Multinucleotide mutations cause false inferences of positive selection

Aarti Venkat[1], Matthew W. Hahn[2], Joseph W. Thornton*[1,3]

(1) Department of Human Genetics, University of Chicago, Chicago IL 60637, USA

(2) Department of Biology and Department of Computer Science, Indiana University, Bloomington IN 47405, USA

(3) Department of Ecology & Evolution, University of Chicago, Chicago IL 60637, USA

*Correspondence: Joseph Thornton, joet1@uchicago.edu

Keywords: adaptation, adaptive evolution, branch-site test, multinucleotide mutations, codon models

ABSTRACT

Phylogenetic tests of adaptive evolution, which infer positive selection from an excess of nonsynonymous changes, assume that nucleotide substitutions occur singly and independently. But recent research has shown that multiple errors at adjacent sites often occur in single events during DNA replication. These multinucleotide mutations (MNMs) are overwhelmingly likely to be nonsynonymous. We therefore evaluated whether phylogenetic tests of adaptive evolution, such as the widely used branch-site test, might misinterpret sequence patterns produced by MNMs as false support for positive selection. We explored two genome-wide datasets comprising thousands of coding alignments – one from mammals and one from flies – and found that codons with multiple differences (CMDs) account for virtually all the support for positive selection inferred by the branch-site test. Simulations under genome-wide, empirically derived conditions without positive selection show that realistic rates of MNMs cause a strong and systematic bias in the branch-site and related tests; the bias is sufficient to produce false positive inferences approximately as often as the branch-site test infers positive selection from the empirical data. Our findings suggest that widely used methods for detecting adaptive evolution often infer a gene to be under positive selection simply because it stochastically accumulated one or a few MNMs. Many, or even most, published inferences of adaptive evolution using these techniques may therefore be artifacts of model violation caused by unincorporated neutral mutational processes. We develop an alternative model that incorporates MNMs and partially reduces this bias, but at the cost of reduced power.

INTRODUCTION

Identifying genes that evolved under the influence of positive natural selection is a central goal in molecular evolutionary biology. During recent decades, likelihood-based phylogenetic methods have been developed to identify gene sequences that retain putative signatures of past positive selection [1-10]. Perhaps the most widely used of these is the branch-site (BS) test of episodic selection, which allows positive selection to affect only some codons on one or a few branches of a phylogeny, and therefore has relatively high power compared to methods that detect selection across an entire sequence or an entire phylogenetic tree [5,6,11]. The BS test has been the basis for published claims of lineage-specific adaptive evolution in many thousands of individual genes [12-16].

The BS and related methods use a likelihood ratio test to compare how well two mixture models of sequence evolution on a phylogeny fit an alignment of coding sequence data. The null model constrains all codons to evolve with rates of nonsynonymous substitution (d_N) less than or equal to the rate of synonymous substitution (d_S), as expected under purifying selection and drift. In the positive selection model, some sites are allowed to have $d_N > d_S$ on a branch or branches of interest. If the increase in likelihood of this model given the data is greater than expected due to chance alone, the null model is rejected and adaptive evolution is inferred. The BS test has been shown to be conservative, with a low rate of false positive inferences, when sequences are generated under an evolutionary process corresponding to the null model [6,11]. It is widely appreciated that likelihood ratio tests can become biased if the underlying probabilistic model is incorrect [17]. The effect on the BS test of a few forms of model violation—such as unequal distribution of selective effects among sites, high sequence divergence, and non-allelic gene conversion—have been previously studied [18-21], and the test has been found to be reasonably robust to most model violations considered [6,22,23].

Recent research in molecular genetics and genomics suggests a potentially important phenomenon that has not been incorporated into models used in tests of positive selection: the propensity of DNA polymerases to produce mutations at neighboring sites. All implementations of the BS and other likelihood-based tests of adaptive evolution use models in which mutations occur only at individual nucleotide sites and are fixed singly and independently. Codons with multiple differences between them can be interconverted only by serial single-nucleotide substitutions, the probability of which is the product of the probabilities of each independent

event. Recent molecular studies have shown, however, that mutations affecting adjacent nucleotide sites often occur during replication, apparently because certain DNA microstructures recruit error-prone polymerases that lack proofreading activity and therefore make multiple errors close together [24-33]; cytosine deaminases can also introduce clusters of co-occurring mutations [34,35]. Consistent with these mechanisms, genetic studies of human trios and mutation-accumulation experiments in laboratory organisms indicate that *de novo* clustered mutations occur more frequently than expected if each occurred independently [24,30-32,36,37], and these MNMs are enriched in transversions. [32,38] Within and between species, one to ten percent of sequence differences are clustered, far more than expected if mutations are independent, and contain transversions at an elevated frequency[24,32,38,39].

We hypothesized that these mutational processes might lead to false signatures of positive selection. Because of the structure of the genetic code, virtually all MNMs in coding sequences are nonsynonymous, and most would comprise multiple nonsynonymous nucleotide changes if they were to occur by single nucleotide steps (**Supplementary Table 1**). The enrichment of transversions in MNMs further increases the propensity for MNMs to produce nonsynonymous changes, because transversions are more likely than transitions to be nonsynonymous. MNMs are therefore likely to produce codons with multiple differences (CMDs) that contain an apparent excess of nonsynonymous substitutions. When these CMDs are assessed using a method that treats all substitutions as independent events, a model that allows d_N to exceed d_S at some sites may have a higher likelihood than one that restricts d_N/d_S to values ≤ 1. Further, the assumption that all mutations have the same transversion-transition rate might exacerbate the tendency to misinterpret MNM-produced nonsynonymous changes as evidence for positive selection. Although CMDs might also be driven to fixation by recurrent positive selection [40-42], these considerations suggest that the possibility that failing to incorporate MNMs in likelihood models might make tests of adaptive evolution susceptible to false positive inferences. The BS and related tests might be particularly sensitive to this problem because they seek signatures of positive selection acting on individual codons on individual branches of the tree [11,43].

RESULTS

To understand the effect of MNMs on the accuracy of the branch-site and related tests of adaptive evolution, we analyzed in detail two previously published genome-wide datasets, which represent classic examples of the application of these tests [12,16,44]. The mammalian dataset consists of coding sequences of 16,541 genes from six eutherian mammals; we retained for analysis only the 6,868 genes with complete species coverage. The fly dataset consists of 8,564 genes from six species in the melanogaster subgroup clade, all of which had complete coverage (**Supplementary Fig. 1**). The fly genes have higher divergence than those in the mammalian dataset, allowing us to examine the performance of the BS test under different evolutionary conditions.

We used the classic BS test to identify genes putatively under positive selection ($P<0.05$) on the human lineage in the mammalian dataset and on all six terminal lineages in flies. 82 genes in humans and 3,938 in flies yielded significant tests (**Supplementary Table 2**). To facilitate further analysis of CMDs, we imposed a quality control filter that kept only those genes in which all CMDs on the branch of interest were reconstructed identically between null and positive selection models; we also applied a multiple testing correction (FDR <0.20). In flies, 443 genes were retained after these steps. Thirty human genes passed the reconstruction filter, but none met the FDR threshold, consistent with previous analyses of these data [16]; nevertheless, we included the 30 initially significant human genes for further analysis because this lineage is the object of intense interest and because its short length contrasts with the fly branches. These two groups constitute the "BS-significant" sets of genes in flies and humans.

CMDs provide virtually all support for positive selection

We sought to determine how much of the evidence for positive selection comes from CMDs. We first observed that CMDs were dramatically enriched in BS-significant genes compared to non-BS-significant genes (**Fig. 1a**). In humans, 97% of BS-significant genes contain CMDs, but only 0.5% of BS-nonsignificant genes do (**Supplementary Fig. 2**). The pattern is similar but less extreme in flies, with the average number of CMDs in BS-significant genes twice that in non-significant genes (**Supplementary Fig. 2**). When CMD-containing codons are excluded from the alignments, the vast majority of genes that were BS-significant lose their signature of selection in both datasets (**Fig. 1b**).

We next calculated the fraction of statistical support for positive selection that comes from CMDs. The total support for positive selection in an alignment is defined as the difference between the log-likelihood of the positive selection model and that of the null model, summed across all codons in the alignment. The fraction of support from CMDs is the support from CMD-containing codons divided by the total support across the entire alignment. CMDs account for >95% of the support for positive selection in virtually all BS-significant genes in both datasets; in about 70% of genes, CMDs provide all the support (**Fig. 1c**).

Finally, we examined the BS test's *a posteriori* identification of sites under positive selection. We found that CMDs were far more likely to be classified as positively selected than non-CMDs. Among genes that were BS-significant on the human lineage, every CMD was inferred to be under positive selection using a Bayes Empirical Bayes posterior probability (PP) cutoff > 0.5. Using a more stringent cutoff of PP>0.9, 66 percent of CMDs were classified as positively selected, compared to 0.07% of non-CMDs. In the fly dataset, CMDs accounted for 90% of codons with BEB>0.9, despite accounting for less than 1% of all codons (**Fig. 1d**).

CMDs are therefore the primary drivers of the signature of selection identified in the BS test. A single CMD provides sufficient statistical support to yield a signature of positive selection on the human lineage, and only a few CMDs in a gene are enough to do the same in flies.

Incorporating MNMs eliminates the signature of positive selection in many genes

If MNMs cause a bias in the BS test, incorporating them into the evolutionary model should reduce the signal of positive selection. We developed a codon model in which double-nucleotide changes are allowed at a rate defined by the parameter δ, which represents the rate of double-nucleotide substitutions relative to that of single-nucleotide substitutions. Simulations under conditions derived from a sample of genes in the mammalian dataset validated the accuracy of parameters estimated using this model (**Supplementary Fig. 3**).

We applied this model to all alignments in the mammalian and fly datasets to estimate the relative rate of MNMs. The average value of δ was 0.026 in mammals and 0.062 in flies, consistent with previously published work quantifying dinucleotide substitutions in these taxa[45]. The estimated δ in BS-significant genes was about twice that in BS-nonsignificant genes in both datasets (**Fig. 2a**).

To determine whether incorporating MNMs eliminates the signature of positive selection, we implemented a version of the BS-test (BS+MNM), which is identical to the classic version except that both the null and positive selection models allow MNMs. We found that 96% of the BS-significant genes on the human lineage lost significance in the BS+MNM test (**Fig. 2b**). In flies, 38% of the BS-significant genes lost significance; a substantial fraction of those that retained significance were enriched in triple substitutions, a process not accounted for in our model (**Fig. 2b**).

MNMs cause false positive inferences on a genome-wide scale

Our observation that the signature of positive selection in many genes is eliminated by use of the BS+MNM model could be due to two causes: failure to incorporate MNMs in the classic BS test may cause a bias towards false positive inferences, or the BS+MNM method may simply have reduced power to identify authentic positive selection.

We addressed this question in two ways. First, we performed power analyses of the BS+MNM test using simulations in which positive selection is present in the generating model. We simulated sequence data on the mammalian and fly phylogenies using genome-wide averages for all model parameters, except we varied the strength of positive selection (ω_2) and the proportion of sites under positive selection. We applied the BS+MNM test to these data and found that it can reliably detect strong positive selection ($\omega_2 > 20$) when it affects more than 10% of sites in a typical gene, or moderate positive selection ($10 < \omega_2 < 20$) that affects a larger fraction of sites (**Supplementary Fig. 4a**). Compared to the classic BS test, however, the BS+MNM test has slightly reduced power to detect positive selection when selection is weak and affects only a small fraction of sites (**Supplementary Figs. 4a - 4c**).

We therefore used simulations to evaluate how frequently realistic rates of multinucleotide mutation produce false positive inferences in the classic BS test. For every gene in the mammalian and fly datasets, we simulated sequence evolution under the null BS+MNM model without positive selection using parameters, including δ, derived from the alignments. We then analyzed these alignments using the classic BS test. In both datasets, the number of genes with significant results ($P<0.05$)—all of which are false positive inferences—was even greater than the number of genes the BS test concluded were under positive selection using the empirical data (**Fig. 3a).** These false inferences are caused primarily by MNM-induced bias,

because simulating data under identical control conditions without MNMs ($\delta = 0$) produced few positive tests. In flies, the proportion of false positive tests when MNMs are present far exceeds 0.05, despite the conservative approach the method uses to calculate P-values [6,11]. More than 1,700 of these false positive tests survive FDR adjustment, compared to just four in the control simulations (**Supplementary Table 2**). In humans, the fraction of false positive inferences is lower, consistent with the test's reduced power in this dataset, but still about three times greater than in the control simulations.

Taken together, these findings indicate that MNMs under realistic evolutionary conditions produce a strong and widespread bias in the BS test towards false inferences of positive selection. This bias is strong enough to potentially account for all genome-wide inferences of positive selection made by the BS test in both humans and flies.

Systematic bias caused by chance MNMs in longer genes

We next sought to identify the causal factors that determine whether a gene yields a false positive result in the BS test because of MNM-induced bias. Most genes are only several hundred codons long, and only a few percent of mutations are MNMs, so on phylogenetic branches of short to moderate length many genes will contain no CMDs caused by multinucleotide mutations. We therefore hypothesized that the propensity for a gene to produce a BS-significant result will depend on factors that increase the probability it will contain one or more fixed MNMs by chance, including its length and the gene-specific rate at which MNMs occur within it.

We first tested for an effect of gene length on the results of the branch-site test. As predicted, we observed that BS-significant genes were on average 100 and 16 codons longer than non-significant genes in the human and fly empirical datasets, respectively (**Fig. 3b**). The genes that produce false positive BS tests in the genome-wide null simulations are also longer than the non-significant genes: the BS-significant genes are on average 26 and 31 codons longer than non-significant genes in the human and fly simulated datasets, respectively (**Supplementary Fig. 5**). These results suggest that genes that present a larger "target" for multinucleotide mutation are more likely to yield positive BS results by chance in both the empirical data and those simulated under the nullions.

To directly test the causal relationship between sequence length and bias in the BS test,

we simulated sequence evolution at increasing sequence lengths, using evolutionary parameters derived from each of the BS-significant genes in the mammalian and fly datasets. For each gene's parameters, we simulated 50 replicate alignments under the BS+MNM null model and then analyzed them using the classic BS test (**Supplementary Fig. 6a**). The false positive rate for any gene's simulations is defined as the fraction of replicates with a significant LRT in the classic BS test, using a P-value cutoff of 0.05. When sequences 5,000 codons long were simulated, 96% of BS-significant genes in the mammalian dataset yielded an FPR greater than 0.05, with a median FPR across genes of 0.39; simulating sequences 10,000 codons long increased this fraction to 100% and the median FPR to 0.56 (**Fig. 3c**). In flies, 99% of genes had FPR>0.05 (median FPR 0.74) when genes 5,000 codons long were analyzed, which increased to 100% of genes (median FPR 0.90) at sequence length of 10,000 codons (Fig. 3c). Control simulations under identical conditions but with δ=0 led to very low FPRs (median 0.02 to 0.03 for both datasets), even with very long sequences (grey dots in **Fig. 3c**). A similar systematic bias and elevated false positive rate also resulted when sequences were simulated under gene-specific conditions, but with δ fixed to its average across the thousands of BS-nonsignificant genes in each dataset (**Supplementary Fig. 6b**).

We next evaluated whether the gene-specific rate of multinucleotide mutation affects a gene's propensity to yield a positive result in the BS test. As predicted, we observed that BS-significant genes in the empirical datasets had higher estimated δ than nonsignificant genes (**Fig. 2a**); similarly, genes producing false positive results in the genome-wide null simulations under empirical conditions also tended to be those with higher δ (**Fig 3d**). To test the effect of the neutral MNM substitution rate on the BS test, we simulated sequences 5,000 codons long under the null BS+MNM model, with a variable δ and all other parameters fixed to their averages across all genes. We found that increasing δ led to a monotonic increase in the frequency of false positive inferences. The FPR was >0.05 percent when δ was only 0.001 and 0.013 on the human and fly lineages, respectively. When δ was equal to its genome-wide average (0.026 and 0.062 in mammals and flies), false positive inferences occurred at rates of 22 and 17 percent, respectively (**Fig. 3e**).

Typical evolutionary conditions are therefore sufficient to cause a strong and systemic bias in the BS test. MNMs are rare, however, so longer genes and those with higher rates of multinucleotide mutation are more likely to undergo this process and manifest the bias. This

view is further supported by the fact that fewer genes are BS-positive on the human branch – which is so short that substitutions of any type are rare, and MNMs even more so – than on the fly phylogeny, where branches are longer, more CMDs are apparent, and hundreds of genes have BS signatures of selection. Taken together, these findings suggest that many genes with BS-significant results in empirical datasets may simply be those that happened to fix multinucleotide substitutions by chance alone.

Transversion-enrichment in CMDs exacerbates bias in the branch-sites test

MNMs tend to produce more transversions than classical single-site mutational processes, so if CMDs are produced by MNMs, they should be transversion-rich. As predicted, the transversion:transition ratio is elevated in CMDs relative to that in non-CMDs by factors of three and two in mammals and flies, respectively, suggesting that an MNM-like process is likely to have produced many of the CMDs (**Fig. 4a**). In the subset of BS-significant genes, CMDs have an even more elevated transversion:transition ratio, as expected if transversion-rich MNMs bias the test (**Fig. 4a**).

To test whether transversion enrichment in MNMs might exacerbate the BS test's bias, we developed an elaboration of the BS+MNM model in which an additional parameter allows MNMs to have a different transversion:transition ratio (κ_2) than single-site substitutions do (κ_1). We validated this implementation by simulating and then analyzing sequence data using the BS+MNM+κ_2 model (**Supplementary Fig. 7**). We estimated the maximum likelihood estimates of the model's parameters for every gene in the mammalian and fly datasets. As predicted, the average genome-wide value of κ_2 is notably higher than κ_1 in both sets of genomes (mean κ_2/κ_1 = 1.2 and 2.67 in the mammalian and fly datasets, respectively), indicating an enrichment of transversions in CMDs.

To determine if this excess of transversions exacerbates the bias of the BS test, we simulated sequence evolution under the BS+MNM+κ_2 null model, using the genome-wide average values for all model parameters and branch lengths, except for κ_2, which we varied. Sequences 10,000 codons long were used, because simulating shorter sequences resulted in a high variance in the realized transversion:transition ratio. We analyzed these data using the classic BS test and calculated the fraction of replicates in which positive selection was inferred. We found that increasing κ_2 caused a monotonic increase in the false positive rate, indicating that

transversion enrichment in MNMs does exacerbate the test's bias (**Fig. 4b**). The effects are strong: when κ_2 was assigned to its genome-wide average value, the false positive rate was 48% and 76% on sequences simulated under conditions derived from the mammalian and fly datasets, respectively. These data indicate that realistic values of the frequency at which MNMs occur and the enrichment of transversions within MNMs can cause a strong bias in the BS test.

MNMs affect newer tests of positive selection

In recent years, newer likelihood-based methods have been introduced to test for episodic site-specific positive selection [2,3,7]. All these methods are based on models of sequence evolution that, like the BS test, do not allow MNMs but instead model CMDs as the result of serial site-specific substitutions. We therefore hypothesized that these methods might also be biased by MNMs. We chose two recent branch-site tests, BUSTED and MEME [2,3], and tested their performance on alignments 5,000 codons long that were simulated using the BS+MNM null model and parameters estimated from the BS-significant gene alignments in humans and flies. To test for MNM-induced bias, we compared results when δ was assigned to three different values: zero, its average across all alignments in the mammalian or fly datasets, or its gene-specific value in each of the BS-significant genes (**Supplementary Fig. 6a**).

We found that BUSTED was very sensitive to an MNM-induced bias. When δ=0, virtually no genes' parameters led to frequent false positive inferences, with a median FPR <0.03 across genes **(Fig. 5a)**. But when δ was assigned to its empirically estimated gene-specific value, the parameters from every gene in humans and the majority in flies yielded false positive rates >0.05, with median FPRs of 0.29 and 0.5, respectively (**Fig. 5a**). Frequent false positive inferences were evident when sequences were simulated using genome-wide average estimates of δ, as well.

We also tested MEME, a fixed-effect likelihood method that tests for an elevated ω at every site rather than testing for positive selection across an entire gene. Genes were inferred as subject to positive selection if one or more codons had a significant signal of adaptive evolution after adjustment for multiple testing. We found that the proportion of genes falsely inferred to under positive selection was quite low compared to BUSTED (Figs. 5a, 5b), but MEME also had extremely low power, consistent with the conclusions of Murrell *et al.* [2] (**Supplementary Figs. 4d, 4e**). MEME does appear to be affected by MNM-induced bias, however, because when data

were simulated under the null model using empirically derived values of δ the proportion of false inferences by MEME increased as the stringency of the FDR cutoff was reduced. In contrast, control simulations in which δ was fixed at zero produced no false positive inferences, even when we raised the FDR cutoff to 0.30 (**Fig. 5b**). These data suggest that MNMs can be misinterpreted by MEME as evidence of positive selection, but the method's extreme conservativism keeps the rate of positive inferences—whether true or false—low.

CMDs that invoke multiple non-synonymous steps drive the signature of positive selection

Finally, we sought further insight into the reasons why CMDs yield a false signature of positive selection in the BS and related tests. In standard models of codon evolution, CMDs are interpreted as the result of two or more serial independent substitutions, even though they can be produced by MNMs in a single mutational event. We hypothesized that CMDs that imply multiple nonsynonymous nucleotide substitutions under these models would provide the strongest support for the positive selection model. We therefore classified CMDs in the empirical datasets by the minimum number of nonsynonymous single-nucleotide substitutions required from the ancestral to derived codon state under standard codon models. As predicted, we found that CMDs that imply more than one nonsynonymous step are dramatically enriched in BS-significant genes (**Fig 6a**).

We also examined the statistical support provided by different kinds of CMDs. As the number of nonsynonymous steps increased, the statistical support provided for the positive selection model also increased (**Fig. 6b**). CMDs that imply one nonsynonymous and one synonymous step typically provide weak to moderate support for the positive selection model, but CMDs that imply two nonsynonymous steps provide very strong support. In many cases, a single CMD in this latter category is sufficient to yield a statistically significant signature of positive selection.

DISCUSSION

Our results demonstrate that the branch-site test and newer related tests suffer from a strong and systematic bias towards false positive inferences. This bias is caused by a mismatch between the methods' underlying codon model of evolution – which assumes that a codon with multiple differences can be produced only by two or more independent substitution events – and

the recently discovered phenomenon of multinucleotide mutation, which produces such codons in a single event. Because of the structure of the genetic code and the high transversion rates that characterize MNMs, most codons produced by this mechanism cause more than one nonsynonymous single-nucleotide change. Confronted with this kind of codon data, the likelihood calculated by the BS test is determined by the product of the probabilities of the individual mutations. Under the null model, the probability of such compound events is extremely small, but it can increase dramatically when d_N/d_S exceeds one, as the positive selection model allows. This increase in likelihood afforded by the positive selection model is much greater than it would be if the substitution were interpreted as the result of a single multinucleotide event. Indeed, our results show that a single codon comprising two nonsynonymous substitutions is often sufficient to yield a statistically significant signature of positive selection in the BS test for an entire gene.

As a result, the BS test functions primarily as a CMD detector. Virtually all statistical support for positive selection comes from CMD-containing sites; removing them from the alignment or incorporating MNMs into the BS test's model eliminates the signature of selection from the majority of genes. The BS test therefore reports positive selection almost exclusively based on sequence patterns that can be produced by either positive selection or neutral evolution under multinucleotide mutation. The test's inability to distinguish between positive selection and neutral fixation of MNMs produces a strong bias towards false inferences of selection when MNMs occur.

The bias is strong and relevant under realistic conditions. Indeed, when sequences were simulated under the null model using parameters estimated from genes in the fly and mammalian datasets, the number of genes with false positive BS tests was approximately the same as the number of genes with positive BS results when the empirical data were analyzed. There is therefore no excess of BS-positive results in these genomes beyond that potentially attributable to MNM-induced bias. Worse, these null simulations did not include the elevated transversion rate that characterizes MNMs, which exacerbates the test's bias. Further, longer genes were more likely to produce false positive results, and the genes with BS-significant results in the empirical data tended to be longer than those without a signature of selection; it is not obvious why, if the BS test's positive results in these genomes were authentic, longer genes would be more likely to be affected by positive selection. Taken together, these results suggest that MNM-

induced bias may explain many or even most of the BS test's inferences of positive selection in these datasets.

Are our findings from these datasets generalizable? MNMs appear to be a property of all eukaryotic replication processes, and the MNM rates that we observed in mammals and flies are in the same range as those previously identified in genetic and molecular studies in a variety of animal, plant, and fungal species. Both datasets comprise a small number of taxa, but the BS test seeks evidence of selection on individual branches, so it seems unlikely that larger trees will somehow inoculate the test against MNM-induced bias. We observed strong bias on lineages with divergence levels ranging from very low (on the human terminal branch) to moderate (the fly branches), so this problem does not appear to be unique to highly diverged sequences or phylogenies with long branches. We must therefore consider the possibility that many – or perhaps even most – of the thousands of previously published reports of positive selection based on the BS-test could be artifacts of MNM-induced bias. Many of the genes with BS-significant results may simply be the ones that happened by chance to neutrally fix one or more multinucleotide mutations.

We do not contend that adaptive evolution is unimportant, or that the BS test is always wrong. Indeed, some of the CMDs in BS-significant genes may have evolved because of authentic positive selection, either by repeated substitution of single nucleotides in a codon or selection on MNMs. But because the BS test functions as a CMD detector, and CMDs can be produced by either positive selection or neutral evolution of MNMs, it provides no reliable evidence of an adaptive history – not even suggestive or *prima facie* evidence. There are numerous cases of strongly supported adaptive evolution involving host-parasite and intracellular genetic conflicts that have produced sequence signatures of positive selection that are likely to be authentic [46]. The persuasive evidence in these cases, however, comes from sources other than the branch-site tests.

If the BS and other tests based on the single-step codon model are unreliable in the face of multinucleotide mutation, what should researchers do? The BS+MNM test could be used to accommodate multinucleotide mutation in the model of codon evolution. However, there are numerous other forms of evolutionary complexity that are not incorporated in our extended BS+MNM model, including MNMs that affect three consecutive nucleotides in a codon, elevated transversion probability within MNMs, and many other kinds of heterogeneity that might bias

the BS+MNM test [47-49]. Considerable additional work is therefore required before our model or improvements upon it can be used with confidence.

A complementary approach is to use functional experiments to explicitly test hypotheses that specific historical changes in molecular sequence caused changes in function or phenotype thought to have mediated adaptation [50,51]. Indeed, the bias we observed may help to explain why some molecular experiments have shown that codons with a high posterior probability of positive selection in the BS test do not contribute to putative adaptive functions, whereas the codon changes that do confer those functions have low or moderate PPs [52]. Experimental tests provide the most convincing evidence of a gene's putative adaptive history, but they require time-consuming laboratory and fieldwork [53,54], so it is not clear how to implement them on a genome-side scale. Future research may develop and validate more robust models to detect positive selection, and these may help to identify candidate genes for which specific, testable hypotheses of past molecular adaptation on specific lineages can be formulated. The phylogenetic tests of selection used to date, however, are misleading for even this purpose.

ACKNOWLEDGEMENTS

We thank the Beagle2, Midway2, and Tarbell supercomputing clusters at the University of Chicago. We are grateful to the members of the Thornton lab for discussion and helpful comments. Funding was provided by NIH R01GM104397 and R01GM121931 (JWT), NSF DEB-1601781 (JWT and AV), NSF DBI-1564611 (MWH), and the Precision Health Initiative of Indiana University (MWH).

AUTHOR CONTRIBUTIONS

Analyses were designed by all authors, performed by AV, and interpreted by all authors. The manuscript was written by AV and JWT with contributions from MWH.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

METHODS

Datasets, quality control, and inference of BS-significant genes: We analyzed two previously published comprehensive datasets of protein-coding alignments on a genomic scale, one in six mammals, the other in six *Drosophila* species **(Supplementary Table 2)**.[16 12,44] We retained only gene alignments with complete coverage in all species and without gross misalignments. We then applied the branch-site test as implemented in CODEML 4.7 to each alignment, assuming the phylogenetic relationships reported in the published studies (**Supplementary Fig. 2**) [12,16]; branch lengths and model parameters were estimated for each alignment by maximum likelihood (ML), and the F3x4 model was used for codon frequencies. We tested each gene in mammals for selection on the terminal branch leading to humans and each gene in flies for selection on each of the six terminal branches[6]. As is standard practice, we calculated P-values using a likelihood ratio test with 1 df (χ_1^2), which makes the test conservative under the null hypothesis. [6] Genes were initially identified as having a putative BS signature of selection at P<0.05. We then applied a correction for multiple testing to a false discovery rate (FDR) <0.20 using the *q-value* package in R (available at http://github.com/jdstorey/qvalue).

To facilitate unambiguous analysis of CMDs, we removed genes containing CMDs for which the ML ancestral reconstructions reported by CODEML at the base of the tested branch differed between the null and positive selection models. In flies, 443 gene-tests ("genes") were retained after these filters and constitute the BS-significant set of genes from this dataset. No genes on the human lineage were significant after FDR correction, so we retained as the BS-significant set from this dataset those genes that passed the ancestral reconstruction filter and had P<0.05 (Supplementary Table 2). The BS-nonsignificant set of genes comprises all genes that pass the alignment and ancestral reconstruction filter that are not in the BS-significant set (*n*=6757, humans; *n*=6883, flies).

Support for positive selection: CMDs were identified in BS-significant and BS-nonsignificant genes as codons with 2 or 3 observed nucleotide differences between the ML states at the ancestral and extant nodes for the branch being tested; non-CMDs are codons with 0 or 1 differences on the branch tested. CMDs were not assessed on branches not tested.

To determine the role of CMDs in significant results from the BS test, we excluded codon positions in BS-significant genes containing CMDs, reanalyzed the data using the BS test, and calculated the fraction of tests that retained a significant result (P<0.05).

We quantified the proportion of statistical support for positive selection in BS-significant genes that comes from CMDs as follows. The site-specific support provided by one codon site in an alignment is the difference between the log-likelihoods of the positive selection model and the null model given the data at that site. Support for positive selection provided by all CMDs in a gene ($support_{CMD}$) is the support summed over all CMD sites in the alignment. The proportion of support provided by CMDs is $support_{CMD} / (support_{CMD} + support_{nonCMD})$. This proportion can be greater than 1 if support by non-CMDs is negative, as occurs if the likelihood of the null model at non-CMD sites is higher than that of the positive selection model, given the parameters of each model estimated by ML over all sites.

Sites were classified *a posteriori* as under positive selection if their Bayes Empirical Bayes posterior probability of being in class 2 (ω_2>1) under the positive selection model in CODEML was >0.5 (moderate support) or >0.9 (strong support).

We categorized observed CMDs by the minimum number of nonsynonymous single-nucleotide steps implied under the Goldman-Yang model between the ancestral and derived states. For each CMD comprising two nucleotide differences, there are two paths by which they can be interconverted by two single nucleotide steps. We determined whether the steps on these paths would be nonsynonymous or synonymous using the standard genetic code and then calculated the mean number of nonsynonymous steps averaged over the two paths. Paths involving stop-codons were not included. We conducted a similar analysis for all possible CMDs in the universal genetic code table.

BS+MNM codon substitution model and test. The codon substitution model of the classic BS test is based on the Goldman-Yang (GY) model [5]. Sequence evolution is modeled as a Markov process, where the matrix element q_{ij}, the instantaneous rate of change from ancestral codon *i* to derived codon *j*, is defined for four types of changes: synonymous transitions and transversions, and non-synonymous transitions and transversions. Three parameters are estimated from the data by maximum-likelihood: ω, the ratio of non-synonymous substitution rate to the synonymous substitution rate (d_N/d_S); π_j , the equilibrium frequency of codon *j;* and κ, the transversion:transition rate ratio. Element q_{ij} is zero for substitutions involving more than one difference, so codons with multiple differences can only evolve through intermediate codons that are a single change away. A scaling factor applied to the matrix ensures that branch lengths are interpreted as the expected number of substitutions per codon.

$$q_{ij} = \begin{cases} \kappa\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \kappa\omega\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ 0 & \text{two or more differences} \end{cases}$$

We developed a modification of the GY model that incorporates MNMs using the parameter, δ, which represents the relative instantaneous rate of double substitutions to that of single substitutions. When δ = 0, the BS+MNM model reduces to the classic BS model that does not incorporate MNMs. Triple substitutions have an instantaneous rate of zero.

The BS+MNM test of positive selection is identical to the BS test, except it utilizes this MNM codon model. We implemented this test by modifying the branch-site test batch file (YangNielsenBranchSite2005.bf) in Hyphy 2.2.6 software by declaring δ a global variable, incorporating it into the codon table, and allowing it to be optimized by ML as it other model parameters are.

$$q_{ij} = \begin{cases} \kappa\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \kappa\omega\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ \omega\delta\kappa^2\pi_j & \text{non-synonymous, both transversions} \\ \omega\delta\pi_j & \text{non-synonymous, both transitions} \\ \omega\delta\kappa\pi_j & \text{non-synonymous, single transversion} \\ \delta\pi_j & \text{synonymous, both transitions} \\ \delta\kappa^2\pi_j & \text{synonymous, both transversions} \\ \delta\kappa\pi_j & \text{synonymous, single transversion} \\ 0 & \text{otherwise} \end{cases}$$

We validated the BS+MNM implementation by simulating 50 replicate alignments using the BS+MNM null model in Hyphy under genome-median parameters (see below). We then used the BS+MNM procedure to find the ML estimate of each parameter, including branch lengths,

given each alignment and the topology of the phylogeny used to generate the sequences. We compared the distribution of estimates over replicates to the "true" values used to generate the sequences (**Supplementary Fig. 3**).

Simulations and analysis of false-positive bias. To characterize bias in the BS and other tests of selection, we conducted sequence simulations in the absence of positive selection under empirically derived conditions. We used the BS+MNM method we implemented in Hyphy to estimate by maximum likelihood (ML) the gene-specific branch lengths and parameters of the null BS+MNM model for every gene in the mammalian and fly datasets. We also calculated the genome-wide median of each parameter over all genes in each dataset (the "genome-average" parameter value). Probability density characterizations for parameters δ and gene length were performed using the *density* function in R.

We simulated sequence evolution under the BS+MNM null model using either gene-specific or genome-median parameters. First, we simulated a "pseudo-genome" without positive selection by simulating one replicate of each of the 6868 and 8564 mammalian and fly alignment, each at its empirical length, using the BS+MNM null model and the ML parameter estimates inferred for that gene from the empirical data. We then ran the BS test on these sequences, testing for signatures of positive selection on the human lineage and each terminal fly lineage **(Supp. Table 2)**. Control simulations were conducted under identical conditions but with δ=0.

To test the effect of gene length on bias in the BS test, we focused on genes in the BS-significant set. For each gene's gene-specific parameters, we simulated 50 replicates alignments of length 5,000 or 10,000 codons. We analyzed these alignments using the BS test, assigning the human branch as foreground for mammalian genes or, for flies, the same branch that produced a significant result when the empirical data were analyzed. The false positive rate (FPR) for any gene's parameters is the fraction of replicates yielding a positive test ($P<0.05$). We also repeated these simulations and analyses using the genome-median value of δ. For control experiments without MNMs, we set δ =0 in the simulations.

To test the effect of the rate at which MNM substitutions are produced on false positive inference rates, we simulated evolution of alignments 5,000 codons long under the BS+MNM null model, using genome-median estimates for all parameters except δ, which we varied. At each value of δ, we simulated 50 replicates. We analyzed each replicate using the BS test for selection on the human or *D. simulans* lineages and calculated the proportion of replicates for each value of δ that yielded a false positive inference ($P<0.05$).

BUSTED and MEME. To examine the accuracy of BUSTED and MEME, we used Hyphy software 2.2.6 (batch files BUSTED.bf and QuickSelectionDetection.bf). We analyzed the 5,000 codon-long alignments simulated under the BS+MNM null model, using parameters estimated by ML for each BS-significant gene, with δ assigned either to its gene-specific estimate, its genome-average, or to zero. We applied BUSTED to the replicate alignments to test for selection ($P<0.05$) on the human lineage or the same fly lineage that was significant for that gene in the BS test of the empirical data.

We analyzed the same alignments using MEME, a site-wise test, for positive selection on the human branch or the same branch that was significant in the BS test of the empirical data. We identified codons in each replicate significant at FDR 0.1, 0.2, or 0.3 using the Benjamini-Hochberg procedure. The proportion of replicates in which at least one replicate had one or more significant sites was then computed.

Power analyses. To characterize the statistical power of the BS and BS+MNM tests, we simulated sequence evolution with positive selection of variable intensity and pervasiveness (**Supplementary Fig. 4**). Specifically, we used BS+MNM in Hyphy to simulate sequence evolution under the positive selection model with the human and *D. simulans* terminal branches as the foreground branches. We used genome-average estimates of all parameters, including gene length (418 and 510 codons for mammals and flies, respectively), but we varied ω_2 and p_2. 20 replicate alignments were simulated under each set of conditions and then analyzed using the BS test, the BS+MNM test, or MEME. For each set of conditions, the true positive rate was calculated as the fraction of replicates yielding a significant test of positive selection (P<0.05 for BS and BS+MNM, FDR<0.20 for at least one site in the alignment for MEME).

$$q_{ij} = \begin{cases} \kappa_1\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \omega\kappa_1\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ \omega\delta\kappa_2^2\pi_j & \text{non-synonymous, both transversions} \\ \omega\delta\pi_j & \text{non-synonymous, both transitions} \\ \omega\delta\kappa_2\pi_j & \text{non-synonymous, single transversion} \\ \delta\pi_j & \text{synonymous, both transitions} \\ \delta\kappa_2^2\pi_j & \text{synonymous, both transversions} \\ \delta\kappa_2\pi_j & \text{synonymous, single transversion} \\ 0 & \text{otherwise} \end{cases}$$

BS+MNM+ κ_2 model: We developed the BS+MNM+ κ_2 model, which incorporates into the BS+MNM model two different transversion:transition rate ratio parameters, κ_1 for single-site substitutions and κ_2 for MNMs. All free parameters of the model are estimated by ML given a sequence alignment. This model was implemented by further modifying our BS+MNM batchfile in Hyphy 2.2.6 software by declaring κ_2 a global variable, incorporating it into the codon table, and allowing it to be optimized by ML as other parameters are in the batch file.

For validation, we estimated the parameters of the BS+MNM+ κ_2 null model by ML for every alignment in each dataset and calculated the genome-average median estimate of each parameter. We then simulated 50 replicate alignments of length 418 and 510 codons in the mammalian and fly datasets respectively, under the BS+MNM+ κ_2 null model using the genome-average for all other model parameters. We then estimated each parameter by ML under the null model given each alignment and compared the distribution of estimates to the parameters used to generate the alignments.

To determine the effect of the MNM-specific transversion:transition rate on false-positive bias in the BS test, we simulated sequences 10,000 codons long under the BS+MNM+κ_2 null model, using genome-median parameters except κ_2, which we varied. For each value of κ_2, we simulated 50 replicates, applied the BS test, and calculated the FPR as the fraction of replicates yielding a positive inference (P<0.05).

Data availability. The empirical alignments reanalyzed in this study are available in the supplementary information of the original publications that generated these data [12, 16, 44].

Code availability. The custom HYPHY batch codes for the BS+MNM and BS+MNM+κ_2 tests are available as supplementary files and at https://github.com/JoeThorntonLab/MNM_SelectionTests.

REFERENCES

1. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725-736 (1994).
2. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764 (2012).
3. Murrell, B. et al. Gene-wide identification of episodic selection. *Mol Biol Evol* **32**, 1365-1371 (2015).
4. Smith, M. D. et al. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**, 1342-1353 (2015).
5. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908-917 (2002).
6. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**, 2472-2479 (2005).
7. Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679 (2005).
8. Kosiol, C., Holmes, I. & Goldman, N. An empirical codon model for protein sequence evolution. *Mol Biol Evol* **24**, 1464-1479 (2007).
9. Whelan, S. & Goldman, N. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**, 2027-2043 (2004).
10. Muse, S. V. & Gaut, B. S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**, 715-724 (1994).
11. Yang, Z. & dos Reis, M. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol* **28**, 1217-1228 (2011).
12. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203-218 (2007).
13. Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. & Hahn, M. W. Adaptive evolution of young gene duplicates in mammals. *Genome Res* **19**, 859-867 (2009).
14. Foote, A. D. et al. Convergent evolution of the genomes of marine mammals. *Nat Genet* **47**, 272-275 (2015).
15. Roux, J. et al. Patterns of positive selection in seven ant genomes. *Mol Biol Evol* **31**, 1661-1685 (2014).
16. Kosiol, C. et al. Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4**, e1000144 (2008).
17. Zhang, J. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol Biol Evol* **16**, 868-875 (1999).
18. Nozawa, M., Suzuki, Y. & Nei, M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A* **106**, 6700-6705 (2009).
19. Casola, C. & Hahn, M. W. Gene conversion among paralogs results in moderate false

detection of positive selection using likelihood methods. *J Mol Evol* **68**, 679-687 (2009).

20. Zhang, J. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol* **21**, 1332-1339 (2004).
21. Anisimova, M. & Yang, Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* **24**, 1219-1228 (2007).
22. Gharib, W. H. & Robinson-Rechavi, M. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol* **30**, 1675-1686 (2013).
23. Zhai, W., Nielsen, R., Goldman, N. & Yang, Z. Looking for Darwin in genomic sequences—validity and success of statistical methods. *Mol Biol Evol* **29**, 2889-2893 (2012).
24. Schrider, D. R., Hourmozdi, J. N. & Hahn, M. W. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol* **21**, 1051-1054 (2011).
25. Saribasak, H. et al. DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J Exp Med* **209**, 1075-1081 (2012).
26. Loeb, L. A. & Monnat, R. J. DNA polymerases and human disease. *Nat Rev Genet* **9**, 594-604 (2008).
27. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. Low fidelity DNA synthesis by human DNA polymerase-eta. *Nature* **404**, 1011-1013 (2000).
28. Seplyarskiy, V. B., Bazykin, G. A. & Soldatov, R. A. Polymerase ζ Activity Is Linked to Replication Timing in Humans: Evidence from Mutational Signatures. *Mol Biol Evol* **32**, 3158-3172 (2015).
29. Stone, J. E., Lujan, S. A., Kunkel, T. A. & Kunkel, T. A. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae. *Environ Mol Mutagen* **53**, 777-786 (2012).
30. Besenbacher, S. et al. Multi-nucleotide de novo Mutations in Humans. *PLoS Genet* **12**, e1006315 (2016).
31. Chen, J. M., Férec, C. & Cooper, D. N. Complex Multiple-Nucleotide Substitution Mutations Causing Human Inherited Disease Reveal Novel Insights into the Action of Translesion Synthesis DNA Polymerases. *Hum Mutat* **36**, 1034-1038 (2015).
32. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res* **24**, 1445-1454 (2014).
33. Arana, M. E., Seki, M., Wood, R. D., Rogozin, I. B. & Kunkel, T. A. Low-fidelity DNA synthesis by human DNA polymerase theta. *Nucleic Acids Res* **36**, 3847-3856 (2008).
34. Pinto, Y. et al. Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity. *Genome Res* **26**, 579-587 (2016).
35. Nowarski, R. & Kotler, M. APOBEC3 cytidine deaminases in double-strand DNA break repair and cancer promotion. *Cancer Res* **73**, 3494-3498 (2013).
36. Chen, J. M., Cooper, D. N. & Férec, C. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: ~0.4% of the single-nucleotide substitution mutation rate. *Hum Mutat* **35**, 392-394 (2014).

37. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-766 (2011).

38. Francioli, L. C. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822-826 (2015).

39. Zhu, W. et al. Concurrent nucleotide substitution mutations in the human genome are characterized by a significantly decreased transition/transversion ratio. *Hum Mutat* **36**, 333-341 (2015).

40. Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S. & Kondrashov, A. S. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature* **429**, 558-562 (2004).

41. Rogozin, I. B. et al. Evolutionary switches between two serine codon sets are driven by selection. *Proc Natl Acad Sci U S A* **113**, 13109-13113 (2016).

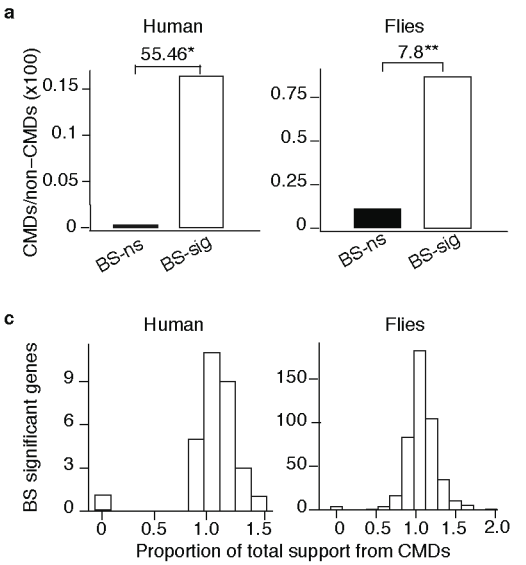
42. Averof, M., Rokas, A., Wolfe, K. H. & Sharp, P. M. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**, 1283-1286 (2000).

43. Suzuki, Y. False-positive results obtained from the branch-site test of positive selection. *Genes Genet Syst* **83**, 331-338 (2008).

44. Larracuente, A. M. et al. Evolution of protein-coding genes in Drosophila. *Trends Genet* **24**, 114-123 (2008).

45. Terekhanova, N. V., Bazykin, G. A., Neverov, A., Kondrashov, A. S. & Seplyarskiy, V. B. Prevalence of multinucleotide replacements in evolution of primates and Drosophila. *Mol Biol Evol* **30**, 1315-1325 (2013).

46. Sironi, M., Cagliani, R., Forni, D. & Clerici, M. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat Rev Genet* **16**, 224-236 (2015).



d

	Human		Flies	
PP	CMD (32)	NonCMD (19563)	CMD (2125)	NonCMD (244877)
<0.5	0	19512	402	242518
0.5-0.9	11	37	578	2224
>0.9	21	14	1145	135

Figure 1 Codons with multiple nucleotide differences (CMDs) drive branch-site signatures of selection.

(a) CMDs are enriched in genes with a signature of positive selection. Codons were classified by the number of nucleotide differences between the ancestral and terminal states on branches tested for positive selection. CMDs have ≥2 differences; non-CMDs have ≤1 difference. The CMD/non-CMD ratio is shown for genes with a significant signature of selection in the BS test (BS-sig) and those without (BS-ns). Fold-enrichment is shown as the odds ratio. *, P=1e-41 by Fisher's exact test; **, P=4e-4 by chi-square test.

(b) Percentage of genes that retain a signature of positive selection when CMDs are excluded from the branch-sites test analysis.

(c) Distribution across BS-significant genes of the proportion of total support for the positive selection model that is provided by CMDs. Total support is the difference in log-likelihood between the positive selection and null models, summed over all codons in the alignment. Support from CMDs is summed over codons with multiple differences. The proportion of support from CMDs can be greater than 1 if the log-likelihood difference between models is negative at non-CMDs.

(d) Most codons classified as positively selected are CMDs. The number of CMDs and non-CMDs in BS-significant genes are shown according to their Bayes Empirical Bayes posterior probability (PP) of being in the positively selected class.

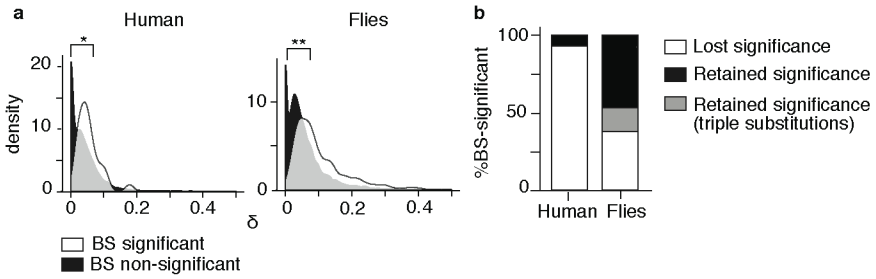


Figure 2 Incorporating MNMs into the branch-sites model eliminates the signature of positive selection in many genes. The mammalian and fly datasets were reanalyzed using a version of the BS test that allows MNMs (BS+MNM) by including a parameter δ, the rate of double substitutions relative to single substitutions.

(a) The distribution of ML estimates of δ across genes with (white) and without (black) a significant result in the classic BS test is shown for empirical alignments. Median estimates of δ for BS-significant and BS-nonsignificant genes are 0.047 and 0.026 in humans, respectively, and 0.107 and 0.062 in flies. *, P=1e-4; **, P=1e-8 by Mann-Whitney U Test.

(b) Proportion of genes with a significant result in the BS test that lose or retain that signature using the BS+MNM test. Genes that remain significant but contain CMDs with three differences, which are not incorporated into BS+MNM, are also shown.

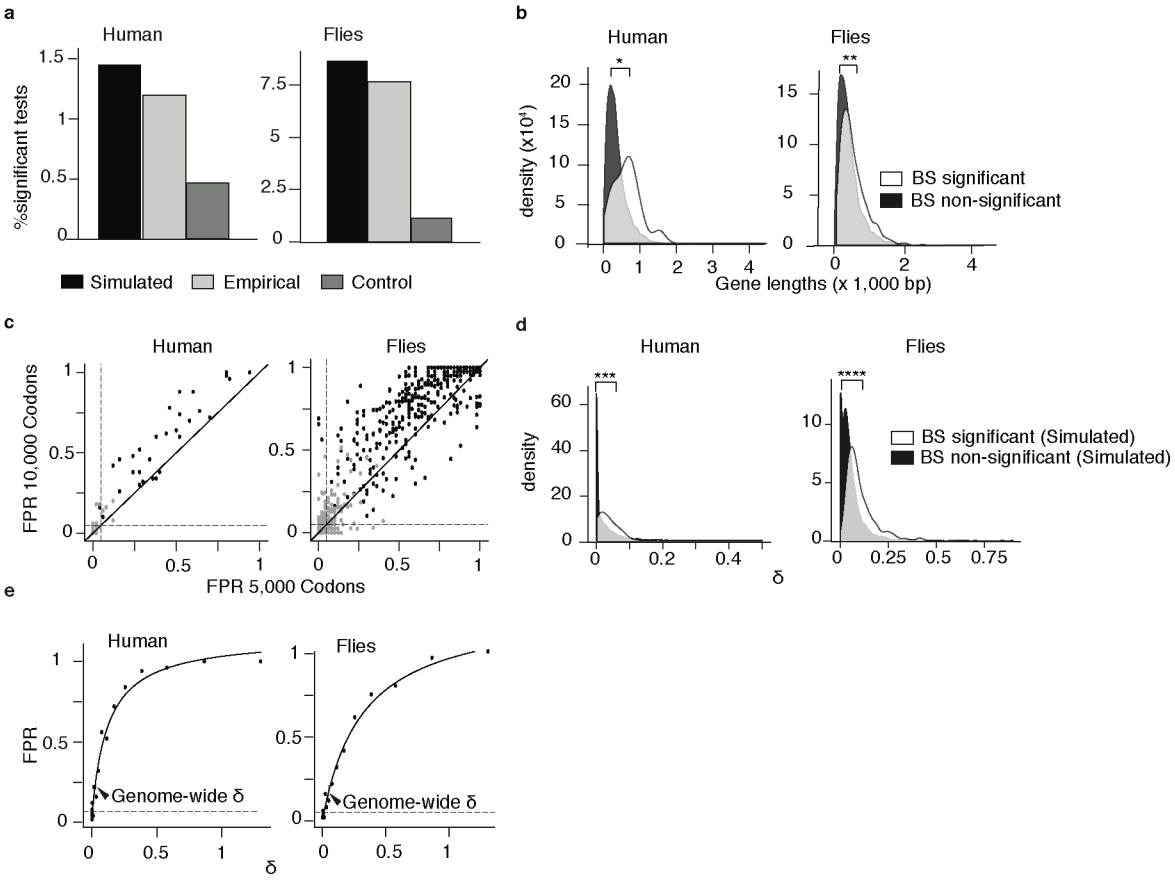


Figure 3 MNMs cause a strong bias in the branch-site test under realistic conditions. For each gene in the mammalian and fly datasets, the parameters of the BS+MNM null model were estimated by maximum likelihood. We then simulated sequence evolution under each gene's inferred null parameters and used the classic BS test on the simulated alignments to test for positive selection on the human and terminal fly lineages.

(a) The fraction of all tests that are BS-significant (P<0.05) is shown for the data simulated under the BS+MNM null model, the original empirical sequence alignments, and a control dataset simulated with $\delta = 0$. Each gene's length in the simulation was identical to its empirical length.

(b) BS-significant genes are longer than BS non-significant genes. The probability density of gene lengths in the two categories is shown for the empirical mammalian and fly datasets. Median lengths in BS-significant and non-significant genes, respectively, were 642 and 343 bp in humans; in flies, 448 and 399 bp. The difference between the two distributions was evaluated using a Mann-Whitney U test. *, P=8e-5; **, P=8e-4.

(c) Systematic bias in the BS test. For each gene with a significant result in the BS test using the empirical data, we simulated 50 replicates using the BS+MNM null model and the ML parameter estimates for that gene at lengths of 5,000 and 10,000 codons; these data were then analyzed using the BS test. The false positive rate (FPR) for any gene's simulation (black points) is the proportion of replicates with P<0.05. Gray points show FPR for control simulations with $\delta = 0$. Dashed lines, FPR of 0.05. The solid diagonal line has a slope of 1.

(d) The distribution of ML estimates of δ across genes with (white) and without (black) a signature of positive selection in the classic BS test is shown for data simulated under the BS+MNM null model. Median δ in BS-significant and BS-nonsignificant genes = 0.03 and 0.0009 in humans, 0.04 and 0.08 in flies. Difference between the distributions was evaluated using a Mann-Whitney U Test: ***, P=1e-199;* ***, P=1e-12.

(e) Increasing the MNM rate increases bias in the BS test. Sequences 5000 codons long were simulated using the BS+MNM model and the median value of each model parameter and branch length across all genes in each dataset, but δ was allowed to vary. The rate of false positives (P<0.05) in 50 replicates at each value of δ is shown. Solid line, hyperbolic fit to the data; dotted line, FPR level of 5%. Arrowhead, median δ across all genes.

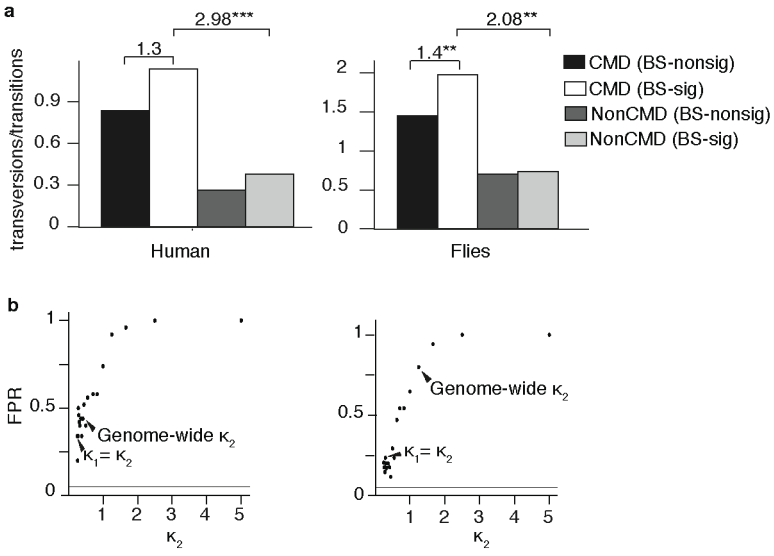


Figure 4 Transversion-enrichment in CMDs biases the BS test.

(a) The ratio of transversions:transitions observed in CMDs and in non-CMDs is shown for BS-significant and BS-nonsignificant genes. Fold-enrichment is shown as the odds ratio. *, P=3e-25; **, P=5e-4 by Fisher's exact test.

(b) Increasing the transversion rate in MNMs increases bias of the BS test. Sequences 10,000 codons long were simulated using an elaboration of the BS+MNM model that allows MNMs to have a transversion:transition rate (κ_2) different from that in single-nucleotide substitutions (κ_1). 50 replicate alignments were simulated under the null model using the average value of every model parameter and branch length across all genes in each dataset, except κ_2 was allowed to vary. The rate of false positives ($P<0.05$) at each value of κ_2 is shown. Arrowheads show the mean value of κ_2 across all genes in each dataset. Dotted line, FPR of 5%.

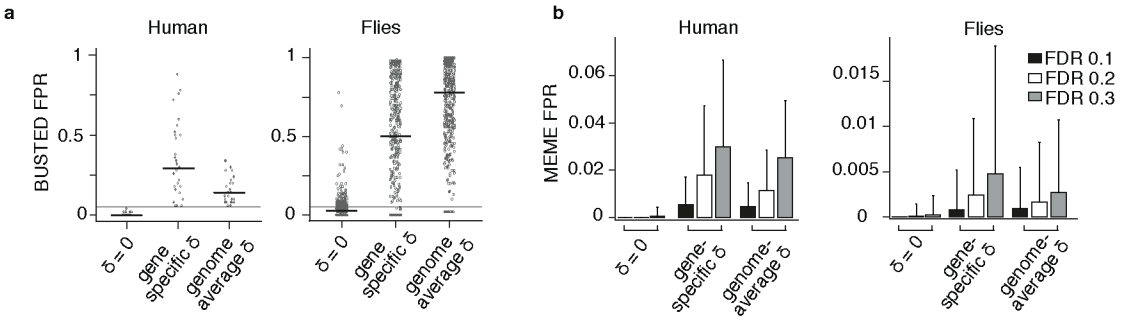


Figure 5 MNMs bias newer tests of positive selection.

(a) False positive inferences under realistic conditions using BUSTED. For every BS-significant gene in each dataset, 50 replicate alignments 5,000 codons long were simulated using the BS+MNM null model and parameter values estimated from the empirical sequences. These alignments were then analyzed for a signature of positive selection (P<0.05) using BUSTED. δ was assigned to its gene-specific estimate, to its average across all genes in each dataset, or to zero. FPR is the proportion of replicate alignments for each gene with P<0.05. Each dot represents the FPR for one gene; black bar, median across genes.

(b) Bias in MEME caused by MNMs. Alignments described in panel (a) were analyzed using MEME, which tests for episodic positive selection at every codon. FPR per gene is the proportion of replicate alignments simulated under that gene's parameters that have at least one significant site after FDR correction, using thresholds of varying stringency. Column height shows the mean FPR across genes; error bars show one standard deviation above the mean; 1 sd below the mean includes zero in all categories.

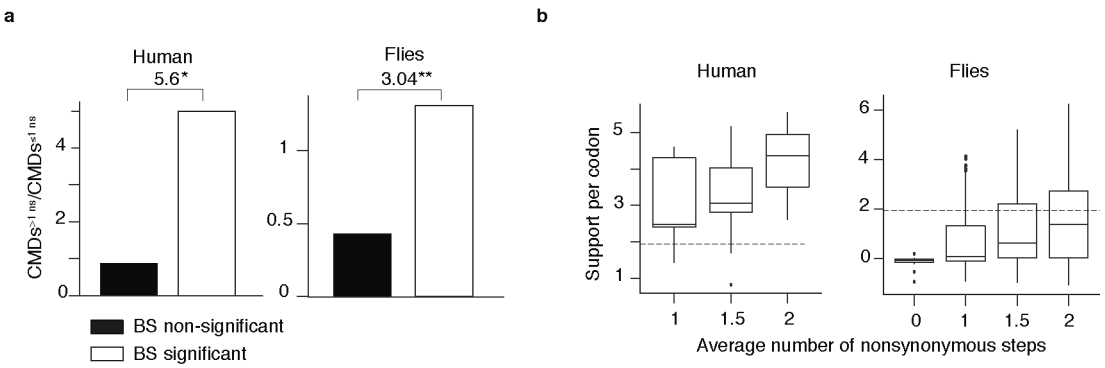


Figure 6 CMDs implying multiple nonsynonymous steps drive the BS test.

(a) For every CMD, the mean of the number of nonsynonymous single-nucleotide steps on the two direct paths between the ancestral and derived states was calculated. In BS-significant and BS-nonsignificant genes, the ratio of CMDs invoking more than one nonsynonymous step to those invoking one or fewer such steps is shown. Fold-enrichment is shown as the odds ratio; *, P=9e-04; **P= 1.6e-67 by Fisher's exact test.

(b) Support for the positive selection model provided by CMDs depends on the number of implied nonsynonymous single-nucleotide steps. Support is the log-likelihood difference between the positive selection and null models of the BS test given the data at a single codon site. Box plots show the distribution of support by CMDs in BS significant genes categorized according to the mean number of implied nonsynonymous steps. Dotted line, support of 1.92, at which the BS test yields a significant result for an entire gene (P<0.05). In human BS-significant genes, no CMDs imply zero non-synonymous changes.

Supplementary Table 1. Paths between codon pairs

Average steps on path (nonsyn, syn)	Number of pairs
0,2	8
0.5,1.5	0
1,1	588
1.5,0.5	308
2,0	548

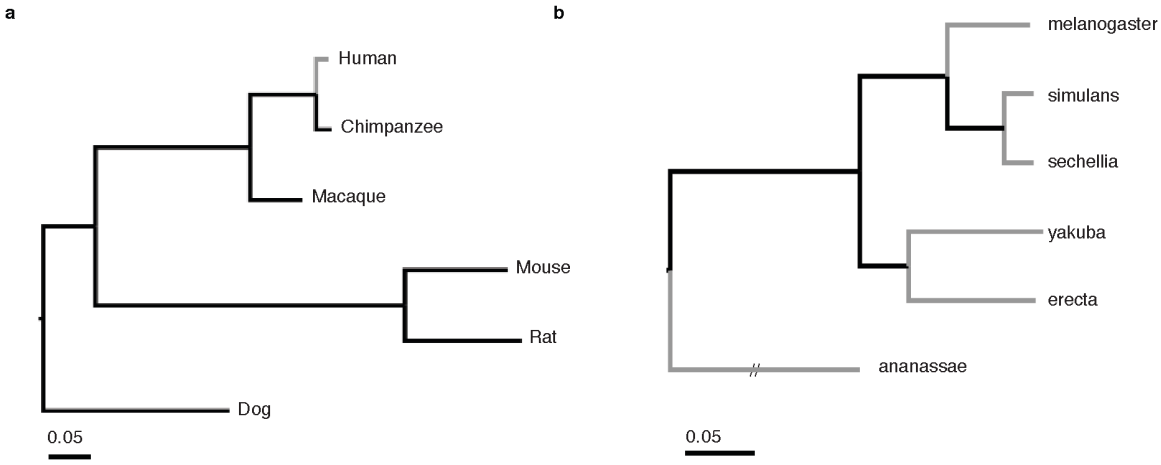
For each possible codon pair separated by 2 nucleotide differences, the universal genetic code was parsed to tabulate the mean number of nonsynonymous and synonymous steps (nonsyn, syn) on the two direct paths between them. Paths with stop codons were excluded.

Supplementary Table 2. Number of gene alignments passing each filtering step in our analysis.

	Humans			Flies		
	Empirical	Simulated	Control	Empirical	Simulated	Control
All genes	16,541	6,868	6,868	8,564	8,564	8564
Genes with complete species coverage	6,868	6,868	6,868	8,564	8,564	8564
BS tests significant at P<0.05 (% of tests)	82 (1.1%)	99 (1.4%)	32 (0.5%)	3,938 (7.6%)	4,444 (8.6%)	582 (1.1%)
BS tests significant after correction (FDR <20%)	0	0	0	2,147	1,755	4
BS-significant genes with unambiguous ancestral codon reconstructions	30*	-	-	443*	-	-

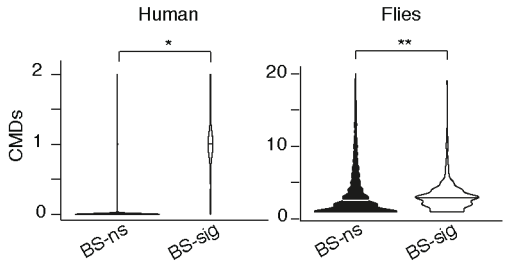
Filtering steps are described in Methods. Empirical data are all genome-wide coding alignments from the studies by Kosiol *et al*., and Larracuente *et al*.[12,44] Simulated data are alignments simulated under the BS+MNM null model using parameters derived from the empirical data. Control data are alignments simulated as above but with MNM rate parameter $\delta = 0$. *, in humans, the empirical BS-significant set includes genes that pass the filter for ancestral reconstruction of CMDs but not for FDR adjustment; in flies, both criteria are met. The total number of tests on the 6 fly lineages is 51,384.

Supplementary Figure 1: Mammalian and fly phylogenies



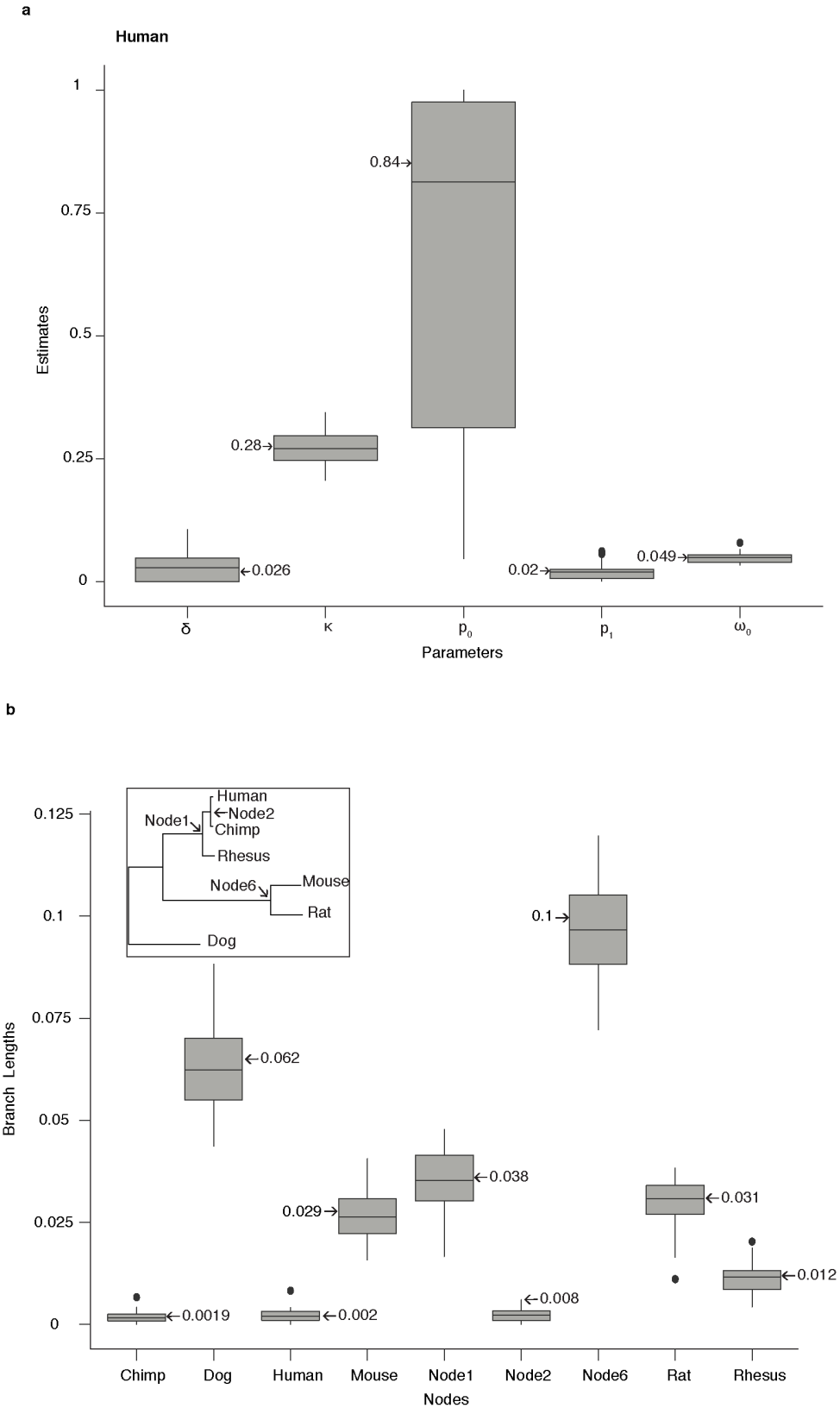
Phylogenies of mammalian and *Drosophila* species used in this analysis. The BS test was used to identify genes under positive selection on each of the lineages colored in grey. Branch lengths are proportional to the median length across all genes analyzed. Scale bar, expected nucleotide substitutions per codon; the hatched branch, shortened for display, has length 0.62.

Supplementary Figure 2: Distibution of CMDs in BS-significant and non-significant genes

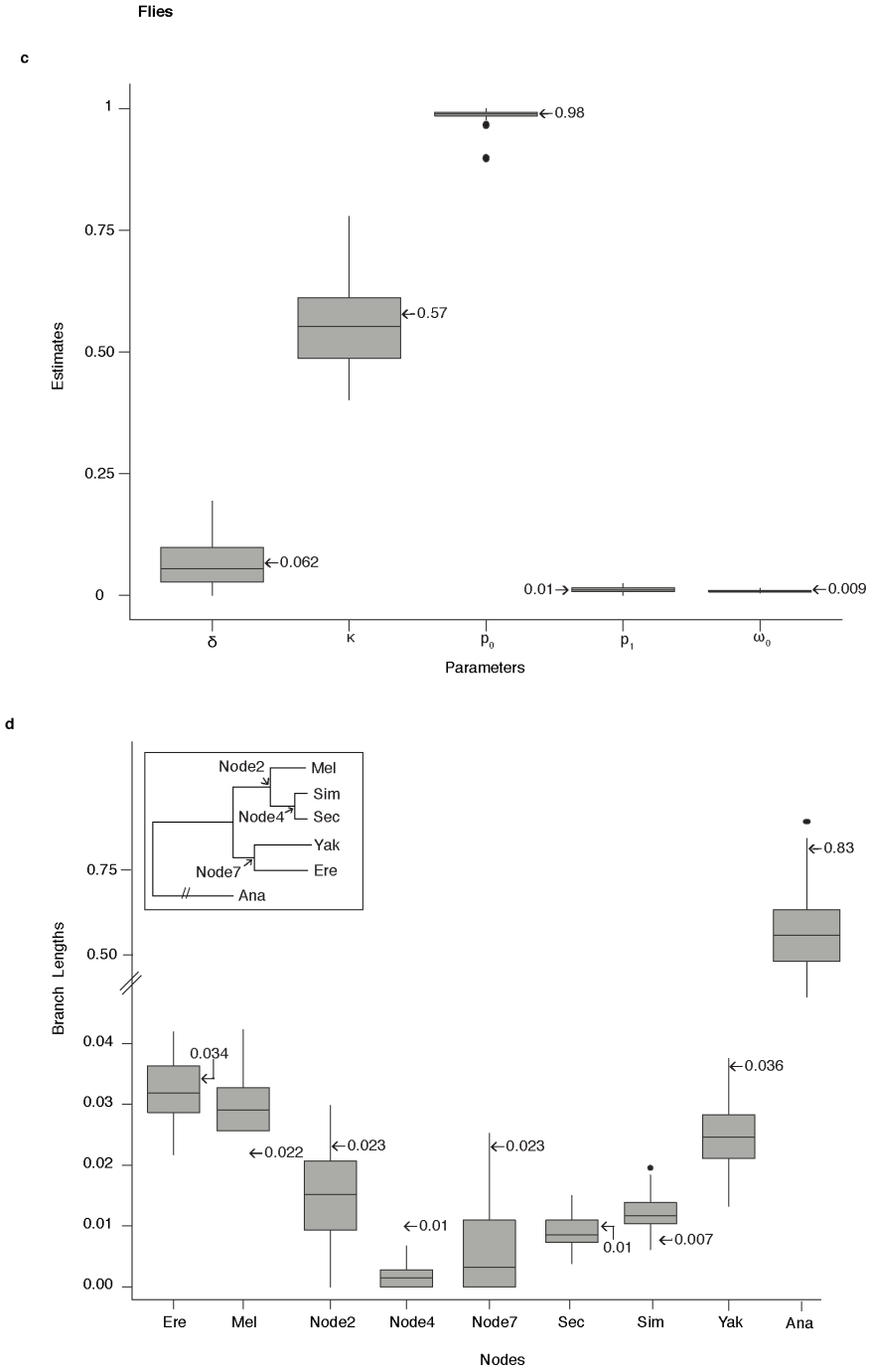


Distribution of the number of CMDs per gene in BS-significant and BS-nonsignificant (ns) genes. Horizontal line in each violin plot indicates the median. *, P<2e-16; **, P=4e-10, Mann Whitney U test.

Supplementary Figure 3: BS+MNM model validation

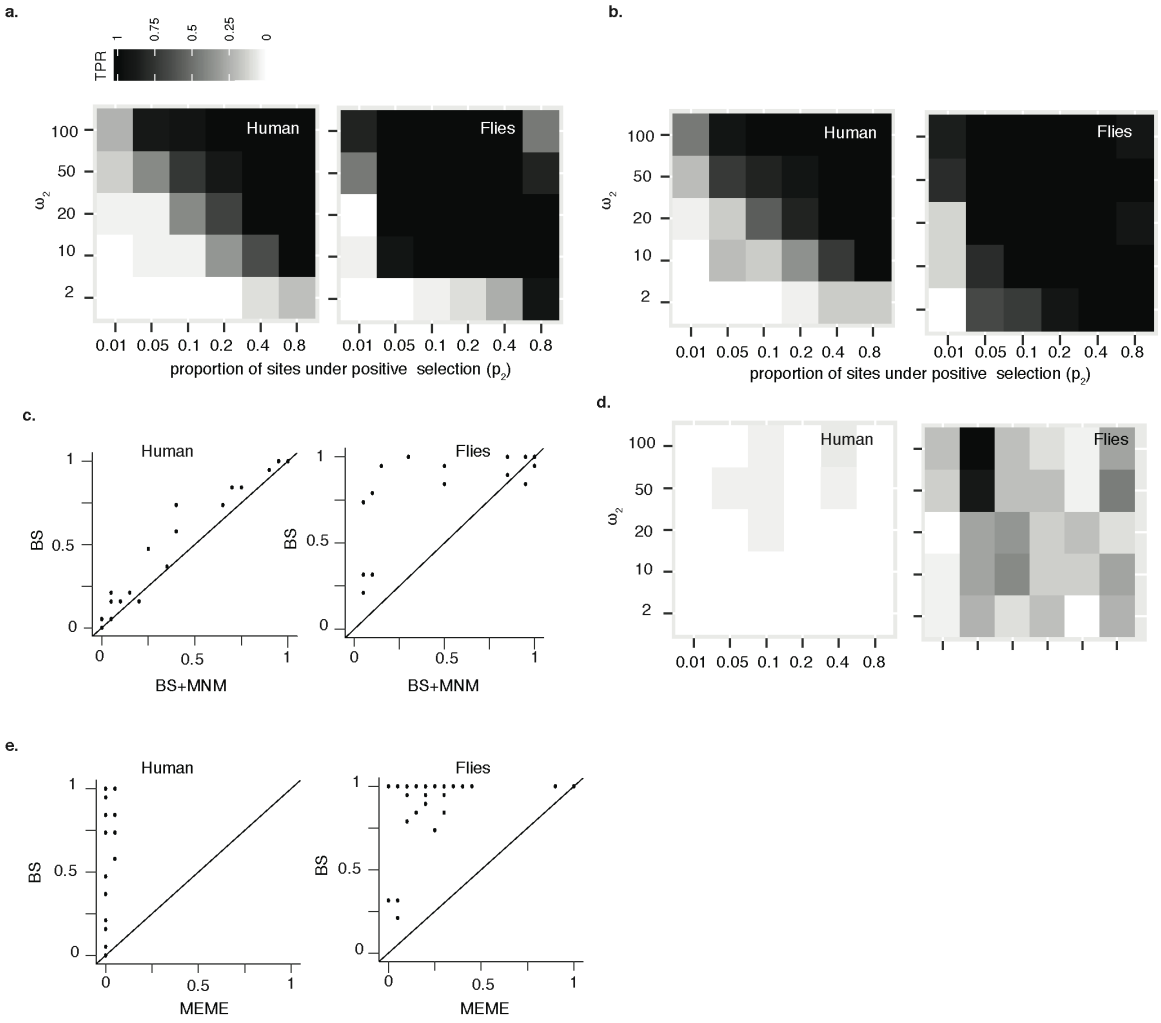


Supplementary Figure 3: BS+MNM model validation



Validation of parameter estimation in the BS+MNM model. 50 replicate alignments were simulated under the BS+MNM null model with genome-average parameters, including gene length; the BS+MNM null model was then used to estimate the parameters from each replicate. Box plots show the distribution of estimates across alignments for model parameters (**a**) and branch lengths (**b**) in the mammalian dataset and in the fly dataset (**c and d).** Arrows indicate the generating parameters used in the simulation. Node names in panels (b) and (d) correspond to those in the phylogenies shown in the inset.

Supplementary Figure 4: Power to detect positive selection: BS+MNM test, classic BS test and MEME



Analysis of power to detect positive selection by various tests.

(a) Power of BS+MNM test. Sequence alignments of genome-average length were simulated using the BS+MNM model with genome-average parameters and branch lengths. The proportion of sites under positive selection (p_2) and the strength of positive selection (ω_2) were varied, with 20 replicate simulations under each set of conditions. The BS+MNM test was applied and the rate of true positive inferences (TPR) was calculated as the fraction of replicates under each condition with a significant result ($P < 0.05$). Cells in the grid are shaded by TPR and the heat map key, shown at top.

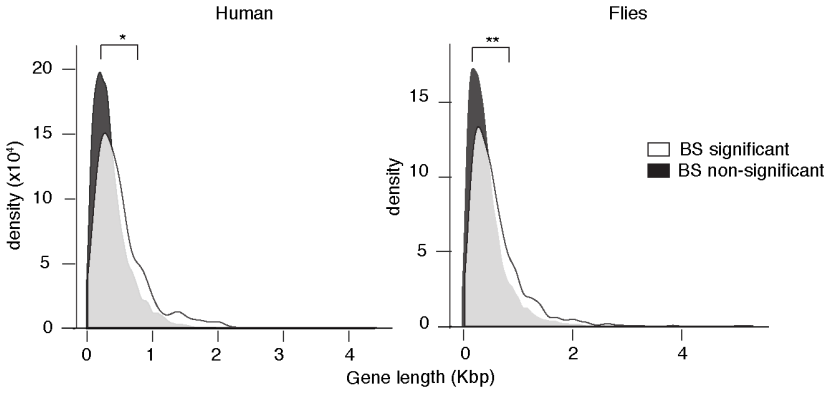
(b) Power of the classic BS test. The data as in (a) were analyzed using the BS test.

(c) Power comparison of the BS+MNM and BS tests. Each dot represents a set of parameter values used to simulate sequence evolution, located by the TPRs achieved by the two tests, as in (a) and (b).

(d) Power of MEME test of episodic selection. MEME was used to analyze the same sequences used in (a). A gene was inferred to be under positive selection if one or more codons in it was significant after correction to FDR <20%.

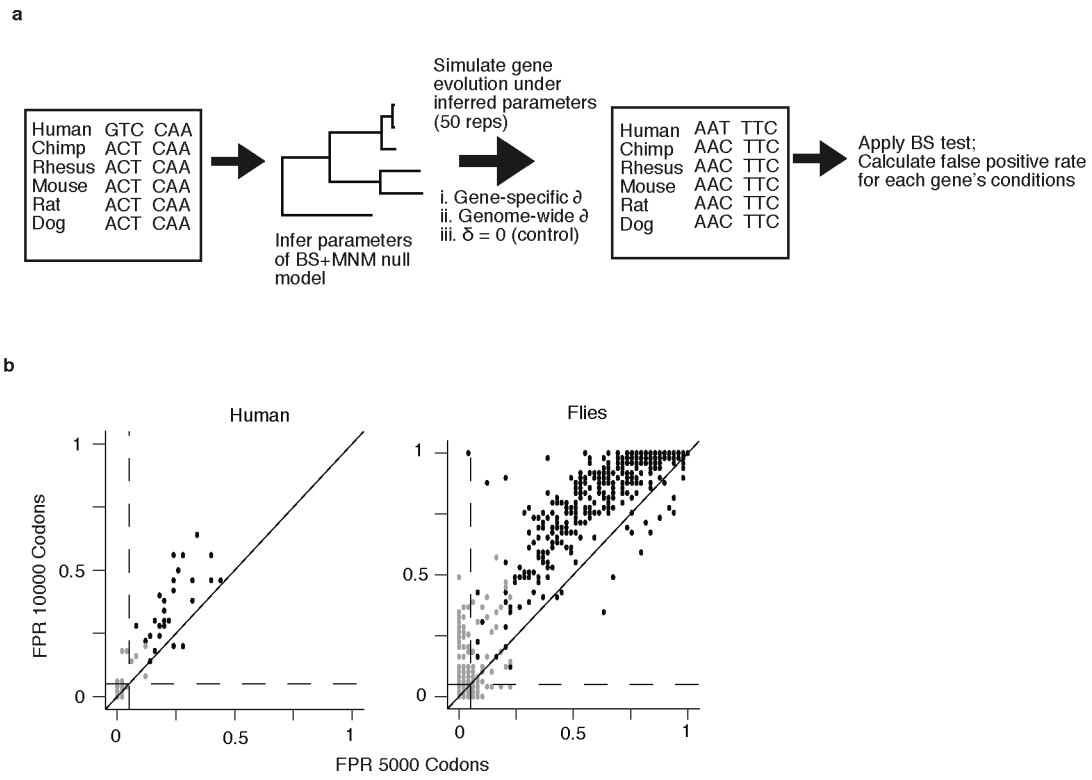
(e) Power comparison of MEME and the BS test. Each dot represents TPRs achieved by the two tests on sequences simulated under one set of parameters, as in (b) and (d).

Supplementary Figure 5: BS-significant genes in the genome-length simulation experiment are longer than BS-non significant genes



Longer genes are more likely to yield false positive BS tests. For each empirical gene in the mammalian and fly datasets, the parameters of the BS+MNM null model were estimated by maximum likelihood. We then simulated sequence evolution under each gene's inferred null parameters and empirical length and used the classic BS test on the simulated alignments to test for positive selection on the human and terminal fly lineages. The distribution of the lengths of genes yielding a BS-significant test ($P<0.05$) or a BS-nonsignificant test is shown. Median lengths in BS-significant and non-significant genes, respectively, were 422 and 343 bp in humans; in flies, 484 and 391 bp. Differences between distributions were evaluated using Mann-Whitney U test. *, P=5e-3; **, P=2e-23.

Supplementary Figure 6: MNMs bias the classic BS test

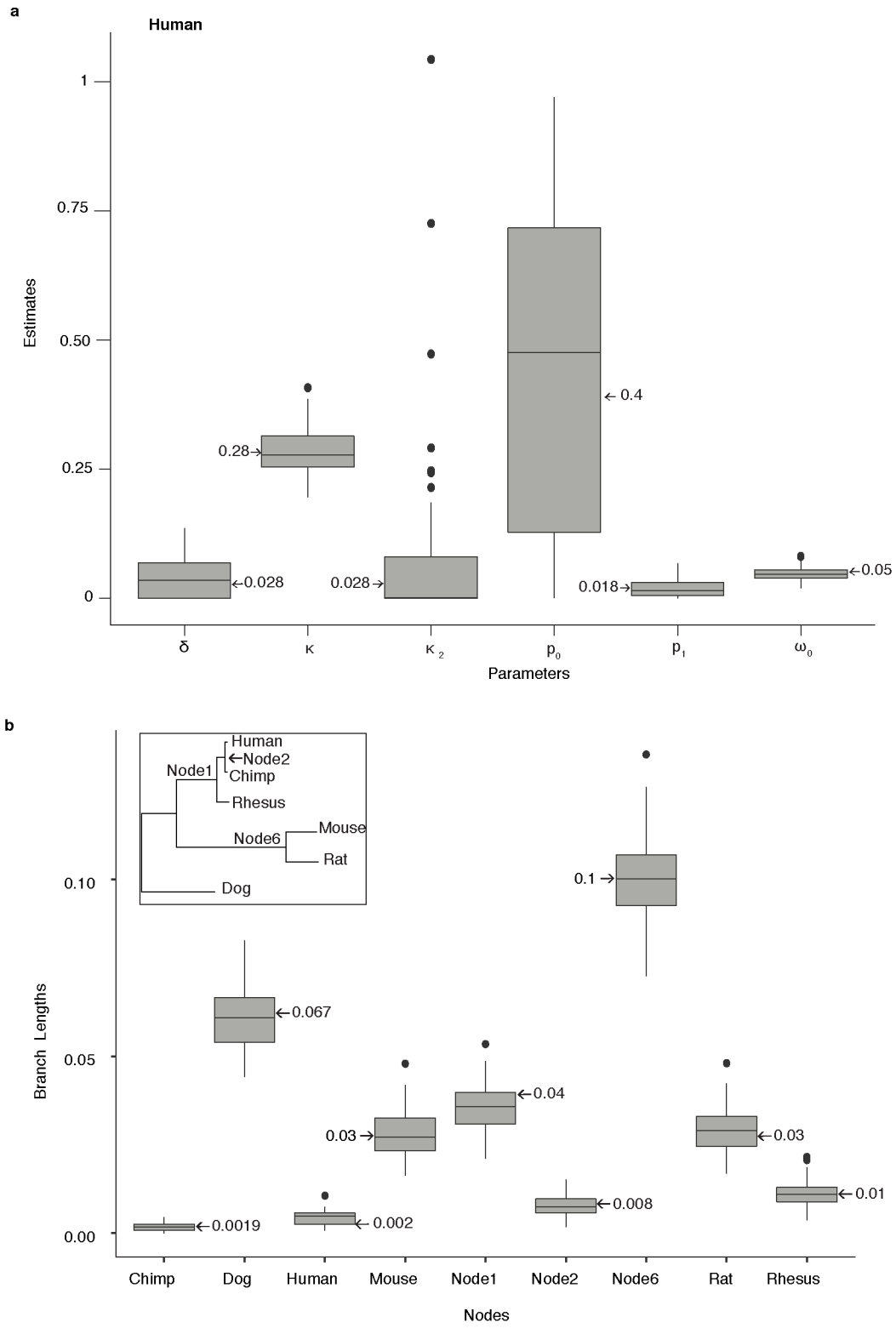


MNMs bias the classic BS test.

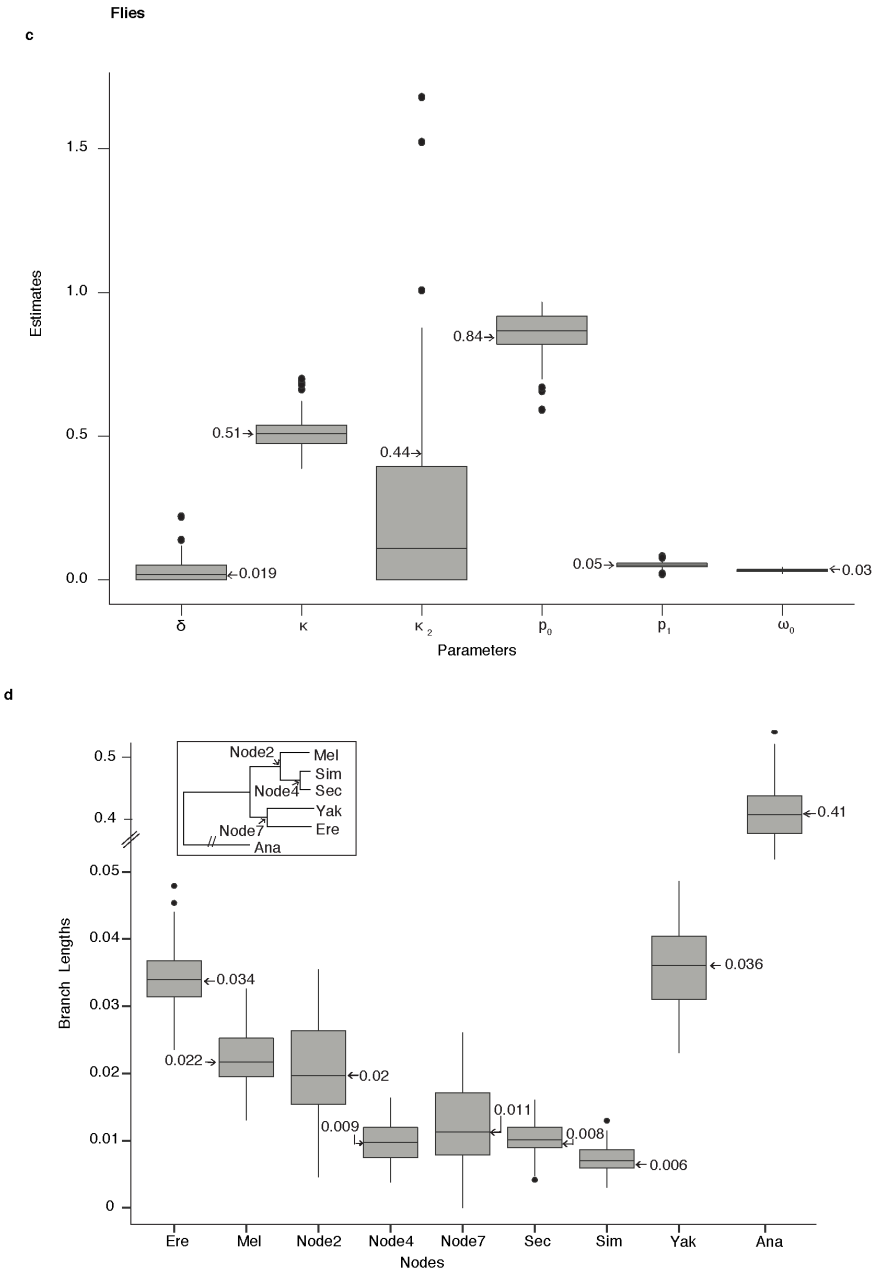
(a) Scheme to simulate genes with MNMs without positive selection. For each BS-significant empirical gene, the parameters of the BS+MNM null model were estimated. Using each set of parameters, 50 replicate alignments were simulated, with δ (the relative rate of MNM substitution) assigned to its gene-specific value, its median across all genes in the dataset (genome-wide average), or to zero. The classic BS test was applied to simulated data, and the false positive rate (FPR) for each set of generating parameters was calculated as the fraction of replicates with a positive result ($P<0.05$).

(b) Systematic bias in the BS test using the genome-average MNM rate. 50 replicate alignments 5,000 or 10,000 codons long were simulated under the BS+MNM null model using gene-specific parameters inferred as in (a). Each black point represents FPRs for sequences 5,000 and 10,000 codons long simulated under one empirical gene's parameters and the genome-wide average δ. Gray points show FPRs for control simulations with $\delta = 0$. Dashed lines, FPR of 0.05. Diagonal line has a slope of 1.

Supplementary Figure 7: BS+MNM+K2 model validation



Supplementary Figure 7: BS+MNM+K2 model validation



Validation of parameter estimates by BS+MNM+κ_2 model. 50 replicate alignments were simulated under the BS+MNM+κ_2 null model using genome-average parameters. Parameters were then re-estimated given each alignment using the same model. Box plots show the distribution of estimates of model parameters (**a**) and branch length (**b**) in humans and in flies (**c, d**). Arrows, generating parameters used to simulate the data. Node names correspond to those in the phylogenies shown in the inset.