# High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes

Penka Markova-Raina and Dmitri Petrov

*Genome Res.* 2011 21: 863-874 originally published online March 10, 2011
Access the most recent version at doi:10.1101/gr.115949.110

# High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes

Penka Markova-Raina[1] and Dmitri Petrov

*Department of Biology, Stanford University, Stanford, California 94305, USA*

We investigate the effect of aligner choice on inferences of positive selection using site-specific models of molecular evolution. We find that independently of the choice of aligner, the rate of false positives is unacceptably high. Our study is a whole-genome analysis of all protein-coding genes in 12 *Drosophila* genomes annotated in either all 12 species (~6690 genes) or in the six *melanogaster* group species. We compare six popular aligners: PRANK, T-Coffee, ClustalW, ProbCons, AMAP, and MUSCLE, and find that the aligner choice strongly influences the estimates of positive selection. Differences persist when we use (1) different stringency cutoffs, (2) different selection inference models, (3) alignments with or without gaps, and/or additional masking, (4) per-site versus per-gene statistics, (5) closely related *melanogaster* group species versus more distant 12 *Drosophila* genomes. Furthermore, we find that these differences are consequential for downstream analyses such as determination of over/under-represented GO terms associated with positive selection. Visual analysis indicates that most sites inferred as positively selected are, in fact, misaligned at the codon level, resulting in false positive rates of 48%–82%. PRANK, which has been reported to outperform other aligners in simulations, performed best in our empirical study as well. Unfortunately, PRANK still had a high, and unacceptable for most applications, false positives rate of 50%–55%. We identify misannotations and indels, many of which appear to be located in disordered protein regions, as primary culprits for the high misalignment-related error levels and discuss possible workaround approaches to this apparently pervasive problem in genome-wide evolutionary analyses.

[Supplemental material is available for this article.]

The large amounts of sequence data generated in the past few years have led to a burst in studies examining important open questions about protein evolution (Nielsen et al. 2005; Chen et al. 2006; Savard et al. 2006; Anisimova et al. 2007; *Drosophila* 12 Genomes Consortium 2007; Heger and Ponting 2007; Kawahara and Imanishi 2007; Vieira et al. 2007; Kosiol et al. 2008; Larracuente et al. 2008; Studer et al. 2008; Lefebure and Stanhope 2009; Dickson et al. 2010; Kunstner et al. 2010). Many of these studies use as a basic premise the alignment of nucleotide and amino acid sequences obtained from different species. Of course, the true alignments are themselves unknown and are typically inferred by one of a number of publicly available alignment programs ("aligners"). It is well understood that these aligners are not perfect (Thompson et al. 1999; Lassmann and Sonnhammer 2002; Nuin et al. 2006; Golubchik et al. 2007; Kemena and Notredame 2009; Morrison 2009). However, to keep the analyses manageable, the estimated alignment is commonly assumed to be the "ground-truth," and the effects of errors in the alignments are rarely controlled for.

Alignment errors are especially tricky in the context of large-scale studies that can include thousands of genes. With such data sets it is intractable to curate or visually inspect a substantial portion of the alignments, as might have been done in the past with smaller data sets. At the same time these are among the most exciting studies, as they can uncover mechanisms difficult to detect at a smaller scale and can lead to important and highly influential conclusions. They are also becoming increasingly common (Bakewell et al. 2007; *Drosophila* 12 Genomes Consortium 2007; Kosiol et al. 2008; Lefebure and Stanhope 2009; Kunstner et al. 2010). It is therefore crucial to understand the prevalence and impact of any existing alignment problems in such studies and to ensure that they do not significantly affect the conclusions.

Not all alignment errors are created equal; their importance depends on the specific question being investigated. For instance, although misalignment of several amino acids within a long gene might not affect the inference of the phylogeny, it could still result in misinference of positive selection at the site of the misalignment. While tentative information about the accuracy and sensitivity of alignment programs is often available (Nuin et al. 2006), it is not necessarily obvious how they would affect each such separate use of the alignments. Previous work has reported significant discrepancies related to aligner choice in phylogeny and site-model selection inference in yeast (Wong et al. 2008) as well as in branch-site model simulations based on mammalian and vertebrate genes (Fletcher and Yang 2010). Errors in alignment were one of the contributors to the high levels of false-positive inference of selection based on branch and branch-site models that have been reported in the mammalian genomes (Mallick et al. 2009; Schneider et al. 2010); they were also observed in the branch-site model simulations (Fletcher and Yang 2010).

Here we explore the effect of alignment errors on the inference of positive selection based on site-specific divergence models, in particular, the commonly used PAML models M7, M8, and M1a, M2a. These metrics might be especially sensitive to alignment problems because both positive selection and misalignments often generate signatures of exceptionally fast evolution. We find that in the 12 *Drosophila* genomes data the results are highly dependent

**[1]Corresponding author.**
**E-mail penka@cs.stanford.edu.**
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.115949.110.

on which alignment program was used. Furthermore, for all of the considered aligners, the rate of false positives caused by an erroneous alignment of nonhomologous codons (as observed during visual inspection of genes with inferred selection) is consistently high, from ~45% to ~80%. We discuss some of the problems and sequence features associated with these alignment errors and conclude that the prevalence of false positives is higher than is likely to be acceptable. The effects of aligning nonorthologous codons must therefore be taken into account before making any inference of site-specific positive selection.

## Results

### Alignments

Our analysis is based on the GLEAN-R reconciled consensus set of predicted gene models in the 12 *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007). We included the longest annotated transcript of each gene and only included genes with an annotation of a single-copy ortholog in all 12 species. The sequences were aligned using five commonly used aligners: AMAP, ClustalW, MUSCLE, ProbCons, and T-Coffee (note that we used the T-Coffee alignments of *Drosophila* 12 Genomes Consortium 2007) (see the Methods section for details). We found a strong correlation between basic properties of alignments generated by all five aligners, such as length and percent alignment identity (Kendall τ rank correlation coefficients ranges between 0.93 and 0.99 for length, and between 0.84 and 0.96 for percent identity). Nevertheless, for some of the genes these properties did vary substantially depending on the aligner used (Supplemental Fig. 1S, c and d). It is also evident that there is a systematic bias in the differences (Supplemental Fig. 1S). The problems that these differences allude to are not hypothetical or inconsequential, as the analysis below will show.

### Rate of protein evolution

The rate of protein evolution was inferred with PAML Model 0 (Yang 1997), ω (or $d_n/d_s$ ratio) of the genes in the 12 species data set varies from $10^{-4}$ to 0.86, with a mean of 0.1 and a median of 0.08. The estimates derived based on the five different alignments are highly correlated (Kendall τ rank correlation coefficients ranging between 0.91 and 0.96, and Spearman's ρ between 0.98 and 0.99). However, there are clear systematic differences in the distributions. For example, the alignments made with AMAP tend to have lower substitution rates (Fig. 1). This is likely related to the fact that they are the least compact among the five alignments, with comparatively longer alignments, longer gaps, and with higher percent identity of the aligned portions (Supplemental Fig. 1S, a and b). Comparing the distributions via Wilcoxon's rank sum test and Kolmogorov-Smirnov goodness-of-fit hypothesis test also indicates significant differences in some cases (AMAP vs. any other aligner, ProbCons vs. T-Coffee or ClustalW).

In the context of the 12 *Drosophila* genomes and the data set used, the ω metric by itself is not useful for detecting positive selection. None of the genes had ω greater than 1 (which would have been indicative of positive selection), and only five genes had ω >0.5. The lack of higher rates might be due to the absence of such genes in the 12 *Drosophila* genomes, or because the genes that evolved fast on all branches either could not be detected and annotated correctly in all 12 species or were lost or duplicated in some

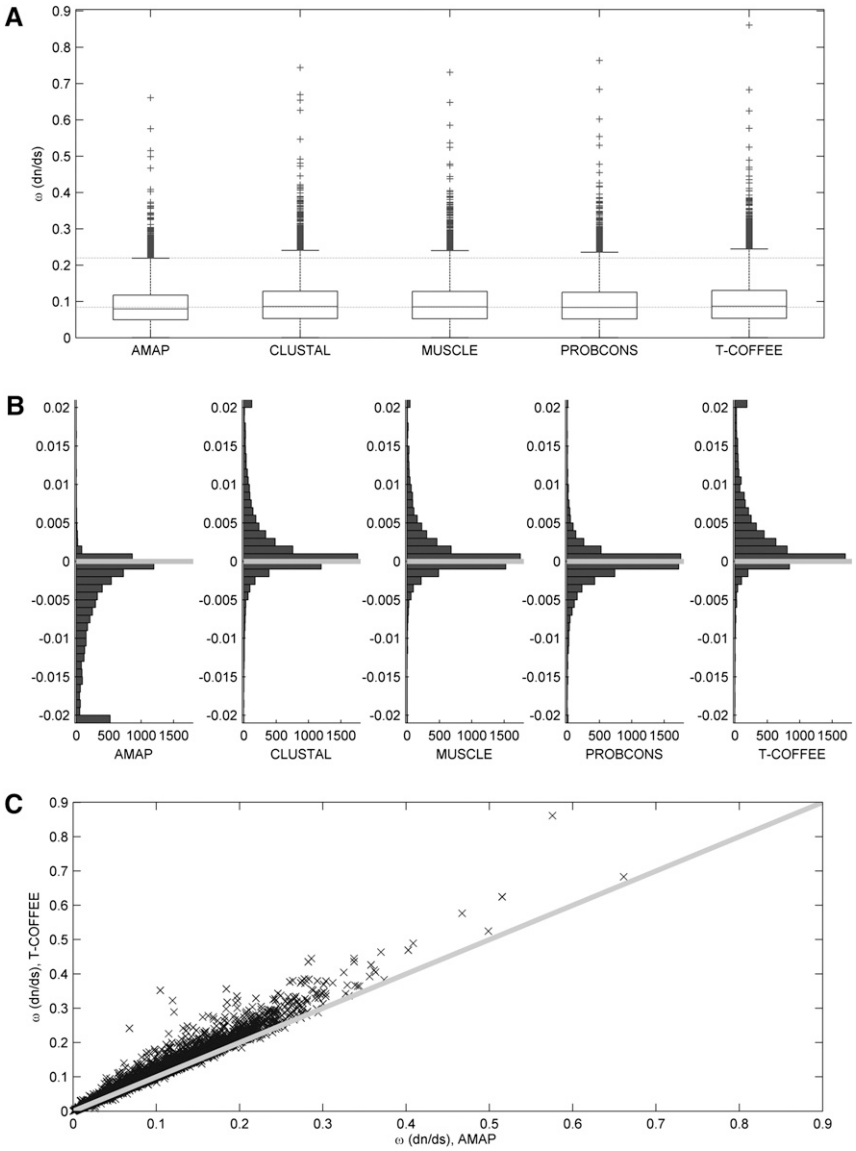


**Figure 1.** Impact of aligner choice on the estimated rates of protein evolution. (*A*) Box plot of ω, or $d_n/ds$,$K_a/K_s$, values per gene depending on the aligner used. (*B*) Frequency distribution of the difference per gene between ω based on a given aligner and the average ω among all five aligners. (*C*) ω of AMAP vs. T-Coffee.

of the species. The impact of aligner choice in the presence of higher rates of protein evolution remains to be determined.

## Genes with positively selected sites

Sites under positive selection were inferred based on the commonly used PAML (Yang 1997) models M1a (neutral model) and M2a (positive), and M7 (neutral) and M8 (positive) (see Methods section). These models allow the ratio of nonsynonymous to synonymous substitutions, ω, to vary among sites, and to treat the ω ratio of each site in the gene as a random variable drawn from a statistical distribution. The neutral evolution models assume ω drawn from a beta distribution (model M7) or from one of two categories: $\omega < 1$ or $\omega = 1$ (model M1a). The positive selection models, M8 and M2a, add a category of $\omega > 1$ to the corresponding neutral model. Whether the positive selection model is statistically significantly more likely than the neutral model is determined by computing the likelihood ratio test (LRT) statistic (i.e., twice the log likelihood difference between the compared models) and comparing it with the $\chi^2$ distribution with two degrees of freedom. The posterior probability that a site was positively selected was determined with the Bayes empirical Bayes method (Yang et al. 2005).

We found that the number of genes inferred to have at least one positively selected site varies substantially with aligner choice (Table 1). For instance, when using Models M7 and M8 and a 95% posterior probability cutoff (Table 1A), the number of such genes varies by ~60% (from 817 to 1290). A switch from ClustalW to T-Coffee, which are two of the most commonly used aligners, corresponds to a 43% increase in the count. The percentage of genes with a selected site ranges between 12% (AMAP) and 19% (T-Coffee), and, overall, 28% of the genes of the 12 species set have a selected site in at least one of its five alignments. These discrepancies are even more pronounced when models M1a and M2a are used and also when the posterior probability cutoff is more stringent at 99%. In addition, even when there is agreement in that selection inferred in a gene, the identity of the inferred sites often varies (data not shown). For instance, some of the genes with inferred sites under selection have multiple such sites, and depending on the aligner there is a variation in the number of sites and their location within the protein (Supplemental Table 2SA, d and g, SB, d and g). Increasing the required number of inferred selected sites to two or three did not change the differences in numbers of genes or the level of consistency among the genes picked by different aligners (M7 and M8 at 95%, data not shown).

Furthermore, we found that the differences due to aligner choice go beyond what is apparent from the difference in numbers in Table 1. Even when the count of genes is similar, it can represent different genes. For each gene with inferred site(s) under selection, Figure 2 shows the number of aligners in whose alignment there was such a site(s). With models M7/M8 and at 95% cutoff, 684 or 36% of the 1902 genes containing a selected site had such a site only in one of the five alignments. At the same time, all five aligner results agree on only 413 (22%) of the genes. This is just 51% of the genes identified in AMAP alignments (the aligner with the lowest count among all five aligners) and 32% of those identified with T-Coffee alignments. At the 99% cutoff, these percentages drop to 48% and 16%, respectively. Note that pairwise comparison shows that these differences persist for all aligner pairs with the overlap ranging between 48% and 86% (Supplemental Table 1S). We performed similar experiments with models M1a and M2a and obtained consistently low overlaps among genes inferred to be under positive selection, depending on the alignment (Table 1, c and d; Supplemental Fig. 3S, a and b).

The number of sites inferred to be under positive selection where the alignment was consistent among all five aligners appears to be very low. We used the CORE consistency scores of the T-Coffee -evaluate_mode option (Notredame and Abergel 2003; Kemena and Notredame 2009), which combines the five previously generated alignments and computes a consistency score for each site in each species (Methods). Only between 4% (T-Coffee) and 18% (AMAP) of the sites inferred to be under positive selection with models M7 and M8 and a cutoff at 95% had the highest possible consistency (score 9 in all species at this site). These corresponded to only 22%–36% of the genes with an inferred positively selected site.

## Impact of removing regions with gaps

Removing regions of the alignments that have a gap (observable deletion or insertion in some of the 12 species, or missing data) before performing PAML analysis is common. It is often done with the hope that the remaining sequence has a better quality alignment and, thus, the results are more reliable. We investigated whether removing these regions would result in a higher consistency in the number of genes with positively selected sites when comparing the alignments generated by the five studied aligners. We found this not to be the case (Table 1, e and f).

When gaps are removed, the number of genes estimated to have a positively selected site was lower compared with when the full sequence was used. This is not unexpected and might be due to some of the positively selected sites having been in these removed "gap" regions, or possibly due to the loss of data affecting the power of PAML to make a statistically significant conclusion. Interestingly, however, we found that ~30% of the genes with a positively selected site were new, i.e.,

**Table 1. Number of genes inferred to have a positively selected site(s). Inference is based on PAML models M7 and M8 (a and b), models M1a and M2a (c and d), and models M7 and M8 after removing gaps (e and f) for 12 species alignments generated with each of the six aligners; and models M7 and M8 for *melanogaster* group alignments (g and h)**

| | 12 genomes, M7/8 | | 12 genomes, M1a/2a | | 12 genomes, M7/8, with removed gaps | | *Melanogaster* group, M7/8 | |
|---|---|---|---|---|---|---|---|---|
| **Aligner** | **95% (a)** | **99% (b)** | **95% (c)** | **99% (d)** | **95% (e)** | **99% (f)** | **95% (g)** | **99% (h)** |
| AMAP | 817 | 213 | 256 | 110 | 558 | 104 | 973 | 257 |
| MUSCLE | 1043 | 306 | 379 | 192 | 764 | 155 | 1134 | 366 |
| ProbCons | 1013 | 281 | 346 | 180 | 801 | 182 | 1128 | 371 |
| T-Coffee | 1290 | 479 | 612 | 353 | 824 | 173 | 1248 (909) | 463 (218) |
| ClustalW | 902 | 261 | 244 | 117 | 666 | 112 | 1269 | 453 |
| Total in 5 | 1902 | 673 | 799 | 441 | 1562 | 384 | 1737 (1723) | 652 (620) |
| PRANK | 468 | 49 | 49 | 16 | 258 | 42 | 581 | 70 |

"Total" indicates the number of genes that were in at least one of the aligners' counts; numbers in brackets in g and h indicate the number of genes if the "masked" instead of unmasked T-Coffee alignments are used. The significance cutoffs (95% and 99% posterior probability) are indicated on the second row.

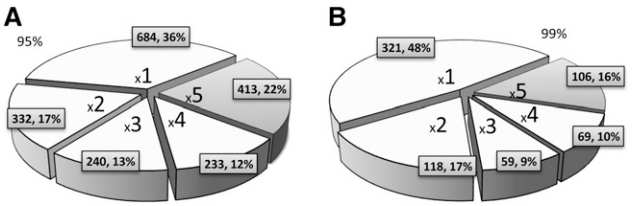


**Figure 2.** Number of genes that were predicted to have a positively selected site with cutoff posterior probability set at 95% (*A*) and 99% (*B*) by exactly 1, 2, 3, 4, or 5 aligners when using PAML models M7 and M8 (out of the genes with at least one such prediction).

they were not detected in the previous analyses where the gapped regions were included.

In addition, note that regions that are retained after gap removal must be annotated as protein-coding in *D. melanogaster*. The quality of *D. melanogaster* gene annotations is high, at least comparatively speaking, among the best one can hope for currently. Alignment errors in the ungapped regions are thus less likely to be due to erroneous annotations of noncoding sequences in some of the 12 *Drosophila* species, although it could still include some errors where a coding region in *D. melanogaster* was aligned to a noncoding sequence from another species.

## Genes inferred to be under selection based on the LRT test

The above analyses relied on the detection of individual sites evolving under positive selection. One might hope that the inference based on a more global statistic such as overall LRT score would be more robust. To test this hypothesis we used three commonly used methods for determining significance of the likelihood ratio statistics. The first approximated the probability distribution of the LRT statistic by a $\chi^2$ distribution with two degrees of freedom, as recommended in the PAML user manual (Table 2, a and b, below). This method does not correct for multiple tests and is not appropriate for a whole-genome analysis—however, it does reflect the differences that might be observed if independent researchers performed tests on the same genes. Overall, ~45% of the genes have inferred selection at significance level 0.05 in at least one of their five alignments; more than half of these would be inferred to be under selection in only three or less of the five alignments, and only 14% in all five alignments (Supplemental Table 4S).

The second method again used a $\chi^2$distribution with two degrees of freedom, but additionally corrected for multiple tests by using the Bonferroni correction (Table 2, c and d). In the third method we calculated q-values and controlled the False Discovery Rate (Table 2, e–g). In both of these cases, the set of genes passing the cutoff threshold shows significant variations.

Finally, the genes with the highest LRT statistic are often of special interest, as they are often believed to be the least susceptible to inference errors and the most attractive targets for further investigation of positive selection. When we compared the 100 genes with the highest LRT for any one aligner, we found that, on average, only 54 of these genes were common when another aligner was used. In the case of T-Coffee, whose alignments were released with the GLEAN-R set (*Drosophila* 12 Genomes Consortium 2007), the overlap with any one of the other four aligners ranged between 35% and 51%.

## Enrichment in GO categories

The GO categories that are over- or under-represented in genes inferred to be under positive selection are of interest and are often reported (*Drosophila* 12 Genomes Consortium 2007; Heger and Ponting 2007; Kawahara and Imanishi 2007). We found that aligner choice affects such analyses as well. For example, the number of noredundant GO Biological Process level 3 terms with significant (FDR-based *P*-values of <0.05) over- or under-representation that we found with FatiGO (Al-Shahrour et al. 2004) ranged from one with AMAP, to five with ClustalW and ProbCons, to seven with T-Coffee, and 15 with MUSCLE. The number of significant GO molecular function level 3 terms ranged between zero and three; there was only one significant GO Cellular Component level 3 term that was independent of aligner and one that was found only for the T-Coffee alignments. Terms that were significant in only some of the alignments included *learning and/or memory* GO:0007611 (only in T-Coffee alignments), *response to drug* GO:0042493 (MUSCLE only), *adult behavior* GO:0030534 (T-Coffee, MUSCLE), etc. When we considered only sites inferred to be under selection with 99% posterior probability instead of 95%, the discrepancies were less pronounced (most likely due to loss of statistical power), but were still present. For example, the number of significant GO biological process level 3 terms was zero with AMAP, two with MUSCLE and ProbCons, and six with ClustalW and T-Coffee. A less-stringent cutoff at *P*-value of 0.1 resulted in a range between zero and 18 terms.

## Visual inspection of 12 species alignments with inferred selected sites

To better understand the reasons for the observed differences we visually inspected the alignments and the attendant inferred positively selected sites in a subset of genes (Fig. 3). Sites with obvious signs of unreliable alignment, such as presence of runs of divergent sites at the protein level or of gaps that were placed differently by the five aligners were designated as "misaligned" (see examples in Supplemental Material). We sliced the data set and picked genes with inferred sites in three different ways, aimed, correspondingly, at the questions of: (1) Does consistency among aligners in the downstream inference of selection indicate a lower rate of error in the positive selection inference, (2) what is the rate of false positives for each aligner, and (3) did any of the five aligners

**Table 2.** Number of genes in the 12-species gene set inferred to be under positive selection at the gene level. Inference is based on: (a, b) comparing LRT with the $\chi^2$ distribution without correction for multiple tests; (c, d) comparing LRT with the $\chi^2$ distribution with a Bonferroni correction for multiple tests; and (e, f, g) *q*-values and FDR (see Methods)

| | $\chi^2$ test | | Bonferroni correction | | FDR | | |
|---|---|---|---|---|---|---|---|
| Aligner 12 species | 0.05 (a) | 0.01 (b) | 0.05 (c) | 0.01 (d) | 0.1 (e) | 0.05 (f) | 0.01 (g) |
| AMAP | 1582 | 1053 | 271 | 226 | 1273 | 1053 | 655 |
| MUSCLE | 1875 | 1309 | 442 | 374 | 1717 | 1413 | 968 |
| ProbCons | 1763 | 1195 | 370 | 309 | 1577 | 1268 | 830 |
| T-Coffee | 2238 | 1714 | 758 | 671 | 2220 | 1920 | 1421 |
| ClustalW | 1634 | 1144 | 401 | 357 | 1410 | 1172 | 805 |
| PRANK | 1112 | 568 | 44 | 31 | 546 | 361 | 110 |

The significance cutoffs (0.1, 0.05, or 0.01) are indicated in the second row.


**Figure 3.** Example alignment from the genes that we performed visual inspection on. The illustrated alignment segment is from the *melanogaster* group alignments of gene *FBgn0036686*. The *top* alignment was made with T-Coffee and has three sites inferred under selection at 99% cutoff. The *bottom* alignment was made with ProbCons; the highest posterior probability in this region for the ProbCons alignment is 60%.

perform significantly better in this context? The complete selection process is described in the Supplemental Materials. For each of the picked genes we visually inspected all sites that were inferred to be positively selected at a 95% cutoff posterior probability in any of the five alignments; the results are summarized in Supplemental Table 2S, A–C.

For all but eight of a total of 83 visualized genes, all sites inferred to be under positive selection included codons that were likely misaligned in one or more species (these are the genes categorized as "M" or "~M" in Supplemental Table 2S, e). The rate of false positives ranged from 71% to 82%, depending on the aligner used. In seven out of the eight apparent true positives, the site under positive selection was detected by PAML using all five alignments. Note, however, that identifying a positively selected site using all five alignments was not a guarantee of a true positive. In fact, 12 out of 19 such cases were clear misalignments (Supplemental Table 2S, A–C, c). Overall, we found no evidence that any of these five aligners generate substantially better results or that consistency of PAML results across aligners can be used as evidence for true positive selection.

### *Melanogaster* group species alignments and impact of masking

Our visual inspection of sites inferred to be under positive selection based on the 12 species alignments indicated a surprisingly high number of false positives and raised a number of questions. Would this still be a problem if only more closely related (and less diverged) species are included in the alignments? Also, we had refrained from applying handcrafted masking and other controls that might normally be done during analysis. We reasoned that this would require customization that would interfere with the ability to cleanly compare the differences caused by the alignment programs. Would applying such typical quality controls have been sufficient to resolve the observed problems?

To examine whether our findings hold in such cases also, we repeated some of the statistical analyses and the visual inspection on genes inferred to be under positive selection in previously published work that included masking and analysis of only closer related *Drosophila* species (*Drosophila* 12 Genomes Consortium 2007; Larracuente et al. 2008). The selection inference analysis in those studies was based on the same gene models that we used; however, only the six *melanogaster* group species were included. The coding sequences had been aligned using T-Coffee and, additionally, two separate steps of masking had been applied for quality control. Thereafter, positive selection was inferred at the gene level based on PAML models M7 and M8 similar to our earlier analysis; most conclusions referred to genes inferred to be under selection with a FDR of 10%.

We first investigated whether positive selection inference using unmasked alignments in the six *melanogaster* group species was more robust than for the 12 species. Comparison of the number of genes with an inferred selected site yielded counts varying by ~30% at 95% cutoff and by as much as ~80% at 99% cutoff (Table 1, g and h). However, the percentage of genes with agreement among at least four of the five alignments was much higher compared with the 12 species alignments, 53% vs. 34% at 95% cutoff and 44% vs. 26% at 99% cutoff (Fig. 2A,B; Supplemental Fig. 2S, e and f). When we repeated the analysis at the gene level (i.e., without requiring that any specific sites are unambiguously identified to be evolving under positive selection with high posterior probability), the number of genes inferred to be under positive selection with a FDR cutoff at 10% ranged between 676 (AMAP) and 1080 (T-Coffee); 691 of these genes were inferred as positively selected—in at least four alignments (Table 3, e). Therefore, overall, the inconsistencies related to the choice of aligner remain, albeit to a lesser degree, even when only the six closer-related *Drosophila* species were included.

We next investigated the impact of each of the two masking steps that had been applied in the published studies. The first step comprised of masking sequences where the divergence between *D. melanogaster* and any non-*melanogaster* species exceeded a carefully determined species-specific cutoff; the resulting alignments were referred to as "masked (GLEAN-R) alignments" in the publication *Drosophila* 12 Genomes Consortium (2007). Applying this masking significantly lowered the prevalence of inferred positive selection (Tables 1, 3). The number of genes with at least one site under selection at 95% significance cutoff dropped by 27%; a 99% cutoff resulted in a 53% drop (Table 1, g and h). When inference was at the gene level, a $\chi^2$ *P*-value cutoff of 0.05 and a FDR at 10% cutoff resulted correspondingly in drops of 20% and 46% (Table 3, a and e). Overall, 32%–58% of the particular genes inferred to be under selection in the unmasked alignments were no longer inferred as such after masking.

The second masking step involved "trimming" the "masked alignments" such that codons missing or masked in more than one species were removed. This step was not described in *Drosophila* 12 Genomes Consortium (2007); however, it had, in fact, been applied prior to the positive selection analysis (TB Sackton, pers. comm.). Unsurprisingly, "trimming" further reduced the prevalence of inferred positive selection (Table 3). For example, when selection inference was based on a $\chi^2$ *P*-value cutoff of 0.05 and FDR at 10% cutoff, the reduction was, correspondingly, 25% and 60%. The reduction in the cases with stricter cutoffs, however, was unexpectedly high, ranging between 60% and 80% for the Bonferroni and FDR cutoffs we included (Table 3, c-g).

While the overall procedure for inferring positive selection used in *Drosophila* 12 Genomes Consortium (2007) was similar to what we did in this study, it differed in how FDR q-values were determined. In that case, the *P*-values had been determined based on simulations, while our approach utilized the $\chi^2$ distribution. To roughly compare those results with ones that could have been observed if other aligners were used, we used the observation that the 10% FDR cutoff in *Drosophila* 12 Genomes Consortium (2007) corresponds to a LRT of 6.03 (based on the raw data available with

**Table 3. Number of genes in the *melanogaster* group gene set inferred to be under positive selection at the gene level. Inference is based on: (a, b) comparing LRT with the $\chi^2$ distribution without correction for multiple tests; (c, d) comparing LRT with the $\chi^2$ distribution with a Bonferroni correction for multiple tests; (e, f, g) *q*-values and FDR (see Methods); and (h) LRT ≥6.03, corresponding to the cutoff LRT for the genes significant for positive selection with FDR of 10% in *Drosophila* 12 Genomes Consortium (2007)**

| | $\chi^2$ test | | Bonferroni correction | | FDR | | | LRT |
|---|---|---|---|---|---|---|---|---|
| Aligner Mel group | 0.05 (a) | 0.01 (b) | 0.05 (c) | 0.01 (d) | 0.1 (e) | 0.05 (f) | 0.01 (g) | 6.03 (h) |
| AMAP | 1230 | 706 | 190 | 156 | 676 | 511 | 335 | 1222 |
| MUSCLE | 1394 | 877 | 305 | 258 | 927 | 714 | 471 | 1379 |
| ProbCons | 1344 | 835 | 289 | 248 | 854 | 672 | 441 | 1338 |
| ClustalW | 1507 | 973 | 380 | 335 | 1065 | 858 | 608 | 1495 |
| T-Coffee, unmasked | 1490 | 982 | 389 | 355 | 1080 | 871 | 610 | 1480 |
| T-Coffee, masked | 1191 | 642 | 150 | 134 | 583 | 405 | 250 | 1184 |
| T-Coffee, masked and trimmed | 890 | 395 | 38 | 30 | 233 | 119 | 51 | 879 |
| PRANK | 828 | 361 | 22 | 14 | 140 | 83 | 35 | 822 |

The significance cutoffs are indicated in the second row.

the article), which is similar to the $\chi^2$results at *P*-value of 0.05. Comparison of the genes inferred to be under selection using this cutoff (Table 3, a and h) confirms that if a different alignment software was used, the genes inferred to be under selection would likely be different, both compared with the unmasked, masked, and "masked and trimmed" T-Coffee alignments. In addition, among the 890 genes inferred to be under selection at 0.05 significance values in the "masked and trimmed" alignments, only 57.5% were inferred with all of the other four aligners; 14.5% were unique to the T-Coffee "masked and trimmed" alignments (Supplemental Table 5S). The substantial differences due to aligner choice and to the quality control steps highlight the fact that different methodologies can have a crucial impact on the estimates of positive selection.

The much lower number of genes inferred to be under selection after masking and/or trimming gives a glimpse of hope that maybe this method has been successful in removing most of the false positives. To investigate this possibility, we performed visual inspection on 25 of the genes inferred in *Drosophila* 12 Genomes Consortium (2007) to be evolving under positive selection at a 1% FDR (as per the q-values data available with that article). For these genes, we reran PAML and inspected all sites inferred to be under positive selection. Visualization was done both on the "masked" alignments, as well as later on the "masked and trimmed" alignments. We found that the problem with false positives here is not as extreme as with the 12 species unmasked alignments, but it remains significant (Supplemental Table 2S, D). Some example visuals from the "masked" alignments are included in the Supplemental Materials. Out of the 25 genes, we annotated 12 (48%) as highly likely to have been inferred to be under selection as a result of an incorrect codon alignment at the site(s) under selection. Furthermore, of the remaining genes, seven have multiple sites, some of which we considered poorly aligned and which were likely to affect the statistical significance. Therefore, our inspection suggests that as many as half of the genes that were inferred to be under positive selection might, in fact, be false positives, even after restricting to six species and applying widely used quality control steps.

## Issues underlying the high levels of false positives

During the visual inspection we noticed a number of repeating problems affecting the inference of positively selected sites, and some representative cases are included in the Supplemental Materials. Common issues causing the misalignments included bad alignments at the CDS start and end related to annotation problems, misinference of intron positions, existence of alternative splicing, amino acid repeats, and the presence of indels in fast evolving pockets located in between well-conserved neighborhoods. Supplemental Tables 2SA, j, B, j, C, j, and 3S, j indicate cases where we observed any of these features in the visualized genes.

We were particularly intrigued by what appeared to be pockets of fast-evolving sequence subregions located within well-conserved neighborhoods, and whether the underlying sequences represented true biological phenomena of localized fast evolution or were artifacts of annotation. To answer this question we examined 15 of the previously visualized genes in which the incorrectly inferred sites under positive selection were located in such fast-evolving subregions (Supplemental Table 3S; Methods). These genes were picked from our analyses of all 12 *Drosophila* species in a way that all aligners, as well as all of the different aligner consistency cases, were represented. The selected sites for seven out of the 15 genes were located immediately near an exon border, and the sites in at least five of these genes were clearly due to a difference in annotation. To further validate this observation we screened all inferred selected sites in all of the genes in 12 species alignments to see whether they were located within 15 amino acid sites of an exon border in the corresponding alignment. There were 285, 338, 360, 416, and 483 genes with such a site correspondingly in AMAP, ClustalW, MUSCLE, ProbCons, and T-Coffee, e.g., between 35% and 41% of the genes with an inferred selected site. Therefore, it appears that annotation differences are an important contributor to false positives in this data set.

For seven of the eight remaining genes, we found published full-length cDNA or 5′/3′ ESTs fully containing the *D. melanogaster* sequence of the fast-evolving subregion. Therefore, we conclude that at least in approximately half of the fast evolving subregions, we observed a genuine fast-evolving sequence.

Furthermore, our visual inspection of the fast-evolving pockets, and especially the ones not located near the exon–intron borders, gave us an impression that these regions often did not have a random sequence composition, specifically containing runs of repeating amino acids and elevated numbers of some amino acids. Because these patterns appeared consistent with those of protein disorder, we investigated whether sites inferred to be under selection were indeed preferentially located in intrinsically disordered regions (Supplemental Table 6S). To do this we first annotated disordered regions in all 12 *Drosophila* species proteins with IUpred (Dosztanyi et al. 2005). IUpred predicts disorder from amino acid sequences by estimating their total pairwise inter-residue interaction energy, and assigns a probabilistic score ranging from 0 (complete order) to 1 (complete disorder) for each amino acid. We used a cutoff value for a disorder of 0.5 as recommended in Dosztanyi et al. (2005). We found that 61%–85% of the sites inferred to be selected were disordered according to this definition if all 12 species sequences are accounted for, which was significantly higher than expected at random (45%–50%).

Lastly, we observed that in almost all cases the sites that were misinfered as positively selected were located close to indel transitions (start or end of an indel) (Supplemental Table 2S, A–C, i). Indels have been associated with both increased number of selected sites as well as alignment ambiguities (Loytynoja and Goldman 2008; Tian et al. 2008; Yang et al. 2009; Zhu et al. 2009; Chen et al. 2010). A computational analysis indeed revealed that in the 12 *Drosophila* alignments the majority (~85%) of the sites inferred to be under selection were located near an indel transition. Less than 15% (depending on the aligner used) of the genes with an inferred selected site did not have an indel within 10 amino acids of the inferred positively selected site. Further research is needed to clarify as to what degree the indels are causing misalignments and appearance of positive selection, and to what degree misalignments due to fast sequence evolution are causing the appearance of indels. In addition, restricting analysis to only inferred selected sites that are not located within indels or their flanking codons does not appear to resolve the lack of consistency among the aligners; e.g., after removing inferred selected sites in the prior analysis, which are in or within five codons of an indel, the resulting number of genes varies between 105 and 181, depending on the aligner.

### PRANK alignments

PRANK (Loytynoja and Goldman 2008), when executed with the "–codon" option, has recently been reported to outperform other alignment programs in simulations and, specifically, to produce alignments with fewer false positives during branch-site inference of selection for these alignments (Fletcher and Yang 2010). PRANK differs from the other aligners we considered in that it takes evolutionary information into consideration during DNA level codon alignment, and it also considers the evolutionary information in determining where to place gaps. The investigators hypothesized that this would likely also apply to site-specific models such as the ones we report on. Therefore, we repeated some of the analyses on the PRANK alignments to evaluate its performance on the 12 *Drosophila* data set.

Using PRANK consistently results in significantly lower prevalence of inferred positive selection compared with the other five aligners (Tables 1–3). There were 468 genes with an inferred site under selection with PAML models M7–M8 at 95% cutoff, and 49 genes at 99% cutoff—compared with 817 and 213 genes, respectively, with the aligner with next lowest numbers (AMAP). A similar pattern is observed with models M1a and M2a (Table 1, c and d), when gaps are removed (Table 1, e and f), in the *melanogaster* group alignments (Table 1, g and h) and when LRT-based statistics are used to identify genes evolving under positive selection (Table 2). In some of the cases the difference was an order of magnitude.

In the case of the *melanogaster* group, the counts after PRANK alignment were fairly similar to when the *Drosophila* 12 Genomes Consortium (2007) "masked and trimmed" alignments are used. However, these two procedures often identify different genes. For example, among the 890 genes inferred based on the "masked and trimmed" alignments with $\chi^2$ test, *P*-value 0.05 and the 828 genes inferred based on the PRANK alignments (Table 3, a), only 551 are in common. Note that 402 of these 551 are also identified as genes under positive selection by the other aligners (without masking) (Supplemental Table 5S, c and d). Similar differences remain when using the Bonferroni correction (Table 3, c).

Finally, we carried out visual inspection of the sites inferred to be under positive selection using PRANK alignments. We selected 20 genes at random, for which at least one site was inferred to be under positive selection (with models M7 and M8, 0.95 cutoff) (Supplemental Table 2S, D). Among these genes, only seven had at least one site that we classified as unlikely to be misaligned at the codon level; an additional two genes had partial misalignment, which would likely affect the significance level. The remaining 11 genes only contained inferred selected sites, which we classified as likely misaligned. We also re-examined the genes whose alignments in the other five aligners were previously visualized. Among these genes, PRANK had a total of eight genes with at least one correctly aligned inferred selected site, nine with only "likely misaligned" such sites, and one with questionable significance (Supplemental Table 2S, l). Therefore, all eight of the genes with a correctly aligned inferred selected site were picked up with the PRANK alignments too, which adds confidence that the false-negative rate in PRANK may not be higher than the other five aligners, while the false-positive rate is substantially reduced. In seven out of eight cases the exact site was inferred as selected in all of the remaining five alignments also, implying low ambiguity during the alignment at that position.

Overall, the rate of "likely misaligned" genes with PRANK alignments was ~50%–55% (55% for Supplemental Table 2S, D; 50% total among Supplemental Table 2S, A–C, l), implying that the level of false-positive inference of selection due to misalignments might be lower for PRANK alignments compared with the other aligners. However, the false-positive rate is still substantial and unacceptably high for most applications of such analyses. Favorably, the overall number of genes and sites are much lower with PRANK, and therefore a manual inspection and visualization would be more feasible in this case.

## Discussion

Codon substitution models provide a comprehensive framework for modeling how protein sequences evolve. These approaches can provide us with the inference of which proteins and which sites within proteins evolve under strong constraint and which underlie adaptation in different lineages. Programs like PAML (Yang 2007) do this reasonably efficiently and in a principled manner. The availability of genomic sequences in many related species has allowed these approaches to be applied in a high-throughput way, and the results of such studies have been extremely influential.

In this study we focused on testing the most basic assumption behind the application of codon substitution models, namely, that we know which codons are orthologous to each other based on the alignment. Indeed, it makes little sense to estimate the number of substitutions that happened between one codon and its neighboring codon, as might happen in an erroneous alignment. We have used coding sequence data from the 12 *Drosophila* genomes Consortium and aligned these sequences using six different aligners (AMAP, ClustalW, ProbCons, T-Coffee, MUSCLE, and PRANK). The alignments were then used to run five different PAML models and infer which genes and which sites within genes evolved under positive selection. Furthermore, we tested whether the comparisons of less divergent sequences were significantly less error prone by comparing error rates in the estimates of positive selection using either all 12 or the more closely related six *D. melanogaster* group species. We also investigated whether post-alignment processing such as gap removal, masking, and trimming would largely eliminate the misalignment and/or misinference of positive selection. In addition to statistical analyses, we carried out visual inspection of alignments for 128 genes. In all

cases, we found a false-positive error rate of the inference of positive selection above 45%, and in some cases above 80%, which is unacceptable for most applications. We also found, unsurprisingly, that the downstream analyses, such as detection of GO under- or over-representation in positively selected genes, are affected by this level of error.

Our findings dovetail with previous findings of high variability in selection inference due to aligner choice in yeast (Wong et al. 2008) and of high levels of false positives in mammals (Mallick et al. 2009; Schneider et al. 2010). Wong et al. (2008) reported that the estimates of phylogeny and the number of inferred positively selected sites in orthologous ORFs from seven yeast species were sensitive to alignment treatment. Specifically, using PAML models M1a/M2a, the number of sites inferred to be under positive selection at the 0.5 posterior probability cutoff varied by aligner choice in 28.4% of the yeast ORFs. Such sensitivity to aligner choice suggests a fairly high error rate, either of false positives or false negatives. In mammalian genomes, high levels of false positives in branch and branch-site model scans for positive selection have also been reported (Mallick et al. 2009; Schneider et al. 2010). These studies found similar underlying problems, as found in the 12 *Drosophila* alignments, with the addition of a high proportion of sites with a sequencing error. We did not focus on detecting sequencing errors, partly because the number of properly aligned positively selected sites was too small to allow such a study.

Among the aligners we have used, PRANK is unique in that it takes evolutionary information into consideration during alignment. Importantly, it has recently been shown to outperform, in simulations, a number of other alignment programs with respect to quality of the alignments and results in fewer false positives during branch-site inference of selection for these alignments (Fletcher and Yang 2010). While site models (such as M7, M8, M1a, M2a) were not examined by that study, the investigators hypothesized that their findings might extend to site models too. Our study confirms that using PRANK in conjunction with site models results in much fewer sites and genes inferred to be under positive selection and a lower level of false positives compared with other aligners. As new data from studies using PRANK becomes available, it is important to be aware that inference of selection using PRANK is not directly comparable to that from previous studies that used other aligners, including the cases in which previous studies have indicated higher levels of selection in a different group of species. Note that in the *Drosophila* data, the rate of false positives due to nonhomologous aligned codons is much higher than was observed in simulations by Fletcher and Yang, and at ~50% is likely to still be unacceptable for most applications.

The analysis we present (together with Wong et al. 2008; Mallick et al. 2009; Fletcher and Yang 2010; Schneider et al. 2010) makes the case that not only are there misalignments and other errors manifested as misalignments at the codon level, but that they might account for a significant number of the sites and genes that appear to be under positive selection. These errors can be significant enough to affect high-level conclusions such as the relative prevalence of genes with positively selected sites, and whether such genes are over- or under-represented in some GO categories. The reported cases span multiple species in different phyla (*Drosophila*, mammals, and yeast), species that are closely related as well as more diverged from each other, and multiple models of divergence-based selection inference. Furthermore, it is easy for such problems to propagate. For example, both the data (Huntley and Clark 2007; Greenberg et al. 2008; Clark and Aquadro 2009; Ridout et al. 2010) and high-level conclusions (Ellegren 2008; Koonin 2009; Singh et al. 2009) from the 12 *Drosophila* species selection inference analysis have been reused multiple times. At a time when whole-genome, large-scale analysis is becoming more prevalent, it is crucial that investigators are well aware of these caveats, so that they can take appropriate measures to avoid them or rephrase their question to avoid the caveats. In the 12 *Drosophila* genomes, in particular, the high levels of false positives we observe combined with the questions raised about the reliability of sites inferred to be under positive selection with site models (Nozawa et al. 2009) puts the reliability of some of these conclusions in question.

Interestingly, we observed that many of the false-positive sites were part of what appeared to be fast evolving pockets in the coding sequence. This was often the case both with the 12 species alignments as well as with the *melanogaster* group alignments (Supplemental Table 2S, A, i, and D, h). While about half of these were caused by annotation-related issues, the remaining half appeared to be due to genuine changes in coding sequences. Both protein disorder and presence of indels (not mutually exclusively) appear to be involved in many of these cases. Although such regions are not appropriate for analysis with PAML once they have diverged so far that they cannot be aligned reliably at the codon level, this does not mean that there has been no positive selection there, nor that these regions are not interesting to study. In fact, a number of studies have reported increased nucleotide diversity, mutation rate, substitution rate, or positive selection next to indels or protein boundary regions (Tian et al. 2008; Yang et al. 2009; Zhu et al. 2009; Chen et al. 2010), as well as positive selection for indel substitutions (Podlaha et al. 2005; Schully and Hellberg 2006). Further research is needed into whether such regions are as common as that observed in the small set we visually inspected and also into a more appropriate, noncodon-based methodology to investigate them. Note that these, as well as other regions of bad alignments, can lead to appearance of spatial clustering of positively selected codons (such as, for example, in the alignments of *FBgn0032627*, *FBgn0050166*, and *FBgn0039025* included in the Supplemental Materials, all of which contain a sequence of misaligned codons, resulting in a clustering of false positives). This might impact the conclusion by *Drosophila* 12 Genomes Consortium (2007) that 63.7% of the genes inferred to be evolving by positive selection at a 10% FDR show evidence for such clustering.

It is notable that the gene annotations and T-Coffee alignments that we considered are the result of a substantial community effort that is rarely available when such sets are constructed and analyzed. The gene annotations in *D. melanogaster* are among the best annotations currently available. Gene annotations for the remaining 11 genomes were predicted independently by multiple research groups, and the GLEAN-R set is a reconciled consensus set based on eight such predictions. It has been verified that most of the predicted single-copy orthologs are expressed (based on microarray experiments on adult flies in six of the species) (*Drosophila* 12 Genomes Consortium 2007). The phylogeny of the 12 species is known (except for, possibly, *D. yakuba* and *D. erecta*) (Pollard et al. 2006) and was used when creating the T-Coffee alignments. In the case of the published PAML *melanogaster* group-based analysis, a customized masking and filtering was applied. While it is clear that errors in sequencing, annotation, and alignment all contribute to the high level of false-positive inference of selection observed (among what should have been sites at the highest confidence of positive selection), they cannot be attributed to lack of effort or lack of state-of-the-art knowledge. Therefore, it is unlikely

that the problems we find are unique to the GLEAN-R data set. On the contrary, they might be more pronounced when fewer resources are available and such thorough efforts are not possible.

Our study focused on false positives. However, errors of codon alignment could have similarly resulted in false negatives. Codon misalignments cannot only raise the nonsynonymous rate estimated for a site, but could also raise the synonymous rate or lower the nonsynonymous rate (both resulting in a lower estimated ω ratio). Furthermore, even sites located in well-conserved, well-aligned neighborhoods can be affected: This is because the PAML-positive selection model M8 only estimates one value for the positive selection rate (ω > 1) $\omega_s$. A number of poorly aligned codons could change the estimate for $\omega_s$ substantially and, subsequently, could affect the probability of a particular site being in a neutral or the selected class. This would be especially relevant when more than one site in a gene appears to be under positive selection.

We have focused on the problems arising out of incorrect or dubious alignments and of the choice of alignment software. This appears to be a discouraging statement for future research given that the use of alignment software for evolutionary analysis of genomic sequences is indispensable and unavoidable, and that the underlying problems related to sequence quality and annotations are similarly complex to resolve. However, in our view, all is not lost—we merely advocate evolutionary analyses that are particularly mindful of the possibility of these misalignments. Applying quality controls such as masking, as well as using evolutionarily aware alignment programs like PRANK, is a good initial step in this direction even if insufficient to fully address the issue. One way to take this a step further would be to separate out the conclusions produced by such analyses into various confidence levels and focus on the most confident conclusions that are unlikely to be affected or caused by misalignments. For example, selected sites that are located in otherwise slowly evolving (and thus better conserved) regions can be considered separately from sites located in fast-evolving pockets or near indels. Our visual inspection strongly suggested that such fast-evolving pockets contribute significantly to the alignment error. Preliminary analysis of the impact of filtering the selected sites in the *melanogaster* group T-Coffee unmasked alignments based on the ω estimates of their neighboring sites indicates that this method can significantly improve the reliability of the inferred sites (in terms of codon homology). Limiting the allowed PAML ω estimate to, at most, 0.5 for all codons within a distance of 10 sites and limiting ω +SE to 1, yielded 120 genes with positive selection. Among 25 genes that we visually inspected, only three appeared to be due to a misaligned codon, which is a substantial improvement to what we observed in the 1% FDR genes that we had inspected (which, moreover, used masked alignments). This method has the advantage that it can reuse the estimates already produced by PAML.

Other methods such as the ones presented in Lunter et al. (2005) and Suchard and Redelings (2006), which estimate posterior probabilities of the columns of an alignment, can be also used for identifying reliably aligned regions and sites in a similar manner, when such highly computationally intensive approaches are not prohibitive given the data under consideration. Alternatively, masking approaches, which are less stringent in separating out reliable sites but are stricter than the "masking and trimming" method, might be acceptable for some applications, depending on the level of false positives that is tolerable and on the divergence of the included species. For example, the filtering methods used in Kosiol et al. (2008), which were based on syntenic pairwise DNA level alignments and a number of criteria including existence of frame-shift indels and the percentage of gaps, are much stricter than the ones in *Drosophila* 12 Genomes Consortium (2007). While such an approach, when appropriate, is likely to be affected by the choice of aligner, given that the locations and number of indels are dependent on the aligning algorithm, it would likely be affected to a lesser degree. Our visual inspection also revealed that many problems are due to either misannotations or pockets of fast evolution, possibly through frequent indels in intrinsically disordered regions. It is clear that these analyses will benefit from improved annotations and possibly from different approaches taken to the evolution of intrinsically disordered protein regions. The simplest interim solution would be to mask such disordered regions prior to alignment. In the future, different alignment procedures or methods for the inference of positive selection in disordered regions need to be developed.

We hope that our findings will reiterate the importance of alignment issues, provide useful insights, and encourage further study of methods for detecting and correcting the effects of dubious alignment regions and codons for site-specific evolutionary analysis.

## Methods

### Data set

The study is based on the GLEAN-R reconciled consensus set of predicted gene models in the 12 *Drosophila* genomes (*Drosophila* 12 Genomes Consortium 2007). We included the longest annotated transcript of each gene and only included genes with annotation of a single-copy ortholog in all 12 species or in all *melanogaster* group species (depending on the analysis). We downloaded the nucleotide sequences from ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments, including 6698 12-species ortholog alignments and 8563 *melanogaster* group species alignments; the downloads contained T-Coffee sequence alignments, which we converted to unaligned, ungapped sequences.

### Alignment software

The alignment programs ("aligners") we compared are AMAP (Schwartz and Pachter 2007), MUSCLE (Edgar 2004), ProbCons (Do et al. 2005), ClustalW (Thompson et al. 1994), T-Coffee (Notredame et al. 2000), and PRANK (Loytynoja and Goldman 2005, 2008). We chose these aligners for the following reasons: T-Coffee because it was used for the alignments released with GLEAN-R set, and because these alignments are likely to be used commonly; ClustalW because it is one of the most widely used aligners; PRANK because it has been recently reported to be superior when used in conjunction with site-evolutionary models for selection (Fletcher and Yang 2010); and the remaining three aligners were picked among other popular aligners. These aligners represent a variety of algorithmic approaches. For computational reasons, we did not include more aligners.

We reused the T-Coffee alignments made available with the GLEAN-R set (*Drosophila* 12 Genomes Consortium 2007) and downloaded both the masked and unmasked alignments from ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments. Except where specified differently, we report results for the guided, unmasked T-Coffee alignment. "Masked and trimmed" *melanogaster* group T-Coffee alignments were graciously provided by Timothy Sackton. The second filtering stage involved additional trimming of the masked alignments, such that codons missing or masked in more than one species were removed (T Sackton, pers. comm.).

PRANK alignments were made with the –codon option and default other parameters. The alignments from the remaining four aligners were generated with default parameters via threading through the amino acid sequence; we utilized Bioperl (Stajich et al. 2002) in the process.

### Evolutionary analysis

Evolutionary analysis was performed with the commonly used PAML package (Yang 1997) (CODEML program). Rate of protein evolution was estimated with model M0. Positively selected sites were identified based on two pairs of models: M1a (NearlyNeutral) and M2a (PositiveSelection), and M7 (Beta) and M8 (Beta and ω) (Nielsen and Yang 1998). These model pairs were chosen because data analyses and computer simulations suggest that they are most effective among PAML models (Anisimova et al. 2001, 2002; Wong et al. 2004).

Models M1a, M2a, M7, and M8 are site models and allow the ratio of nonsynonymous to synonymous substitutions ω to vary across sites; ω of each site in the gene is considered a random variable drawn from a statistical distribution. The (nearly) neutral evolution models, M1a and M7, assume a ω drawn from a beta distribution (model M7) or from one of two classes: $\omega < 1$ or $\omega = 1$ (model M1a). The positive selection models, M8 and M2a, add a class $\omega > 1$ to the corresponding neutral model. Whether the positive selection model is more likely than the neutral is determined by computing the LRT statistic (twice the log likelihood difference between the compared models) and comparing it against the $\chi^2$ distribution $\chi^2$. In cases where the positive selection model is statistically significantly more likely, we used the posterior probabilities that the site belongs to the $\omega > 1$ class, as estimated by CODEML with the Bayes empirical Bayes method (Yang et al. 2005), to determine the genes with at least one site under selection. Additional information about the PAML analysis is included in the Supplemental Material.

We considered two cutoff levels for the probability that a site inferred to be positively selected is truly under positive selection, 95% and 99%. All reported results are with a cutoff of 95% unless stated otherwise.

Q-values and False Discovery Rate (FDR) were calculated by using the q-value package (Storey 2002) in R. *P*-values were calculated based on the LRT compared against the $\chi^2$ distribution $\chi^2$. The distribution of *P*-values was U-shaped and, therefore, we used the Bootstrap method for estimating $\pi 0$ (Dabney and Storey 2004); we used default values for the remaining q-value parameters. The critical values for the Bonferroni correction were calculated online with Uitenbroek(1997). The LRT- and FDR-related values for the genes inferred to be under selection in *Drosophila* 12 Genomes Consortium (2007) were downloaded from ftp://ftp.flybase.net/12_species_analysis.

The analysis where regions with a gap in one or more of the species sequences is removed was done by setting CODEML's *cleandata* variable to 1.

### Visual inspection

We visually inspected the alignments by using the Jalview software package (Clamp et al. 2004). We chose the genes to inspect via the following procedure. In the 12 species alignments, we initially picked five genes at random (by using a random number generating Perl script) from each of five "mismatch categories", where the mismatch categories were defined by categorizing genes according to the number of aligners (one to five, excluding PRANK) that lead to a positively selected site being inferred in the gene. This resulted in a total of 25 genes for the initial inspection in Supplemental Table 2S, A. An additional set of 58 genes were selected such that for each aligner five to seven genes were picked at random among each of: (1) all of the genes with an inferred selected site at this aligner's alignment (Supplemental Table 2S, B) and (2) the genes that have an inferred selection site only when aligned with this aligner (Supplemental Table 2S, C). Among these, one gene among those marked as "misaligned," was selected at random from each group, each aligner, to be further investigated for transcript and annotation problems (Supplemental Table 3S).

In the above randomization process, we did not include the PRANK alignments that we considered separately. Supplemental Table 2S, D, contains the information for these genes inferred to be under positive selection after PRANK alignments, and the selection again was done at random among the genes with an inferred selected site in the PRANK alignments. Among alignments within the *melanogaster* group only, we picked 25 genes at random from the genes inferred to be evolving under positive selection at a 1% FDR in *Drosophila* 12 Genomes Consortium (2007). The data in Supplemental Table 2S, E, is based on the GLEAN-R masked alignments of these 25 genes, as at the time of the initial analysis we were not aware that the actual alignments used for PAML analysis in *Drosophila* 12 Genomes Consortium (2007) included trimming of the alignment (codons missing or masked in more than one species were removed, because this had not been mentioned in the publication; T Sackton, pers. comm.). Repeating the analysis on these trimmed alignments did not change the assigned category for any of these 25 genes.

### Other analyses

Regions or sites with intrinsic disorder were annotated with IUpred (Dosztanyi et al. 2005). IUpred predicts disorder from amino acid sequences by estimating their total pairwise inter-residue interaction energy, based on the assumption that IUP sequences do not fold due to their inability to form sufficient stabilizing inter-residue interactions. The energy values are then transformed into a probabilistic score ranging from 0 (complete order) to 1 (complete disorder). The threshold value characteristic of disorder is 0.5 (Dosztanyi et al. 2005), and we used it for the cutoff in our analysis. Unless otherwise stated explicitly, IUpred was run on the *D. melanogaster* sequence only, and we checked for both short and long disorder (score higher than 0.5 for either of them classified the site as disordered).

The consistency scores of the alignments per gene, Consistency of the Overall Residue Evaluation (CORE) (Notredame and Abergel 2003; Kemena and Notredame 2009), were produced with the T-Coffee -evaluate_mode, based on the five alignments previously generated with AMAP, Clustal, MUSCLE, ProbCons, and T-Coffee. They were obtained based on averaging the scores of each of the aligned pairs involving a residue within a column.

## Acknowledgments

We gratefully acknowledge members of the Petrov lab, Rajat Raina, Nadia Singh, Chuong Do, Tim Sackton, Amanda Larracuente, Adam Siepel, and Andy Clark for fruitful discussions and suggestions regarding this project. We thank three anonymous reviewers for their useful suggestions. Support for this work was provided by NIH grants T32 GM63495 and T5 LM07033 to P.M-R., NIH grants GM077368, GM089926, and HG002568-07 to D.P., and NSF grant CNS-0619926 for the Stanford BioX2 Cluster.

## References

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20:** 578–580.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18:** 1585–1592.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* **19:** 950–958.

Anisimova M, Bielawski J, Dunn K, Yang Z. 2007. Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol Biol* **7:** 154. doi: 10.1186/1471-2148-7-154.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci* **104:** 7489–7494.

Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Natl Acad Sci* **103:** 5977–5982.

Chen C-H, Chuang T-J, Liao B-Y, Chen F-C. 2010. Scanning for the signatures of positive selection for human-specific insertions and deletions. *Genome Biol Evol* **2009:** 415–419.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. *Bioinformatics* **20:** 426–427.

Clark NL, Aquadro CF. 2009. A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Mol Biol Evol* **27:** 1152–1161.

Dabney A, Storey J, 2004. QVALUE: The Manual. Version 1.0. http://www.genomics.princeton.edu/storeylab/qvalue/manual.pdf.

Dickson RJ, Wahl LM, Fernandes AD, Gloor GB. 2010. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *PLoS ONE* **5:** e11082. doi: 10.1371/journal.pone.0011082.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **15:** 330–340.

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347:** 827–839.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Ellegren H. 2008. Comparative genomics and the study of evolution by natural selection. *Mol Ecol* **17:** 4586–4596.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. *Mol Biol Evol* **27:** 2257–2267.

Golubchik T, Wise MJ, Easteal S, Jermiin LS. 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol* **24:** 2433–2442.

Greenberg AJ, Stockwell SR, Clark AG. 2008. Evolutionary constraint and adaptation in the metabolic network of *Drosophila*. *Mol Biol Evol* **25:** 2537–2546.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. *Genome Res* **17:** 1837–1849.

Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol* **24:** 2598–2609.

Kawahara Y, Imanishi T. 2007. A genome-wide survey of changes in protein evolutionary rates across four closely related species of *Saccharomyces* sensu stricto group. *BMC Evol Biol* **7:** 9. doi: 10.1186/1471-2148-7-9.

Kemena C, Notredame C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25:** 2455–2465.

Koonin EV. 2009. Darwinian evolution in the light of genomics. *Nucleic Acids Res* **37:** 1011–1034.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4:** e1000144. doi: 10.1371/journal.pgen.1000144.

Kunstner A, Wolf JBW, Backstrom N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, et al. 2010. Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Mol Ecol* **19:** 266–276.

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.

Lassmann T, Sonnhammer EL. 2002. Quality assessment of multiple alignment programs. *FEBS Lett* **529:** 126–130.

Lefebure T, Stanhope MJ. 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. *Genome Res* **19:** 1224–1232.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci* **102:** 10557–10562.

Loytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320:** 1632–1635.

Lunter G, Miklos I, Drummond A, Jensen JL, Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6:** 83. doi: 10.1186/1471-2105-6-83.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res* **19:** 922–933.

Morrison DA. 2009. Why would phylogeneticists ignore computerized sequence alignment? *Syst Biol* **58:** 150–158.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3:** e170. doi: 10.1371/jiournal.pbio.0030170.

Notredame C, Abergel C. 2003. Using multiple alignment methods to assess the quality of genomic data analysis. In *Bioinformatics and genomes: Current perspectives* (ed. MA Andrade), pp. 30–50. Horizon Scientific, Wymondham, UK.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205–217.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci* **106:** 6700–6705.

Nuin PA, Wang Z, Tillier ER. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* **7:** 471. doi: 10.1186/1471-2105-7-471.

Podlaha O, Webb DM, Tucker PK, Zhang J. 2005. Positive selection for indel substitutions in the rodent sperm protein Catsper1. *Mol Biol Evol* **22:** 1845–1852.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet* **2:** e173. doi: 10.1371/journal.pgen.0020173.

Ridout KE, Dixon CJ, Filatov DA. 2010. Positive selection differs between protein secondary structure elements in *Drosophila*. *Genome Biol Evol* **2010:** 166–179.

Savard J, Tautz D, Lercher MJ. 2006. Genome-wide acceleration of protein evolution in flies (Diptera). *BMC Evol Biol* **6:** 7. doi: 10.1186/1471-2148-6-7.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2010. Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* **2009:** 114–118.

Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. *J Mol Evol* **62:** 793–802.

Schwartz AS, Pachter L. 2007. Multiple alignment by sequence annealing. *Bioinformatics* **23:** e24–e29.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative genomics on the *Drosophila* phylogenetic tree. *Annu Rev Ecol Evol Syst* **40:** 459–480.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12:** 1611–1618.

Storey J. 2002. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* **64:** 479–498.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* **18:** 1393–1402.

Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22:** 2047–2048.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* **27:** 2682–2690.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455:** 105–108.

Uitenbroek DG. 1997. Simple interactive statistical analysis bonferroni calculator. http://www.quantitativeskills.com/sisa/calculations/bonfer.htm.

Vieira FG, Sanchez-Gracia A, Rozas J. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome Biol* **8:** R235. doi: 10.1186/gb-2007-8-11-r235.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* **319:** 473–476.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555–556.

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

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22:** 1107–1118.

Yang H, Wu Y, Feng J, Yang S, Tian D. 2009. Evolutionary pattern of protein architecture in mammal and fruit fly genomes. *Genomics* **93:** 90–97.

Zhu L, Wang Q, Tang P, Araki H, Tian D. 2009. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol Biol Evol* **26:** 2353–2361.

*Received October 1, 2010; accepted in revised form February 18, 2011.*