ARTICLES
https://doi.org/10.1038/s41559-018-0584-5
nature ecology & evolution

Multinucleotide mutations cause false inferences of lineage-specific positive selection

Aarti Venkat[1], Matthew W. Hahn[2,3] and Joseph W. Thornton[1,4]*

Phylogenetic tests of adaptive evolution, such as the widely used branch-site test (BST), assume that nucleotide substitutions occur singly and independently. Recent research has shown that errors at adjacent sites often occur during DNA replication, and the resulting multinucleotide mutations (MNMs) are overwhelmingly likely to be non-synonymous. To evaluate whether the BST misinterprets sequence patterns produced by MNMs as false support for positive selection, we analysed two genome-scale datasets—one from mammals and one from flies. We found that codons with multiple differences account for virtually all the support for lineage-specific positive selection in the BST. Simulations under conditions derived from these alignments but without positive selection show that realistic rates of MNMs cause a strong and systematic bias towards false inferences of selection. This bias is sufficient under empirically derived conditions to produce false positive inferences as often as the BST infers positive selection from the empirical data. Although some genes with BST-positive results may have evolved adaptively, the test cannot distinguish sequence patterns produced by authentic positive selection from those caused by neutral fixation of MNMs. Many published inferences of adaptive evolution using this technique may therefore be artefacts of model violation caused by unincorporated neutral mutational processes. We introduce a model that incorporates MNMs and may help to ameliorate this bias.

Identifying genes that evolved under the influence of positive natural selection on phylogenetic time scales is a central goal in studies of molecular evolution. Of the many methods developed for this purpose[1–10], the most widely used is the branch-site test (BST)[5,6]. This technique has been the basis for published claims of lineage-specific adaptive evolution in many thousands of genes[11–15].

The BST uses a likelihood ratio test to compare two probabilistic models of sequence evolution, given an alignment of coding sequences. The null model constrains all codons to evolve with rates of non-synonymous substitution (d_N) less than or equal to the rate of synonymous substitution (d_S), as expected under drift and purifying selection alone. In the positive selection model, some sites are allowed to have $d_N > d_S$ on one or more branches of interest. If the positive selection model increases the likelihood more than is expected by chance, the null model is rejected and adaptive evolution is inferred. The BST is conservative in the absence of model violation, with a low rate of false positive results when sequences are generated according to the null model[6,16]. Although likelihood ratio tests can be biased if the underlying probabilistic model is incorrect[17], the BST has been found to be reasonably robust to several forms of model violation[6,18–24].

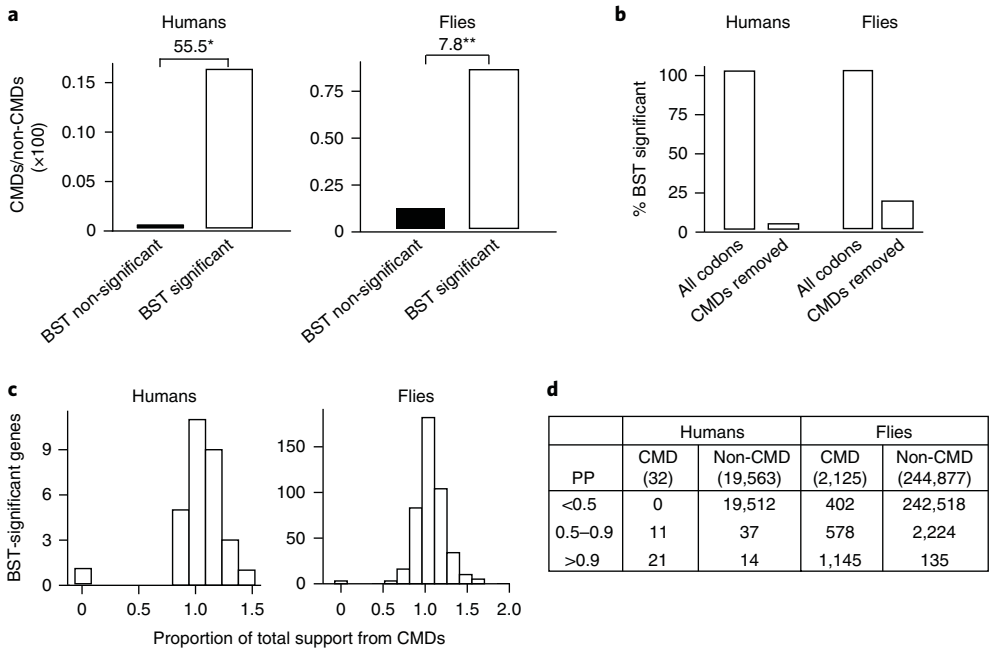
A recently discovered genetic phenomenon—the propensity of DNA polymerases to produce mutations at neighbouring sites—has not been evaluated for its effect on the BST. All current models for identifying positive selection assume that mutations are fixed singly and independently at individual nucleotide sites: codons with multiple differences (CMDs) can be interconverted only by serial single-nucleotide substitutions, the probability of which is the product of the probabilities of each independent event. But molecular studies of replication show that some polymerases are prone to making adjacent mutations[25–33]. In human trios and laboratory organisms, de novo mutations often occur in tandem or at nearby sites more frequently than expected if each occurred independently[25,32–36]. The precise frequency at which multinucleotide mutations (MNMs) occur is difficult to estimate, but a recent study concluded that about 0.4% of mutations, polymorphisms, and substitutions in humans are at directly adjacent sites (counting each tandem pair as one event)[34]. In *Drosophila melanogaster*, analysis of rare polymorphisms and mutation-accumulation experiments estimated that 1.3% of all mutations are at adjacent sites[37]. Tandem MNMs therefore appear to account for on the order of 1% of mutations.

We hypothesized that MNMs might lead to false signatures of positive selection in the BST and related tests. Because of the structure of the genetic code, virtually all MNMs in coding sequences are non-synonymous, and most would require multiple non-synonymous changes if they were to occur by single-nucleotide steps (Supplementary Table 1). Furthermore, MNMs tend to be enriched in transversions[35,38,39], and transversions are more likely than transitions to be non-synonymous. MNMs are therefore likely to produce CMDs containing an apparent excess of non-synonymous substitutions, even in the absence of positive selection. When these data are assessed, assuming that all substitutions are independent, a model that allows d_N to exceed d_S at some sites may have significantly higher likelihood, potentially leading to false inferences of positive selection. CMDs can also be fixed by positive selection[16,40–42], but current methods may fail to distinguish selected CMDs from those produced by neutral fixation of MNMs. Simulations suggest that MNMs may increase the rate of positive inference in the BST and related selection tests[43,44], but there has been no comprehensive analysis of the effect of MNMs under realistic, genome-scale conditions.

Results

We analysed two previously published genome-scale datasets, which represent classic examples of the application of the BST[12,14,45].

[1]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [2]Department of Biology, Indiana University, Bloomington, IN, USA. [3]Department of Computer Science, Indiana University, Bloomington, IN, USA. [4]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA.
*e-mail: joet1@uchicago.edu



d

	Humans		Flies	
PP	CMD (32)	Non-CMD (19,563)	CMD (2,125)	Non-CMD (244,877)
<0.5	0	19,512	402	242,518
0.5–0.9	11	37	578	2,224
>0.9	21	14	1,145	135

Fig. 1 | 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, whereas non-CMDs have ≤1 difference. The CMD/non-CMD ratio is shown for genes with a significant signature of selection in the BST and those without. Fold-enrichment in BST-significant versus non-significant gene sets is shown above the columns as the odds ratio. $^{*}P = 4 \times 10^{-4}$ by χ^2 test; $^{**}P = 1 \times 10^{-41}$ by Fisher's exact test. **b**, Percentage of genes that retain a signature of positive selection when CMDs are excluded from the BST analysis. **c**, Distribution across BST-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 CMDs. The proportion of support from CMDs can be greater than one if the log-likelihood difference between models is negative at non-CMDs. **d**, Most codons classified as positively selected are CMDs. The numbers of CMDs and non-CMDs in BST-significant genes are grouped by the Bayes empirical Bayes posterior probability (PP) that they are in the positively selected class.

The mammalian dataset consists of coding sequences of 16,541 genes from six species; we retained for analysis only the 6,868 genes with complete species coverage. The fly dataset consists of 8,564 genes from six *Drosophila* species, all of which had complete coverage (Supplementary Fig. 1).

We used the BST to identify genes putatively under positive selection ($P < 0.05$) on the human lineage in the mammalian dataset and on each of the six terminal lineages in flies. A total of 82 genes in humans and 3,938 tests in flies yielded significant tests ($P < 0.05$; Supplementary Table 2). Filtering for data quality and correcting for multiple testing (false discovery rate (FDR) < 0.2) yielded 443 fly genes for further analysis. In total, 30 human genes passed the quality filter, but none survived the multiple testing correction, consistent with previous analyses[14]. Nevertheless, we included the 30 initially significant and high-quality genes because this lineage is the object of intense interest and because its short length contrasts with the fly branches, allowing us to examine the performance of the BST under different conditions. These two sets constitute the 'BST-significant' genes in flies and humans.

CMDs provide virtually all support for positive selection. We found that CMDs are the primary drivers of BST-positive results. CMDs are dramatically enriched in BST-significant genes relative to non-BST-significant genes (Fig. 1a and Supplementary Fig. 2). When CMD-containing sites are excluded from the alignments, the vast majority of genes that were BST-significant lose their signature of selection (Fig. 1b). In virtually all BST-significant genes, >95% of the statistical support for positive selection, defined as the fraction of the total log-likelihood difference between the positive selection and null models, comes from CMDs; in about 70% of genes, CMDs provide all the support (Fig. 1c). CMDs also account for 60 and 90% of sites inferred a posteriori to have been positively selected (posterior probability > 0.9) in humans and flies, respectively, although they represent <1% of all codons (Fig. 1d).

Incorporating MNMs eliminates the signature of positive selection in many genes. CMDs could be enriched in BST-positive genes because of an MNM-induced bias or because they were fixed by positive selection. To distinguish between these possibilities, we implemented a version of the BST that is identical to the classic version, but its model allows double-nucleotide changes using an additional parameter δ, which scales the rate of each double-nucleotide substitution relative to single-nucleotide substitutions. We evaluated our implementation of this BS + MNM model using simulations under realistic conditions and found that parameters are estimated with reasonable accuracy (Supplementary Fig. 3). When fit to all empirical mammalian and fly alignments, the BS + MNM null model provides a statistically significant likelihood increase for 22% of human genes and 57% of fly genes compared with the classic null model without MNMs. This test has a low rate of false positive inferences (Supplementary Table 3). In both datasets, the average estimated value of δ is about twice as high in the subset of BST-significant genes compared with BST-non-significant genes (Fig. 2a).

Next, we evaluated the empirical alignments for positive selection using the BS + MNM test, which incorporates MNMs into the null and positive selection model. We found that 94% of the tests on the human lineage that were significant using the classic BST lost significance (Fig. 2b and Supplementary Table 4). In flies, 38% of the tests lost significance, and a substantial fraction of the remaining genes were enriched in triple substitutions—a process not accounted for in our model (Fig. 2b and Supplementary Table 4).

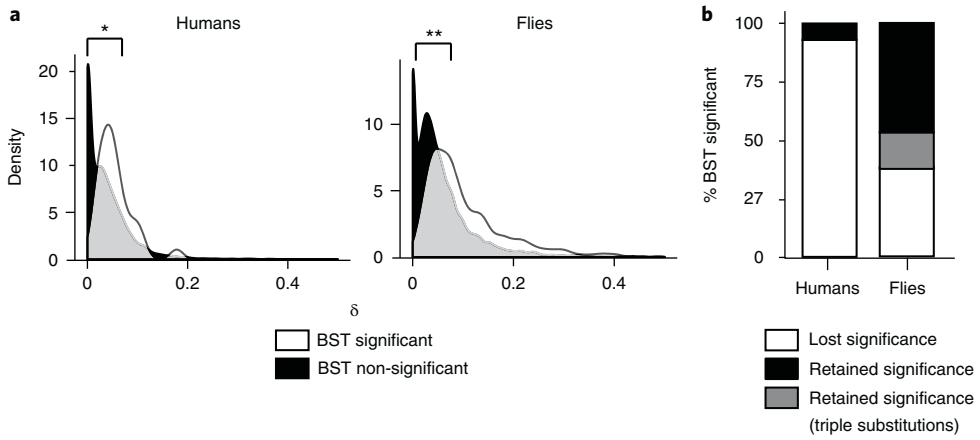


Fig. 2 | Incorporating MNMs into the branch-site model eliminates the signature of positive selection in many genes. The mammalian and fly datasets were reanalysed using a version of the BST that allows MNMs (BS + MNM) by including a parameter δ, a multiplier on the rate of each double substitution relative to single substitutions. **a**, The distribution of maximum likelihood estimates of δ across genes is shown, for genes that yield a significant (white) or non-significant (black) result in the classic BST. Median estimates of δ in BST-significant and BST-non-significant genes are 0.047 and 0.026 in humans, respectively, and 0.107 and 0.062 in flies. $^{*}P = 6.7 \times 10^{-4}$; $^{**}P = 1 \times 10^{-8}$ by Mann-Whitney *U*-test. **b**, Proportion of tests with a significant result in the BST that lose or retain that signature using the BS + MNM test. Genes with tests that remain significant but contain CMDs with three differences, which are not incorporated into the BS + MNM model, are also shown.

MNMs cause false positive inferences on a genome-wide scale. Our finding that incorporating MNMs eliminates the signature of positive selection from many genes could have several causes, including: (1) the BS + MNM model may have reduced power to identify authentic positive selection compared with the BST; (2) the BS + MNM model may ameliorate a false positive bias in the BST caused by MNMs; or (3) the additional parameter δ may allow the BS + MNM model to fortuitously fit other forms of sequence complexity, potentially reducing a bias in the BST caused by other model violations.

To evaluate these possibilities, we first analysed the test's power. We simulated sequences under the BST's positive selection model, using genome-wide average values for all parameters but varying the strength of positive selection (ω_2) and the proportion of sites under positive selection. We found that the BS + MNM test reliably detects strong positive selection ($\omega_2 > 20$) when it affects ~10% of sites in a typical gene, or moderate positive selection ($10 < \omega_2 < 20$) on a larger fraction of sites (Supplementary Fig. 4a,b). Its power is similar to that of the classic BST, with a slight reduction under only a few conditions on the fly lineage (Supplementary Fig. 4c). Thus, although some genes may have lost their signature of selection because of reduced power in the BS + MNM test, this is unlikely to be the primary cause of the dramatic reduction in the number of positive results when the test is used.

Next, we used simulations without positive selection to directly evaluate whether realistic rates of MNM increase the BST's propensity to deliver false positive inferences. For every gene in the mammalian and fly datasets, we simulated sequence evolution under the null BS + MNM model using parameters derived from the alignments, including δ, gene length and selection parameters. The fraction of substitutions that occurred at tandem sites on the branches of interest in the simulations (1.6% in humans and 3.2% in flies) was comparable to or slightly higher than the fraction of tandem substitutions phylogenetically inferred on these branches in the empirical alignments (1.3% in humans and 1.6% in flies), presumably because the BS + MNM model captures some but not all aspects of real sequence evolution (Supplementary Table 5). We then analysed the simulated alignments using the classic BST. In these experiments, every BST-positive result is false.

The number of genes yielding false positive results was greater than the number of genes that the BST inferred to be under positive selection using the empirical data (Fig. 3a). In flies, almost 9% of genes were falsely inferred to be under positive selection ($P < 0.05$), despite the conservative approach the method uses to calculate *P* values[6,16], compared with just 1% in control simulations without MNMs ($\delta = 0$). Over 1,700 of these false positive tests (an average of almost 300 genes per lineage) survived FDR adjustment, compared with a total of just 4 positive tests in the control simulations (Supplementary Table 2). In humans, the fraction of false positive inferences was lower, consistent with the test's reduced power in this dataset, but still about three times greater than in the control simulations. These false inferences were caused specifically by unincorporated MNMs—not some other form of model violation—because all other parameters were identical between the models used for simulation and analysis.

These findings indicate that MNMs under realistic evolutionary conditions produce a strong and widespread bias in the BST towards false inferences of positive selection. This bias is strong enough to cause the BST to make false inferences of positive selection at about the same rate as it infers selection in the real genomes of humans and flies.

Systematic bias caused by stochastic fixation of neutral MNMs. Only a few percent of mutations are MNMs, and most genes are only several hundred codons long, so on phylogenetic branches of short-to-moderate length many genes will evolve zero fixed MNMs. If neutral fixation of MNMs is a major cause of bias in the BST, a gene's propensity to produce a BST-significant result should depend on factors that increase the probability that it will contain one or more CMDs by chance, including its length and the gene-specific rate of MNM in that gene.

To evaluate this possibility, we first tested for an association between gene length and BST-positive results. As predicted, BST-significant genes were on average 100 and 16 codons longer than non-BST-significant genes in the human and fly datasets, respectively (Fig. 3b). A similar pattern was evident in the null simulations under empirically derived conditions (Supplementary Fig. 5); this finding cannot be attributed to an increase in the power to detect true positive selection in longer genes because no positive

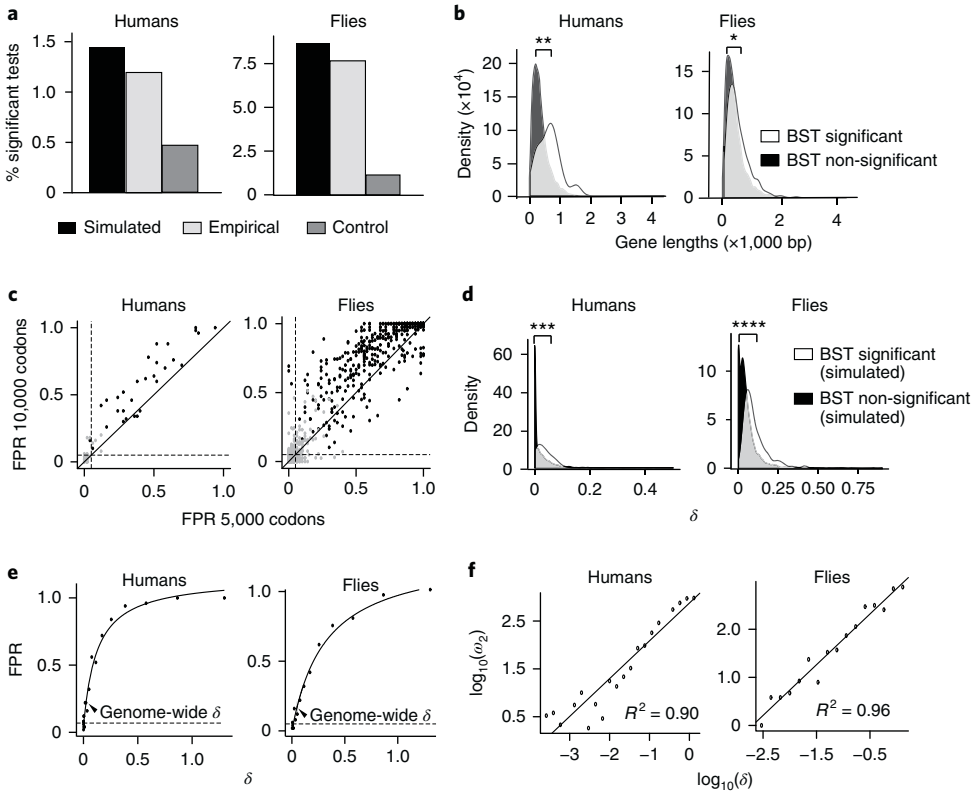


Fig. 3 | MNMs cause a strong bias in the BST under realistic conditions. For each gene in the mammalian and fly datasets, the parameters of the BS + MNM null model were estimated by maximum likelihood. Sequences of the original gene length were then simulated under these parameters and analysed using the classic BST. **a**, The fraction of all BST-significant tests ($P < 0.05$) is shown for the data simulated under the BS + MNM null model, the empirical data and a control dataset simulated with $\delta = 0$. **b**, BST-significant genes are longer than BST-non-significant genes. The probability density of gene lengths in the two categories is shown for the empirical datasets. Median lengths were 642 and 343 base pairs (bp) in humans, and 448 and 399 bp in flies, respectively. Mann-Whitney *U*-test for differences in the distributions: $^{*}P = 8 \times 10^{-4}$; $^{**}P = 8 \times 10^{-5}$. **c**, Systematic bias in the BST. For each empirical BST-significant gene, we simulated 50 replicate alignments using the BS + MNM null model parameters at lengths of 5,000 and 10,000 codons, then analysed them using the BST. The false positive rate (FPR) for any gene's simulation (black points) is the proportion of replicates that are significant ($P < 0.05$). Grey points represent the FPR for control simulations with $\delta = 0$. Dashed line, FPR of 0.05. The solid diagonal line has a slope of 1. **d**, The distribution of maximum likelihood estimates of δ across genes with (white) and without (black) a signature of positive selection in the classic BST is shown for data simulated under the BS + MNM null model. The median values of δ in BST-significant and BST-non-significant genes were 0.03 and 0.0009 in humans, and 0.08 and 0.04 in flies, respectively. Mann-Whitney *U*-test for the difference between distributions: $^{***}P = 1 \times 10^{-12}$; $^{****}P = 1 \times 10^{-199}$. **e**, Increasing the MNM rate exacerbates the bias in the BST. Sequences 5,000 codons long were simulated using the BS + MNM null model and the median value of each model parameter and branch length across all genes in each dataset, with variable δ. The FPR ($P < 0.05$) in 50 replicates at each value of δ is shown. The solid line represents the hyperbolic fit to the data. The dashed line shows FPR = 5%. The arrowhead points to the median value of δ across all genes. **f**, Relationship between δ and inferred ω_2. Sequences simulated in **e** were used to infer the positive selection parameter ω_2 under the BST-positive selection model. The best-fit linear regression line and coefficient of determination (R^2) are shown.

selection was present. To directly test the causal relationship between sequence length and false positive bias in the BST, we simulated multiple replicate alignments under the BS + MNM null model at increasing sequence lengths (L) using evolutionary parameters derived from each BST-significant gene (Supplementary Fig. 6). At $L = 5{,}000$ codons, 96% of human genes produced an unacceptable false positive rate (FPR > 0.05), with a median FPR of 0.39. Doubling the sequence length exacerbated the bias, with every gene now yielding an unacceptable FPR (median = 0.56; Fig. 3c). The same pattern was evident in flies, with even higher FPRs (median FPRs = 0.74 and 0.90 at $L = 5{,}000$ and 10,000, respectively). Control simulations under identical conditions but with $\delta = 0$ led to very low FPRs, even with very long sequences. Although these experiments involve lengths greater than those of most real genes, they establish that the probability that a gene will yield a false positive BST result is directly related to the target size it provides for chance fixation of MNMs.

Next, we evaluated whether the rate of MNM affects a gene's propensity to yield a positive result in the BST. As predicted, BST-significant genes in the empirical datasets had higher estimated δ than non-BST-significant genes (Fig. 2a). In the null simulations using the BS + MNM model under conditions derived from each empirical alignment, genes producing false positive BST results also tended to have higher δ (Fig. 3d). To directly test the causal relationship between the frequency of neutral MNMs and false positive bias in the BST, we simulated multiple replicate alignments under the BS + MNM null model with empirically derived parameters, but with variable δ. As δ increased, the rate of false positive inferences increased monotonically (Fig. 3e); the inferred value of the parameter ω_2, which represents the inferred intensity of positive selection in the model, also increased with δ (Fig. 3f).

We examined whether the branch-site unrestricted statistical test for episodic diversification (BUSTED)[2]—a recent method to identify episodic site-specific selection events across an entire tree—was also biased by MNMs. When sequences of length $L = 5{,}000$ were

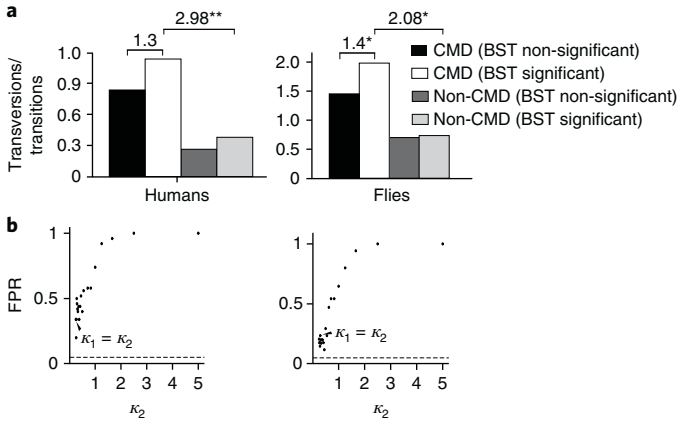


Fig. 4 | Transversion enrichment in CMDs biases the BST. a, Ratio of transversions to transitions observed in CMDs and non-CMDs for BST-significant and BST-non-significant genes. Fold-enrichment is shown above the columns as the odds ratio. $^{*}P = 5 \times 10^{-4}$; $^{**}P = 3 \times 10^{-25}$ by Fisher's exact test. **b**, Increasing the transversion rate in MNMs increases the bias of the BST. Sequences 10,000 codons long were simulated using an elaboration of the BS + MNM model that allows MNMs to have a transversion-to-transition rate (κ_2) different from that in single-nucleotide substitutions (κ_1). A total of 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, but 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 FPR when sequences were simulated with κ_2 equal to κ_1. The dashed line represents an FPR of 5%.

simulated under empirical conditions, BUSTED yielded an unacceptably high FPR for every gene in humans and most genes in flies (median FPRs = 0.29 and 0.50, respectively; Supplementary Fig. 7). In control simulations with $\delta = 0$, virtually no genes had a high rate of false positive inferences (FPR < 0.03). The biasing effect of neutral MNMs on tests of branch-specific positive selection is therefore not unique to the classic BST.

Taken together, these data indicate that MNMs under typical evolutionary conditions cause a strong and systemic bias in the BST and related tests. MNMs are rare, so whether any specific gene manifests the bias depends on factors that determine the probability of stochastic fixation of MNMs within it. Consistent with this conclusion, fewer genes are BST-positive on the very short human branch—on which substitutions are infrequent and CMDs even more rare—than on the fly phylogeny's longer branches. Many genes with BST-significant results may simply be those that happened to fix multinucleotide substitutions by chance.

Transversion enrichment in CMDs exacerbates bias in the BST. MNMs tend to produce more transversions than single-site mutational processes, so if CMDs are produced by MNMs, they should be transversion-rich[35,38,39]. As predicted, we found that the transversion-to-transition ratio is elevated in CMDs relative to non-CMDs by factors of three and two in mammals and flies, respectively (Fig. 4a). In the subset of BST-significant genes, CMDs have an even more elevated transversion-to-transition ratio, as expected if transversion-rich MNMs bias the test (Fig. 4a). These data are consistent with the hypothesis that a transversion-prone MNM process produced many of the CMDs in BST-significant genes, but it is also possible that positive selection could have enriched for transversions.

To directly test whether transversion enrichment in MNMs exacerbates the BST's bias, we developed an elaboration of the BS + MNM model in which an additional parameter allows MNMs to have a different transversion-to-transition-rate ratio (κ_2) compared with single-site substitutions (κ_1). We simulated sequence data using this model under empirically derived conditions without positive selection, varying the value of κ_2. We then analysed these data using the classic BST. We found that increasing κ_2 caused a rapid and monotonic increase in the FPR. The effect is strong: for example, when κ_2/κ_1 is increased from its baseline value of 1 to 2, the FPR approximately doubles (Fig. 4b). Thus, realistic rates of MNM generation and transversion enrichment together cause even stronger bias in the BST than MNMs alone.

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 BST and related tests. We hypothesized that CMDs implying multiple non-synonymous substitutions under standard models would provide the strongest support for the positive selection model. As predicted, we found that CMDs that imply more than one non-synonymous step are dramatically enriched in BST-significant genes (Fig. 5a). Furthermore, CMDs implying more non-synonymous single steps provided greater statistical support for the positive selection model (Fig. 5b). CMDs implying one non-synonymous and one synonymous step typically provide weak to moderate support, but a single CMD that implies two non-synonymous steps is often sufficient to yield a statistically significant signature of positive selection for an entire gene (Fig. 5b).

Discussion

This work establishes that the BST suffers from a strong and systematic bias towards false positive inferences. This bias is caused by a mismatch between the method's 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 MNM, which produces such codons in a single event. Under the BST's null model, the probability of two fixation events within a codon is extremely small, but it can increase dramatically when d_N/d_S exceeds one, as the positive selection model allows. Under realistic conditions, neutral fixation of just one or two MNMs in a gene is enough to yield a significant result in the BST.

As a result, CMDs in real sequences are the primary drivers of positive results by the BST. Virtually all statistical support for positive selection in the genome-scale alignments we studied comes from CMD-containing sites. These CMDs could have been produced by either neutral fixation of MNMs or positive selection, but the BST provides no reliable basis to distinguish between these possibilities.

The BST's bias is strong and pervasive under realistic, genome-scale conditions. Simulations without positive selection

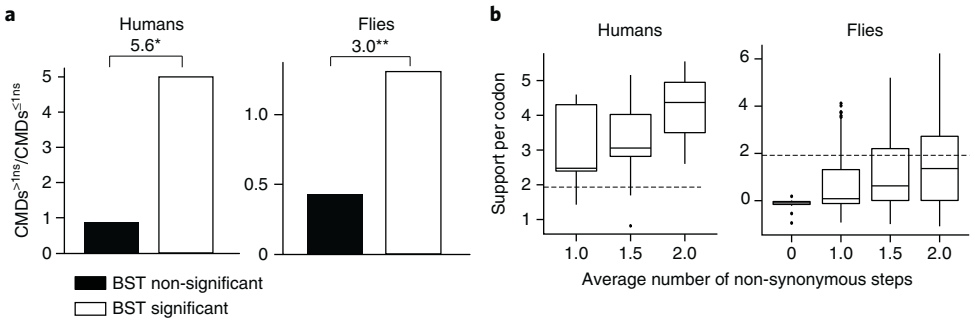


Fig. 5 | CMDs implying multiple non-synonymous steps drive the BST. a, For every CMD in every gene, the mean number of non-synonymous single-nucleotide steps on the two direct paths between the ancestral and derived sequence states on the branch of interest was calculated. In BST-significant and BST-non-significant genes, the ratio of CMDs invoking more than one non-synonymous step ($CMD^{>1ns}$) to those invoking one or fewer such steps ($CMD^{\leq 1ns}$) is shown. Fold-enrichment in BST-significant versus non-significant gene sets is shown above the columns as the odds ratio. $^{*}P = 9 \times 10^{-4}$; $^{**}P = 1.6 \times 10^{-67}$ by Fisher's exact test. **b**, Support for the positive selection model provided by CMDs depends on the number of implied non-synonymous single-nucleotide steps. Support is the log-likelihood difference between the positive selection and null models of the BST given the data at a single codon site. Box plots show the distribution of support by CMDs in BST-significant genes categorized according to the mean number of implied non-synonymous steps. The dashed line represents support of 1.92, which yields a significant result for an entire gene ($P < 0.05$). In human BST-significant genes, there are no CMDs that imply zero non-synonymous changes.

under empirically derived conditions showed that MNMs cause the BST to produce very frequent false positive inferences. In both the human and fly empirical datasets, the number of genes that the BST infers to be positively selected does not exceed the number expected to be produced by MNM-induced bias alone. Furthermore, these simulations did not include the elevated transversion rate that characterizes MNMs, which exacerbates the test's bias. Our results suggest that MNM-induced bias may explain many or most of the BST's inferences of positive selection in these datasets.

We do not contend that the BST is always wrong or that molecular adaptive evolution does not occur. Some of the CMDs in BST-significant genes may have evolved because positive selection fixed MNMs or several single-site mutations within a codon in serial. The test cannot distinguish between sequence data produced by these two scenarios, so it provides no reliable evidence that a gene evolved adaptively. It also cannot reliably estimate the fraction of genes in a large set that evolved under positive selection. There are numerous examples of strongly supported adaptive evolution—particularly involving host–parasite genetic conflicts—in which sequence signatures of positive selection are likely to be authentic[46–51], but the convincing evidence in these cases comes from sources other than the BST. The bias we discovered may help explain why some studies have found that codons with a high posterior probability of positive selection in the BST have no effect on putatively adaptive functions, whereas those that do confer those functions have low or moderate posterior probabilities[52–54].

Our results are likely to be generalizable. MNMs appear to be a property of replication processes in all eukaryotes, and the MNM rates that we observed in mammals and flies are in the same range as those identified in in a variety of eukaryotic species[25,34,37]. We observed strong bias on lineages with divergence levels ranging from very low (on the human terminal branch) to moderate (on the fly branches), so this problem does not appear to be unique to highly diverged sequences. The major factors determining whether a gene returns a false positive result in the BST test are those that affect the probability that one or more MNMs will be stochastically fixed (that is, gene length, MNM rate and overall substitution probability). We must therefore consider the possibility that some—and potentially many—of the thousands of genes previously reported to be under positive selection based on the BST could simply be those that happened by chance to neutrally fix one or more MNMs.

If the BST is so prone to error, what should researchers do? The BS + MNM test may be a promising means to accommodate MNMs, but there are many other forms of evolutionary complexity that are not incorporated into this model[55–57]. More work is therefore required before the BS + MNM test or related techniques[9] can be used with confidence. An alternative strategy—using functional experiments to explicitly test hypotheses about the genes and substitutions that drove molecular adaptation—can produce strongly supported inferences, but it is not clear how to implement time-consuming bench and field work on a genome-wide scale[50,58–60]. Future research may develop and validate more robust models to detect positive selection, and these may help to identify candidate genes for which specific hypotheses of past molecular adaptation on specific lineages can be formulated and tested. The primary method used for this purpose until now is unreliable.

Methods

Datasets, quality control and inference of BST-significant genes. We analysed 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)[12,14,45]. We retained only genes that did not have gross misalignments and that had complete coverage in all fly species or all primate species. We applied the BST as implemented in CODEML 4.7 to each alignment, assuming the phylogenetic relationships reported in the published studies (Supplementary Fig. 1)[12,14]. Branch lengths and model parameters were estimated for each alignment by maximum likelihood, using the F3X4 model for codon frequencies. We tested each gene in mammals for selection on the terminal branch leading to humans. In flies, each gene was tested separately for selection on each of the six terminal branches. We expressed the fraction of positive inferences across genes as the proportion of all tests conducted[6]. As is standard practice, we calculated *P* values using a likelihood ratio test with 1 d.f. (χ_1^2), which makes the test conservative under the null hypothesis[6]. Genes were initially identified as having a putative BST signature of selection at $P < 0.05$. We then applied a correction for multiple testing to an 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 at positions with alignment gaps. We also removed genes for which the maximum likelihood ancestral reconstructions reported by CODEML at the base of the tested branch differed between the null and positive selection models, yielding a set of genes with CMDs that do not depend on which model is chosen. In flies, 443 genes were retained after these filters and constitute the BST-significant set of genes from this dataset. These genes produce 458 positive tests in the BST, because a gene can yield a significant test on more than one branch. No

genes on the human lineage were significant after FDR correction, so we retained as the BST-significant set from this dataset those genes that passed the ancestral reconstruction filter and had $P<0.05$ (Supplementary Table 2). The BST-non-significant set of genes comprises all genes that pass the alignment and ancestral reconstruction filter that are not in the BST-significant set (humans: $n=6{,}757$; flies: $n=6{,}883$). We also repeated our analysis of CMD enrichment (see below) using a gene set that had not been filtered for reconstruction consistency; our conclusions were unchanged (Supplementary Table 6).

Support for positive selection. CMDs were identified in BST-significant and BST-non-significant genes as codons with two or three observed nucleotide differences between the maximum likelihood states at the ancestral and extant nodes for the branch being tested. Non-CMDs are codons with zero or one difference on the branch tested. CMDs were not assessed on branches not tested.

To determine the role of CMDs in significant results from the BST, we excluded codon positions in BST-significant genes containing CMDs, reanalysed the data using the BST, 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 BST-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 ($\text{support}_{\text{CMD}}$) is the support summed over all CMD sites in the alignment. The proportion of support provided by CMDs is $\text{support}_{\text{CMD}}$ / ($\text{support}_{\text{CMD}}+\text{support}_{\text{non-CMD}}$). This proportion can be greater than one 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 maximum likelihood over all sites.

Sites were classified a posteriori as under positive selection if their Bayes empirical Bayes posterior probability of being in class 2 ($\omega_2>1$) 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 non-synonymous 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 non-synonymous or synonymous using the standard genetic code and then calculated the mean number of non-synonymous 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 BST is based on the Goldman–Yang model[5]. Sequence evolution is modelled 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 change: synonymous transitions and transversions, and non-synonymous transitions and transversions (see q_{ij}, equation (1)). Three parameters are estimated from the data by maximum likelihood: ω; the ratio of the 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-to-transition-rate ratio.

$$q_{ij}=\begin{cases}\kappa\pi_j & \text{synonymous transversion}\\ \pi_j & \text{synonymous transition}\\ \kappa\omega\pi_j & \text{non-synonymous transversion}\\ \omega\pi_j & \text{non-synonymous transition}\\ 0 & \text{two or more differences}\end{cases} \quad (1)$$

Element q_{ij} is zero for substitutions involving more than one difference, so CMDs can only evolve through intermediate codon states. A scaling factor applied to the matrix ensures that the expected number of substitutions per codon equals the branch length. The BST null model is a mixture of submodels, each based on the Goldman–Yang model, in which ω is constrained to values less than or equal to one on all branches; the BST positive selection model is a mixture of Goldman–Yang submodels in which one mixture component allows ω to exceed one on a specified set of foreground branches.

We developed a modification of the Goldman–Yang model that incorporates MNMs using the parameter δ, which scales the instantaneous rate of a double-nucleotide substitution relative to that of a single-nucleotide substitution (see q_{ij}. equation (2)). When $\delta=0$, the BS + MNM model reduces to the classic BST model that does not incorporate MNMs (q_{ij}, equation (1)). Triple substitutions have an instantaneous rate of zero. Double- and single-nucleotide substitutions differ in the total number of codon pairs in each category, the number of non-synonymous codon pairs, and the average number of transversions across codon pairs, so δ does not represent the expected ratio of the total number of double-nucleotide to single-nucleotide substitutions. As in the Goldman–Yang model, the matrix is scaled given the parameter values used so that the expected total number of substitution events on a branch equals the branch length.

$$q_{ij}=\begin{cases}\kappa\pi_j & \text{synonymous transversion}\\ \pi_j & \text{synonymous transition}\\ \kappa\omega\pi_j & \text{non-synonymous transversion}\\ \omega\pi_j & \text{non-synonymous transition}\\ \omega\delta\kappa^2\pi_j & \text{non-synonymous, 2 transversions}\\ \omega\delta\pi_j & \text{non-synonymous, 2 transitions}\\ \omega\delta\kappa\pi_j & \text{non-synonymous, 1 transversion, 1 transition}\\ \delta\pi_j & \text{synonymous, 2 transitions}\\ \delta\kappa^2\pi_j & \text{synonymous, 2 transversions}\\ \delta\kappa\pi_j & \text{synonymous, 1 transversion, 1 transition}\\ 0 & \text{otherwise}\end{cases} \quad (2)$$

The BS + MNM test of positive selection is identical to the BST, but it uses the BS + MNM codon model for both the null and positive selection models. We implemented this test by modifying the BST batch file (YangNielsenBranchSite2005.bf) in HyPhy 2.2.6 software (https://github.com/veg/hyphy) by declaring δ a global variable, incorporating it into the codon table and allowing it to be optimized by maximum likelihood.

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 maximum likelihood 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 with the values used to generate the sequences (Supplementary Fig. 3).

To test whether there is statistical support in the empirical sequence data for the BS + MNM null model relative to the standard BST null model, we performed a likelihood ratio test with one d.f., comparing the fit of the BS + MNM null model and the BST null model. For each of the 6,868 human genes, we tested whether the BS + MNM null model fit the data better than the BST null model at $P<0.05$ and also applied an adjustment for multiple testing (FDR < 0.2). We performed similar likelihood ratio tests for all fly alignments on each of the six terminal lineages. To determine whether this test might be prone to falsely inferring support for the BS + MNM model, we simulated control sequences under the null BST model with parameters derived from the empirical sequences and performed the likelihood ratio test as described above. Only 2% of these simulated genes in humans and 2.6% of genes in flies yielded significant support for BS + MNM at $P<0.05$. Zero human genes and 0.006% of fly genes retained significance after multiple testing adjustment (FDR < 0.2) (Supplementary Table 3).

Simulations and analysis of false positive bias. To characterize bias in the BST 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 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 the 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 6,868 and 8,564 mammalian and fly alignments, each at its empirical length, using the BS + MNM null model and the maximum likelihood parameter estimates inferred for that gene from the empirical data. We then ran the BST on these sequences, testing for signatures of positive selection on the human lineage and each terminal fly lineage (Supplementary Table 2). Control simulations were conducted under identical conditions, but with $\delta=0$.

To test the effect of gene length on bias in the BST, we focused on genes in the BST-significant set. For each gene's gene-specific parameters, we simulated 50 replicate alignments of length 5,000 or 10,000 codons. We analysed these alignments using the BST, 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 analysed. The 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 $\delta=0$ in the simulations.

To test how the rate at which MNM substitutions are produced affects false positive inference rates of the BST, we simulated the 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 analysed each replicate using the BST for selection on the human or *Drosophila simulans* lineages and calculated the proportion of replicates for each value of δ that yielded a false positive inference ($P<0.05$).

We computed the observed proportion of tandem substitutions as a fraction of all substitutions on the human and *D. melanogaster* lineages in both empirical and simulated datasets. For each of the 6,868 genes in the curated mammalian dataset, we aligned the human gene to the phylogenetically inferred sequence of the human–chimp ancestor, identified all substitutions as differences between these sequences, and calculated the proportion of tandem substitutions as the number of substitutions at adjacent sites divided by the sum of substitutions at adjacent sites and those at non-adjacent sites across all sites in the dataset. Substitutions at adjacent sites were counted as a single tandem substitution. For each of the 8,564 genes in the fly dataset, we aligned the *D. melanogaster* sequence to the *D. melanogaster–D. simulans* ancestor and followed the procedure described above. For simulated sequences, we repeated this procedure using the corresponding ancestral and terminal sequences simulated under the BS + MNM null model and parameters estimated from each gene in the empirical datasets, including δ.

BUSTED. To examine the accuracy of BUSTED, we used HyPhy software 2.2.6 (batch files BUSTED.bf and QuickSelectionDetection.bf). We analysed the 5,000-codon-long alignments simulated under the BS + MNM null model using parameters estimated by maximum likelihood for each BST-significant gene, with δ assigned to either its gene-specific estimate, its genome average or 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 BST of the empirical data.

Power analyses. To characterize the statistical power of the BST and BS + MNM tests, we simulated sequence evolution with positive selection of variable intensity and pervasiveness (Supplementary Fig. 4). Specifically, we used the BST positive selection model in HyPhy to simulate sequence evolution with the human and *D. simulans* terminal branches as the foreground branches. We used genome-wide 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, parameters that represent the intensity of positive selection and the proportion of sites positively selected class, respectively. A total of 20 replicate alignments were simulated under each set of conditions and then analysed using the BST, the BS + MNM test or BUSTED. 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 BST and BS + MNM; FDR < 0.20 for at least one site in the alignment for BUSTED).

BS + MNM + κ_2 model. We developed the BS + MNM + κ_2 model, which incorporates into the BS + MNM model (q_{ij}, equation (2)) two different transversion-to-transition-rate ratio parameters—κ_1 for single-site substitutions and κ_2 for MNMs (see q_{ij}, equation (3)). All free parameters of the model are estimated by maximum likelihood given a sequence alignment. This model was implemented by further modifying our BS + MNM batch file in the 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 maximum likelihood as other parameters are in the batch file.

$$q_{ij} = \begin{cases} \kappa_1\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \omega\kappa_1\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ \omega\delta\kappa_2^2\pi_j & \text{non-synonymous, 2 transversions} \\ \omega\delta\pi_j & \text{non-synonymous, 2 transitions} \\ \omega\delta\kappa_2\pi_j & \text{non-synonymous, 1 transversion, 1 transition} \\ \delta\pi_j & \text{synonymous, 2 transitions} \\ \delta\kappa_2^2\pi_j & \text{synonymous, 2 transversions} \\ \delta\kappa_2\pi_j & \text{synonymous, 1 transversion, 1 transition} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

For validation, we estimated the parameters of the BS + MNM + κ_2 null model by maximum likelihood for every alignment in each dataset and calculated the genome-average median estimate of each parameter (Supplementary Fig. 8). We then simulated 50 replicate alignments of length 418 and 510 codons in the mammalian and fly datasets respectively, using the BS + MNM + κ_2 null model with all model parameters set to their genome-wide median values. For each replicate, we estimated each parameter by maximum likelihood under the null model given each alignment. We then compared the distribution of estimates with the parameters used to generate the alignments. We found that most parameters were estimated accurately, but estimates of κ_2 had high variance (Supplementary Fig. 8a,b), presumably because the data in a single gene, in which CMDs are typically rare, is inadequate to support a robust estimate of this parameter. We therefore limited our use of this model to generating sequences by simulation rather than making inferences from sequence data.

To determine the effect of the MNM-specific transversion-to-transition rate on false positive bias in the BST, we simulated sequences under the BS + MNM + κ_2 null model using genome-median parameters except κ_2, which we varied. Sequences 10,000 codons long were used because simulating shorter sequences resulted in high variance in the realized transversion-to-transition ratio. For each value of κ_2, we simulated 50 replicates, applied the BST and calculated the FPR as the fraction of replicates yielding a positive inference ($P < 0.05$).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

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.

Data availability. The empirical alignments reanalysed in this study are available in the supplementary information files of the original publications that generated and analysed these data[12,14,45].

Received: 26 July 2017; Accepted: 18 May 2018;
Published online: 2 July 2018

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. Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
3. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
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. 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).
12. Drosophila 12 Genomes Consortium et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218 (2007).
13. Foote, A. D. et al. Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
14. Kosiol, C. et al. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).
15. Roux, J. et al. Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* **31**, 1661–1685 (2014).
16. Yang, Z. & dos Reis, M. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.* **28**, 1217–1228 (2011).
17. Zhang, J. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* **16**, 868–875 (1999).
18. 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).
19. 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).
20. Nozawa, M., Suzuki, Y. & Nei, M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc. Natl Acad. Sci. USA* **106**, 6700–6705 (2009).
21. 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).
22. 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).

23. Kosakovsky Pond, S. L. et al. A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* **28**, 3033–3043 (2011).
24. Zhang, J. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol. Biol. Evol.* **21**, 1332–1339 (2004).
25. Schrider, D. R., Hourmozdi, J. N. & Hahn, M. W. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* **21**, 1051–1054 (2011).
26. Saribasak, H. et al. DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J. Exp. Med.* **209**, 1075–1081 (2012).
27. Loeb, L. A. & Monnat, R. J. DNA polymerases and human disease. *Nat. Rev. Genet.* **9**, 594–604 (2008).
28. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. Low fidelity DNA synthesis by human DNA polymerase-η. *Nature* **404**, 1011–1013 (2000).
29. 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).
30. 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).
31. 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).
32. Besenbacher, S. et al. Multi-nucleotide de novo mutations in humans. *PLoS Genet.* **12**, e1006315 (2016).
33. 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).
34. 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).
35. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **24**, 1445–1454 (2014).
36. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).
37. Assaf, Z. J., Tilk, S., Park, J., Siegal, M. L. & Petrov, D. A. Deep sequencing of natural and experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new mutations. *Genome Res.* **27**, 1988–2000 (2017).
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. 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).
41. 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).
42. Rogozin, I. B. et al. Evolutionary switches between two serine codon sets are driven by selection. *Proc. Natl Acad. Sci. USA* **113**, 13109–13113 (2016).
43. De Maio, N., Holmes, I., Schlötterer, C. & Kosiol, C. Estimating empirical codon hidden Markov models. *Mol. Biol. Evol.* **30**, 725–736 (2013).
44. Suzuki, Y. False-positive results obtained from the branch-site test of positive selection. *Genes Genet. Syst.* **83**, 331–338 (2008).
45. Larracuente, A. M. et al. Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* **24**, 114–123 (2008).
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).
47. Elde, N. C., Child, S. J., Geballe, A. P. & Malik, H. S. Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature* **457**, 485–489 (2009).
48. Patel, M. R., Loo, Y. M., Horner, S. M., Gale, M. & Malik, H. S. Convergent evolution of escape from hepaciviral antagonism in primates. *PLoS Biol.* **10**, e1001282 (2012).
49. Demogines, A., Abraham, J., Choe, H., Farzan, M. & Sawyer, S. L. Dual host–virus arms races shape an essential housekeeping protein. *PLoS Biol.* **11**, e1001571 (2013).
50. Barber, M. F. & Elde, N. C. Nutritional immunity. Escape from bacterial iron piracy through rapid evolution of transferrin. *Science* **346**, 1362–1366 (2014).
51. Machkovech, H. M., Bedford, T., Suchard, M. A. & Bloom, J. D. Positive selection in CD8$^+$ T-cell epitopes of influenza virus nucleoprotein revealed by a comparative analysis of human and swine viral lineages. *J. Virol.* **89**, 11275–11283 (2015).
52. Field, S. F., Bulina, M. Y., Kelmanson, I. V., Bielawski, J. P. & Matz, M. V. Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J. Mol. Evol.* **62**, 332–339 (2006).
53. Yokoyama, S., Tada, T., Zhang, H. & Britt, L. Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proc. Natl Acad. Sci. USA* **105**, 13480–13485 (2008).
54. Zhuang, H., Chien, M. S. & Matsunami, H. Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. *Proc. Natl Acad. Sci. USA* **106**, 21247–21251 (2009).
55. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31**, 1956–1978 (2014).
56. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).
57. Pond, S. K. & Muse, S. V. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22**, 2375–2385 (2005).
58. Chan, Y. F. et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302–305 (2010).
59. Barrett, R. D. & Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* **12**, 767–780 (2011).
60. Siddiq, M. A., Loehlin, D. W., Montooth, K. L. & Thornton, J. W. Experimental test and refutation of a classic case of molecular adaptation in *Drosophila melanogaster*. *Nat. Ecol. Evol.* **1**, 0025 (2017).

Acknowledgements

We are grateful to the members of the Thornton laboratory for discussion and helpful comments. We thank the Beagle2, Midway2 and Tarbell supercomputing clusters at the University of Chicago. We also thank the developers of HyPhy for presenting an open source platform that allows customization of standard analyses. Funding was provided by NIH R01GM104397 and R01GM121931 (to J.W.T.), NSF DEB-1601781 (to J.W.T. and A.V.), NSF DBI-1564611 (to M.W.H.), and the Precision Health Initiative of Indiana University (to M.W.H.).

Author contributions

The analyses were designed by all authors, performed by A.V. and interpreted by all authors. The manuscript was written by A.V. and J.W.T. with contributions from M.W.H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0584-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.W.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): JOSEPH W THORNTON

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We did not collect any new data in this study. We analyzed two comprehensive, genome-wide protein coding alignment datasets that have been previously published. i) The mammalian dataset consists of 16541 genes with orthologs across six eutherian mammals, and ii) The fly dataset consists of 8563 genes across each of the six species in the melanogaster subgroup clade.
Data analysis	We used PAML 4.7 (CODEML binary) to run the branch-site test on the previously published alignment datasets. HYPHY 2.2.6 was used to test newer methods such as BUSTED using the batch file BUSTED.bf. We modified the Yang-Nielsen branch-site batch file in HYPHY (YangNielsen_BranchSite2005.bf) to incorporate multinucleotide mutations using the delta parameter (BS+MNM model). An additional parameter describing a different transversion:transition rate in MNM codons, kappa2, was then a further modification of the BS+MNM model (BS+MNM+k2). All simulations that tested the effect of delta or kappa2 on inferences of positive selection were then conducted using the BS+MNM or BS+MNM+k2 models. Our custom HYPHY batch files are available for download at https://github.com/JoeThorntonLab/MNM_SelectionTests.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data analyzed in this study has been published previously, and our manuscript cites these studies from which the raw data can be obtained.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed two comprehensive, genome-wide protein coding alignment datasets that have been previously published -- i) the mammalian dataset consists of 16541 genes with orthologs across six eutherian mammals, and ii) the fly dataset consists of 8563 genes across six species in the melanogaster subgroup clade.
Data exclusions	We applied stringent quality control metrics, focusing on alignment quality and ancestral codon reconstructions. We excluded alignments that did not pass these metrics. All quality control procedures, including the sample size at every stage, are described in the methods and supplementary results.
Replication	Estimates of false positive rate of the branch-site and related tests under defined conditions were made by performing 50 replicate simulations for each gene under each condition. For validation of new models, 50 replicates under genome-wide average conditions were used. For power analyses, we simulated and analyzed 20 replicate alignments for each set of conditions. To determine the genome-wide number of false positive inferences produced from null-simulated data, we simulated one replicate alignment for each of the one replicate of each of the 6868 mammalian and 8564 fly genes. All replication procedures are described in the text.
Randomization	Randomization is not relevant to this study.
Blinding	There was no blinding in this study.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging