that are required for ADH function. Moreover, the same amino acid residue often changed in independent lineages. One possible explanation for these patterns is that selection favored diminution of canonical ADH activity early in the history of these novel proteins. In both Adh-Twain and jingwei (and presumably Adh-Finnegan), the 5′ regulatory apparatus derives from the non-Adh parent. As a result, the expression patterns of these genes mirror that of the non-Adh parent, not Adh. Adh overexpression experiments in Drosophila have shown that delayed development can result from misexpression of this gene (e.g., ref. 30). Residual expression of Adh-like activity, shortly after the formation of a CFG, was potentially deleterious. To explain the parallelism observed both at the level of residue and in the state of the derived amino acid, we suggest that there was selection for reduced Adh function rather than loss of function. The large number of conserved ADH residues suggests that many amino acid sites are potentially mutable to loss-of-function alleles. Moreover, most amino acid changes at these particular sites probably result in proteins that are amorphs, or nearly so. These amorph mutations are unlikely to contribute to a successful new gene. There are, however, a few residues that are mutable to partial loss-of-function alleles, and among those, a subset of mutations that actually result in partial loss of function rather than complete or nearly complete loss of function. Although rare, these later mutations, which can reduce Adh activity while retaining a viable protein, are much more likely to contribute to a functional new gene. This hypothesis explains the early adaptive evolution of the Adh region, the similarity of the amino acid changes, and the fact that these substitutions occurred at conserved sites. Moreover, this hypothesis also may partially explain why these CFG proteins were not observed on Adh allozyme gels, although it is difficult to exclude the effects of reduced gene expression in the CFGs (14, 31–33).

Our hypothesis, however, does not require that these CFG proteins lose the ability to catalyze the oxidation of alcohols, nor does it require that these three genes are evolving to the same new function. It requires only that the CFG catalytic activity shifts enough so that the CFG’s activity no longer substantially overlaps and interferes with that of Adh (a shift in activity). Even if a CFG was beneficial at its inception, the misexpression of Adh presumably lessened the adaptive advantage conferred by this new gene. This pleiotropic drag is reduced easily by changing those sites that are critical for canonical Adh activity.

An alternative explanation for the observed pattern of convergence is that the amino acid substitutions in the CFGs restore structural aspects of the Adh protein after fusion with new 5′ regions. This hypothesis seems less parsimonious for two reasons. First, this model does not explain the allozyme data. Second, the genes fused to the 5′ prime ends of the Adh differ in size, amino acid composition, and ancestral function. It is unlikely that such different 5′ structures would engender common structural changes in Adh.

If our shift hypothesis is true, our model suggests that genes participating in gene-fusion events are examples of neofunctionalization rather than redundancy or subfunctionalization. As neormorphs, these genes are likely examples of genes that are associated with adaptive phenotypes. Recently, Zhang et al. (34) presented biochemical evidence that the jingwei protein has indeed gained new function relative to Adh. They showed that jingwei has expanded its substrate range to include long-chain primary alcohols. Relative to Adh, jingwei shows increased specificity for long-chain alcohols such as farnesol and geraniol and decreased specificity for 1-propanol. This change in substrate preference in jingwei is consistent with the shift model, although it is important to note that jingwei and Adh both catalyze oxidation/reduction of a broad range of substrates in vitro.

The ultimate test of the shift model is whether additional copies of novel Adh-derived proteins show similar patterns of evolution and whether other classes of novel chimeric proteins with a shared ancestral function also show rampant parallelism. This, of course, requires that more CFGs be identified. We recently detected another potential Adh-derived CFG in the genomic sequence of Drosophila ananassae on the equivalent of chromosome 2R. Like jingwei and Adh-Twain, this gene appears to be derived from an Adh retrosequence with an ORF extending 5′ of the normal Adh start codon. Comparison of this gene with the D. ananassae Adh shows that the Adh retrogene has diverged similarly to the three previously described CFGs. Of the changes that occurred in the D. ananassae retrogene, 22% occurred at the conserved sites described above. Nine of the 11 early amino acid changes observed in jingwei also occur in the D. ananassae gene. Six of the 14 early amino acid changes in Adh-Twain also are found in the D. ananassae gene. Fifteen of the 24 early substitutions in Adh-Finnegan also are found in the D. ananassae novel gene. Although additional evidence is needed to prove that this potential fusion gene is an actual gene, the fact that the D. ananassae gene exhibits the exact same patterns of substitution as observed in the other three Adh-derived CFGs supports our model. Overall, our data suggest that the ancestral function of a
protein can severely limit the diversity of evolutionary trajectories it can follow as it evolve toward some novel function. Clearly, however, functional data from all four novel Adh-derived genes will be needed to further test this idea. Several additional hypotheses can be tested as new CFGs are identified and as we develop a deeper understanding of the CFGs already identified. First, Devor and Moffat-Wilson (ref. 35; see also ref. 36) have shown that short, conserved, widely expressed genes tend to contribute more retrogenes in humans. These short, widely expressed genes also may contribute more new CFGs as well. Adh, for instance, is a short, conserved, widely expressed gene. Second, Conant and Wagner (37) have recently shown that a subset of gene duplications show strong asymmetric amino acid divergence. One might expect this pattern to be true of duplication-derived fusion genes as well (Adh and the CFGs discussed here fit this pattern.). Third, the shift in protein sequence may be accompanied by a shift in expression pattern. Although too little is known to answer this question conclusively, the current data suggest that a shift in expression pattern may occur but is not required. Jingwei has a much more limited expression pattern than Adh, although Adh can be found in most of the tissues where Jingwei is expressed. Like Jingwei, Adh-Finnegan in some species is expressed only in a subset of the tissues that express Adh. Adh-Twain, however, appears not to have shifted its expression, because Adh-Twain, like Adh, is broadly expressed. Nevertheless, the fine-scale expression patterns of these genes are too poorly known to for a strong test for a shift in expression. Many of these questions will become tractable as more genomes are sequenced and more CFGs are discovered.

We thank M. Antezana for suggesting the structural hypothesis for divergence; members of the Begun laboratory and the Evolution Discussion Group at the University of California, Davis; M. Long and K. Thornton for thoughtful comments on an earlier draft of the manuscript; K. Thornton for a helpful statistical suggestion; and two insightful reviewers for several suggestions that improved the manuscript and helped us think critically about our shift model. This work was supported by National Science Foundation grants (to C.D.J. and MCB 9973804 to D.J.B.).