Berrio et al. BMC Genomics (2020) 21:359
https://doi.org/10.1186/s12864-020-6752-4

BMC Genomics

METHODOLOGY ARTICLE Open Access

Identifying branch-specific positive selection throughout the regulatory genome using an appropriate proxy neutral

Alejandro Berrio[1*], Ralph Haygood[2] and Gregory A. Wray[1]

Abstract

Background: Adaptive changes in *cis*-regulatory elements are an essential component of evolution by natural selection. Identifying adaptive and functional noncoding DNA elements throughout the genome is therefore crucial for understanding the relationship between phenotype and genotype.

Results: We used ENCODE annotations to identify appropriate proxy neutral sequences and demonstrate that the conservativeness of the test can be modulated during the filtration of reference alignments. We applied the method to noncoding Human Accelerated Elements as well as open chromatin elements previously identified in 125 human tissues and cell lines to demonstrate its utility. Then, we evaluated the impact of query region length, proxy neutral sequence length, and branch count on test sensitivity and specificity. We found that the length of the query alignment can vary between 150 bp and 1 kb without affecting the estimation of selection, while for the reference alignment, we found that a length of 3 kb is adequate for proper testing. We also simulated sequence alignments under different classes of evolution and validated our ability to distinguish positive selection from relaxation of constraint and neutral evolution. Finally, we re-confirmed that a quarter of all non-coding Human Accelerated Elements are evolving by positive selection.

Conclusion: Here, we introduce a method we called *adaptiPhy*, which adds significant improvements to our earlier method that tests for branch-specific directional selection in noncoding sequences. The motivation for these improvements is to provide a more sensitive and better targeted characterization of directional selection and neutral evolution across the genome.

Keywords: Adaptation, Positive selection, Analytical method, *adaptiPhy*, Proxy, Neutral

Background

An accurate and comprehensive characterization of the genomic distribution of adaptive substitutions is essential for understanding the genetic basis for trait divergence between species [1–5]. Tests for positive selection at the interspecies scale developed during the 1980s focused on ω, the ratio of nonsynonymous to synonymous substitution rates in protein coding regions [6, 7]. These methods were first applied at a whole genome scale soon after the release of reference genome assemblies for human, chimpanzee, and macaque [8–11], and provided the earliest relatively unbiased views of positive selection on protein-coding regions. At the same time, a growing appreciation for the contribution of regulatory mutations to adaptation [12–14] prompted the development of methods to test for positive selection in noncoding regions.

Two general approaches were devised to test for positive selection in the absence of a genetic code. One seeks regions that contain many substitutions along the

* Correspondence: alebesc@gmail.com
[1]Department of Biology, Duke University, Biological Sciences Building, 124 Science Drive, Durham, NC 27708, USA
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

human branch (since the most recent common ancestor with chimpanzees) but are otherwise highly conserved among mammals or vertebrates [15–19]. The other seeks an elevated rate of substitution along the human lineage in a query region hypothesized to contain regulatory elements relative to a nearby reference (proxy neutral) region thought to contain few functional elements [20, 21]. Both approaches test for branch-specific accelerated substitution, but differ in the reference point against which they assess acceleration: the first tests for accelerated substitution within otherwise conserved regions against a putatively neutral region that is usually obtained from local Ancient Repeats (ARs) or fourfold degenerate sites (4D) [15, 17–19], while the second employs a putatively non-functional local intron of the genome as a neutral reference against which to identify branch-specific accelerated substitution [20, 21]. Wong & Nielsen [20] defined the parameter ζ as the ratio of substitution rates in the query region to those in the associated neutral region; ζ is thus analogous to ω. To detect significance in the departures from neutrality, both approaches typically use maximum likelihood estimation and likelihood ratio tests (LRTs) that compare a null model allowing neutrality against an alternative model that additionally allows for positive selection.

These two general approaches have complementary strengths and weaknesses. The first approach is less sensitive, in that it does not make use of an appropriate proxy for neutral regions along the human lineage. This approach allowed the discovery of Human Accelerated Regions (HARs) [16]. However, there is no reason to suppose adaptive evolution along the human lineage has been confined to regions under purifying selection in most or all other species, particularly since such regions constitute a small fraction of the genome (Fig. 1a). Additionally, this method may fail to distinguish regions evolving under relaxation of constraint from those that have experienced positive selection. The first method has been implemented in *phyloP* [22], which is straightforward to execute and allows running selection tests using different approaches, such as LRT, SPH, Score and Genomic Evolutionary Rate Profiling (GERP) [16, 23–25]. *phyloP* has been extensively used for more than a decade, and has been applied to conserved DNA regions using neutral proxies based on four-fold degenerate (4D) sites e.g., [26] or local ancient repeats (ARs) e.g., [27, 28]. The second method runs in *HyPhy* [29] and requires more computational and analytical effort than *phyloP*. On the other hand, it is more broadly applicable because it can query any genomic region regardless of whether that region was previously under functional constraint.

At the time these approaches were first developed, very little information existed about the location of functional elements in any genome. This limited the ability to identify suitable proxy neutral regions, i.e., those likely to be free from either purifying or positive selection. Inadvertently using constrained or accelerated regions as neutral proxies can potentially introduce artificial adaptive signals or reduce sensitivity, respectively. In addition, not knowing the location of regulatory elements meant that testing for positive selection at a genome-wide scale was intractable due to the need for massive correction for multiple testing. Prior to the invention of functional genomic assays for chromatin status, the best method for identifying putative regulatory elements was sequence conservation [15].

The ENCODE project [30, 31] and other efforts [32] to identify regulatory elements throughout the human genome mean that it is now possible to focus tests for positive selection on likely functional noncoding elements and to identify appropriate proxy neutral regions. Highly conserved noncoding regions overlap with only about 1.09% of the ~ 3 million known DNase I Hypersensitive Sites (DHSs) in the human genome [33] and just 1.06% of ~ 1.7 million enhancers and promoters from the HoneyBadger2 regions published by the ENCODE and Roadmap Epigenomics projects (Fig. 1a). Accordingly, it seems likely that a substantial fraction of functional DNA elements do not occur in regions of strong conservation. At the same time, ~ 18.3% of the human genome currently has no known regulatory or functional annotation, despite extensive study, providing a principled basis for choosing proxy neutral sequences that may be superior to 4D sites and ARs.

Here we introduce *adaptiPhy*, an improved analytical method that can test for branch-specific directional selection on any collection of query segments based on accurate alignments from three or more species. We implemented a series of technical and computational modifications to our previously published method [21] using openly available software from PHAST [22, 34] and functional genomic datasets from ENCODE [35]. We tested the performance of *adaptiPhy* in regions that have been already tested (i.e., published Human Accelerated Regions, or HARs) and among simulated sequences evolving at neutral rates, positive selection in one branch of the tree, or two branches of the tree, and relaxation of constraint. Significant improvements include: 1) better genome-wide representation of putatively neutral proxies based on functional annotations; 2) ability to test for selection in nearly any noncoding region of the genome, including functionally dense regions; 3) the ability to increase stringency by filtering reference sequences; and 4) improved understanding of how branch number and the size of query and reference regions impact test sensitivity. We demonstrate that this approach can be applied productively to focal collections of genomic regions commonly encountered in contemporary genomics and

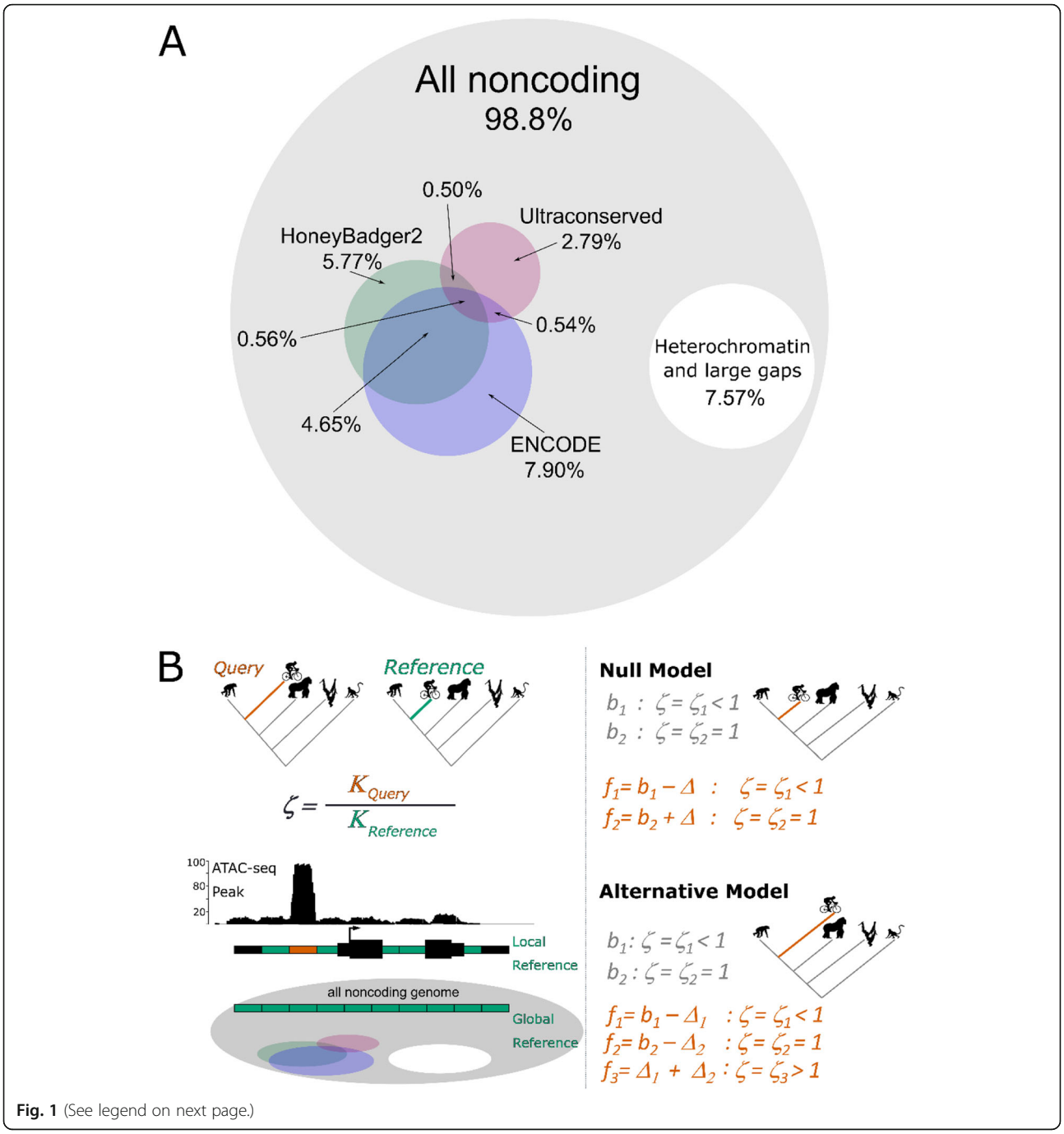

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Overlaps between functional regions of the human genome and the evolutionary model. a. *Left*, Venn diagram showing percentage of the human genome that overlaps with known non-coding functional annotations and conserved regions of the human genome. *Right*, scaled Venn diagram showing the proportion of conserved regions with respect to known functional annotations and gapped DNA representing telomeric and centromeric sequences (white). **b.** Graphic summary of our improved method. *Left panel*, the evolutionary ratio "ζ" is computed as the ratio of the substitution rate in a query (K_{query}) with respect to the substitution rate in a reference ($K_{reference}$) region. Queries can be obtained from functional annotations such as ATAC-seq or ChIP-seq peaks (red box), while reference alignments can either be taken by sampling local non-functional elements in the vicinity of the query or from a genome-wide random sampling of non-functional and putatively neutral regions of the genome (green boxes). *Right panel,* to test for positive selection in a query region on a foreground branch (red) of the tree, we fit, via maximum likelihood, both a null model and an alternative model to the alignments of the reference and query regions. In both models, on all branches, all sites in the reference region evolve neutrally. In both models, on the background branches, a fraction b1 > = 0 of sites in the query region evolve under purifying selection, at rate $\zeta_1 < 1$ relative to sites in the reference region, and a fraction b2 = 1 - b1 of sites in the query region evolve neutrally, at relative rate $\zeta_2 = 1$. In the null model, evolution on the foreground branch is the same as on the background branches, except a fraction $\Delta >= 0$ of sites in the query region that evolve under purifying selection on the background branches may evolve neutrally on the foreground branch, that is, the model allows for relaxation of constraint on the foreground branch. In the alternative model, fractions $\Delta_1 >= 0$ and $\Delta_2 >= 0$ of sites in the query region that evolve under purifying selection and neutrally, respectively, on the background branches may evolve under positive selection on the foreground branch, at rate $\zeta_3 > 1$. A likelihood-ratio test indicates whether the alternative model fits the alignments significantly better than the null model. As explained under "Materials and methods", we conservatively approximate this test as a chi-squared test with one degree of freedom

genetics research, such as the open chromatin landscape of a specific cell type or trait-associated regions from a genome-wide association study.

Results

Global nonfunctional sequences provide appropriate neutral reference sequences

To identify appropriate proxy neutral regions, we began by identifying all putatively non-functional regions (NFRs) of length 300 bp (similar in length to many regulatory elements) throughout the human genome. NFRs are regions devoid of any coding sequence, noncoding RNA, open chromatin region, ChIP peak, or other functional unit; we also masked repeats (see Methods for inclusion criteria). We then tallied the number of NFRs located within 10, 40 and 100 kb of a set of 1000 random DHS sites, non-coding Human Accelerated Elements (ncHAE), and a control subset of "global" NFRs from throughout the genome. For the longest region (i.e. 100 kb), on average, there are only 7.8 local NFRs per DHS, 28.8 NFRs per ncHAE, and 64.4 NFRs per global NFR (Fig. 2a). Moreover, 58.5% of DHSs and 43.3% of ncHAEs had *no* local NFRs within 100 kb. Thus, the number of local NFRs that can be used as neutral reference regions is often insufficient for extensive testing of positive selection. Some previous studies used ancient repeats (ARs) as neutral proxies [e.g.,28].We found that on average, there are only 3.3 ARs within 100 kb per DHS, and 44% of DHS regions had no AR within 100 kb. Thus, identifying sufficient ARs to use as a local reference for each DHS is also difficult.

Next, we asked whether local ARs, local NFRs, or global NFRs can be used to build an appropriate reference for testing positive selection. To build a local reference region, we concatenated all the NFRs or ARs within 100 kb of a given query region. Next, we computed the substitution rate of each concatenated sequence of local ARs, local NFRs, and NFRs across the genome. For a set of query regions in our sample, we found a wide distribution of substitution rates among concatenated local references including local NFRs, global NFRs, and ARs (Fig. 2b).

We thus sought to test whether filtering global NFRs by their relative substitution rate over the entire tree can provide an improved neutral proxy for estimating positive selection. We filtered out global NFRs representing the top and bottom quartiles of relative substitution rates (Fig. S1), then concatenated 10 NFRs per query. When we compared the substitution rates of this new set of putative neutral references, we found that the distribution of substitution rates of the filtered global NFRs is narrower and its median is skewed to the right (Fig. 2c). This suggests that global NFRs can provide a set of putatively neutral elements if appropriately filtered. This approach also allows conservativeness and sensitivity to be modulated by tuning the filtration step accordingly. Moreover, given that we use relative values of substitution rate, this filtering step can be applied to any region of the genome regardless of the amount of functional annotation.

To assess the impact of using local versus global NFRs on testing for positive selection, we sampled three sets of queries: (1) widespread and specific DHSs (open in > 124 and exactly 1 ENCODE cell types, respectively); (2) a set of ncHAEs to be used as positive controls; and (3) a set of putatively non-functional DNA elements to be used as negative controls. The correlation in *P*-values is high among the 3531 DHSs that could be analyzed using both local and global neutral proxies (Spearman's Rank test $\rho = 0.80$; $P < 2.2 \times 10^{-16}$; Fig. 3a). Of these, only 2.63% scored high for positive selection ($P < 0.05$) for global proxies, while 5.12% scored high for positive using the local proxy alone. Likewise, the correlation of *P*-

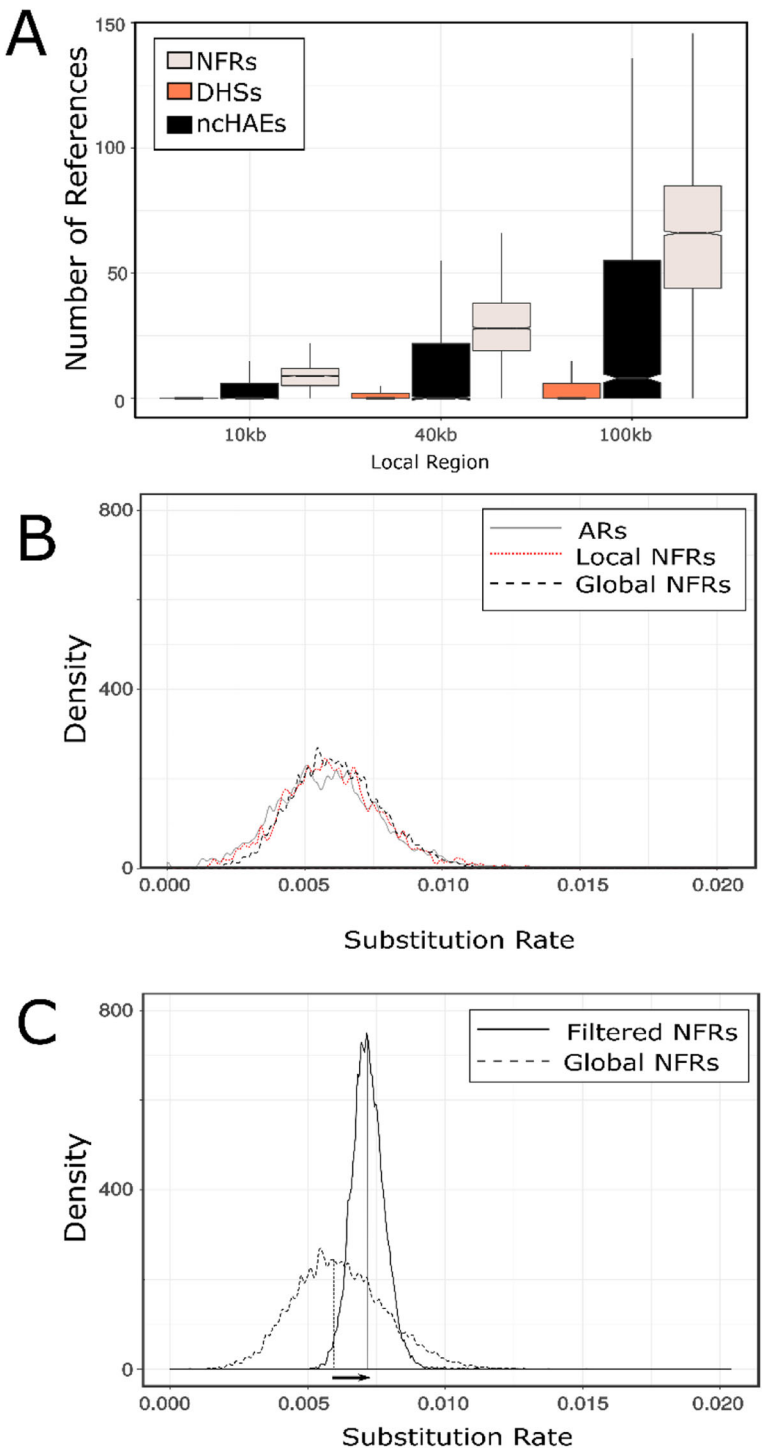

Fig. 2 Finding a neutral proxy. **a**. Distribution of the number of local reference alignments around each DNA element within three different distances: 10 kb, 40 kb and 100 kb. **b.** Density distribution of relative substitution rates among concatenated Ancestral Repeats (ARs) around 100 kb of each DHS element in our list, concatenated local NFRs around each DHS, concatenated global NFRs before filtering out trees with low and high substitution rates. **c.** Density distribution of relative substitution rates in the concatenated global NFRs before and after filtration step. The arrow depicts the change in the median distribution of substitution rate of global reference alignments before versus after filtering

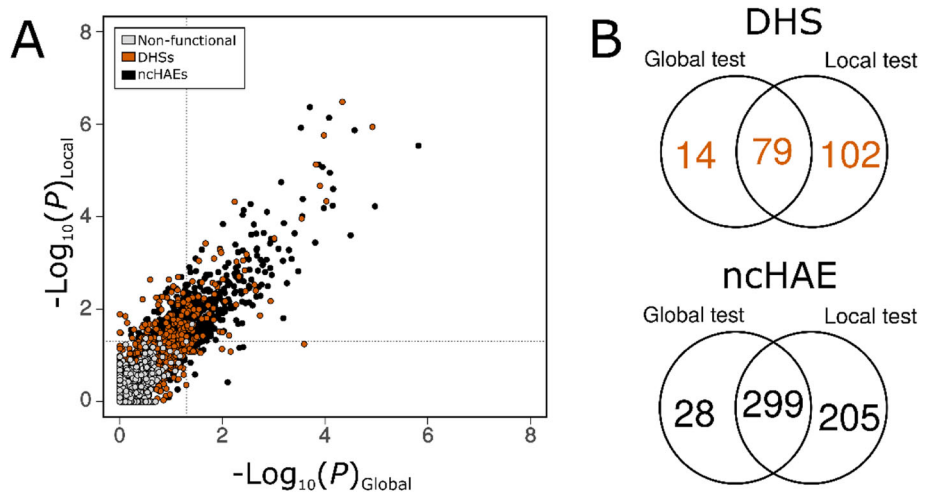

Fig. 3 Global proxy as a useful neutral proxy. **a**. Correlation between local and global tests of selection among different classes of DNA elements; the Spearman rank correlation coefficient are highly significant and very highfor DHSs ($P < 2.2 \times 10^{-16}$, $\rho = 0.80$) and ncHAEs ($P < 2.2 \times 10^{-16}$,$\rho = 0.86$), while correlation is low for NFRs ($P < 2.2 \times 10^{-16}$, $\rho = 0.52$). Inner dashed lines depict a significance level of $P = 0.05$. **b.** Venn Diagram of the overlap of regions scoring high for positive selection using the global test vs the local test for DHS elements (top) and ncHAEs (bottom)

values in the global and local sets is high among the 1291 ncHAEs that could be tested using both local and global proxies (Spearman's Rank test $\rho = 0.86$; $P < 2.2 \times 10^{-16}$). Of these, only 25.33% of the ncHAE regions tested positive globally, while 39.04% tested positive for selection using the local tests (Fig. 3b). Thus, local proxies in general identify more putative cases of positive selection but have limited applicability within function-dense regions of the genome while global proxies can be used to test any query region but possibly with lower sensitivity.

Sensitivity is high given practical query length, reference length, and branch number

Earlier, we observed that the distribution of substitution rates among all global reference regions used in this study is narrow and high (Fig. 2b). More specifically, we observed an average human branch length of 0.0072 substitutions per site using global neutral proxies, which is appreciably faster than the average substitution rates of local elements (average branch length = 0.0055). When we evaluated the effect of query and neutral proxy length on the sensitivity of the estimation of positive selection using empirical data, we measured the effect of reference length on substitution rate; we tested reference alignments of 300 bp, 900 bp, 3 kb, 9 kb, and 30 kb (Fig. 4a). As expected, the median of the branch lengths of the global references does not increase or decrease as they get longer; rather, they reach an equilibrium at 0.00725 with reduced variation (Fig. S5).

Functional genomic approaches such as ChIP-seq and ATAC-seq identify putative regulatory regions with window lengths that are usually between 150 and 800 bp and skewed towards shorter lengths [33, 36–40]. In order to assess the ability of *adaptiPhy* to identify positive selection throughout the biologically meaningful range of putative regulatory element sizes, we tested the effect of query lengths. Importantly, the ability to detect selection is not strongly affected by differences in query length, and remains similar down to ~ 150 bp (Fig. 4b). This finding suggests that our set of reference sequences is able to detect signatures of positive selection in regions where the query is longer than the actual functional element under selection, and in particular across most of the size range of known regulatory elements and open chromatin regions in the human genome.

Finally, we tested the impact of using three or five species to detect positive selection, as more branches might be expected to provide a better estimation of the background substitution rate in the reference. We found that adding one or two species above the minimum of three (two ingroup and one outgroup) provides only a negligible improvement in sensitivity of *adaptiPhy* (Fig. 4c). In contrast, the sensitivity of *phyloP* is more dependent on the number of species, improving markedly with additional taxa (Fig. 4c). Thus, *adaptiPhy* may be preferable in situations where the minimum number of reference genome assemblies is available.

The test discriminates between four different types of evolutionary scenarios

To determine whether *adaptiPhy* can correctly detect selection under different evolutionary scenarios, we simulated reference alignments evolving neutrally and query alignments evolving under four selection regimes, namely neutral on both background branches (BG) and

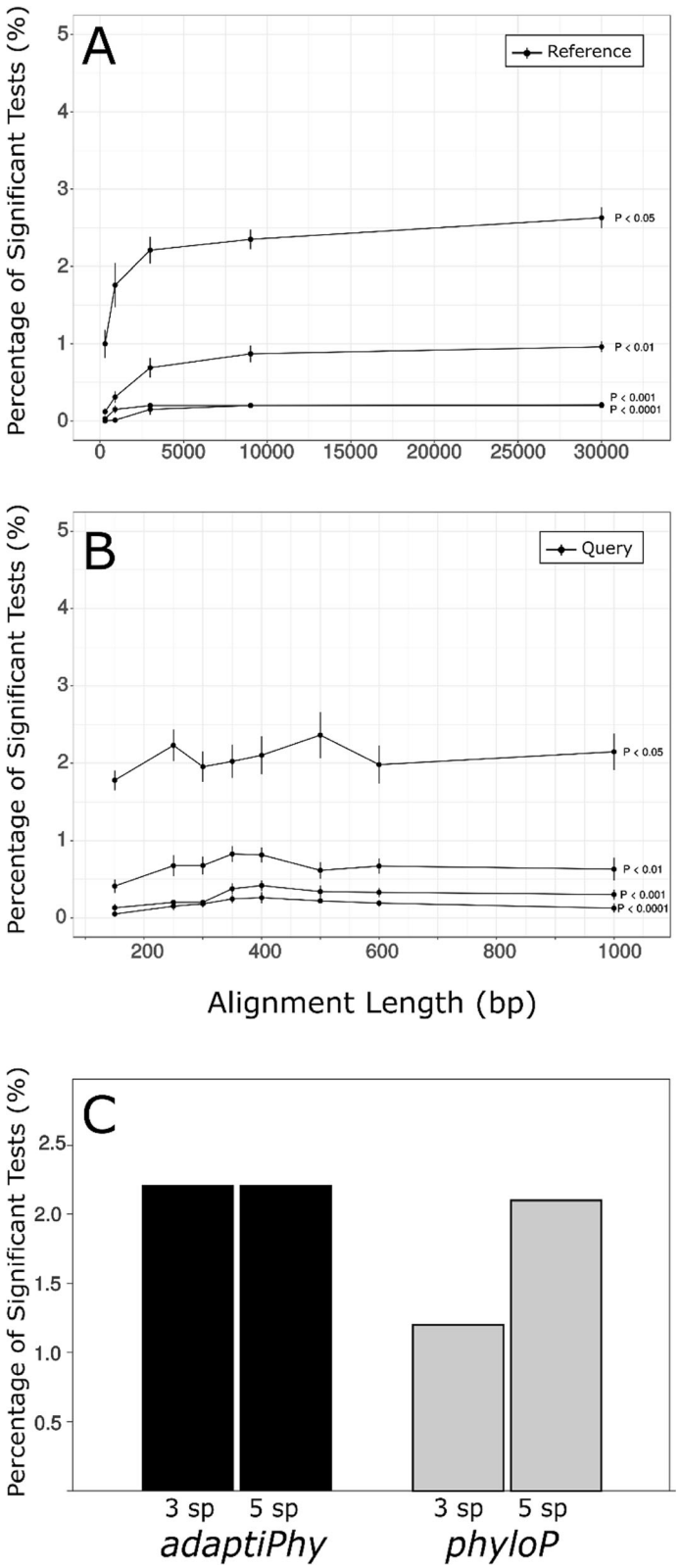

Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Sensitivity test and the effect of reference and query length, and number of species in the estimation of selection in a subset of 1000 random DHSs. **a**. Effect of reference length on the power to identify selection at different significance levels. **b.** Effect of query length on the power to identify selection at different significance levels. **c.** Percentage of significant tests in a sample of 1000 DHSs that were analyzed with *phyloP* and *adaptiPhy*. Bars labeled as 5 sp. depict tests done for five branches, while 3 sp. labels depict three-species tests

a foreground branch (FG) of a five-species tree, purifying selection on the BG and neutral on the FG (i.e., relaxation of constraint), neutral on the BG but positive selection on the FG, and positive selection in the FG and one BG species, while the other BG species remain neutral. We found that *adaptiPhy* accurately discriminates positive selection from relaxation of constraint and neutral evolution (Fig. 5). Of the simulated neutral regions, 1.8% were incorrectly identified as positive selection using *adaptiPhy* compared to 2.8% using *phyloP*. In addition, *phyloP* fails to distinguish positive selection from relaxation of constraint in almost half of the simulated cases, while our approach fails in only 2.2% of cases. In contrast, of the simulated regions under positive selection, *adaptiPhy* and *phyloP* identified 99.7 and 99.4% of sequences simulated to be under positive selection, respectively. When positive selection occurs in the FG and in one of the BG species, *adaptiPhy* and *phyloP* identified 99.2 and 83.5% of the cases, respectively (Fig. 5). Across all the evolutionary scenarios tested, the sensitivity of *adaptiPhy* to detect positive selection is 0.99 and specificity is 0.98, while the sensitivity of *phyloP* is 0.91 and specificity is 0.74. While it is difficult to evaluate whether the same would occur with real data, these results from simulations suggest that *adaptiPhy* may produce a lower false positive rate than *phyloP*, particularly for instances of relaxed selection.

The test reconfirms many previously identified human accelerated elements

To test whether *adaptiPhy* and *phyloP* replicate results from previous scans for positive selection in noncoding regions using our global NFRs [15, 17–19], we queried a consolidated set of 2649 ncHAEs [41]. Since our findings indicate that our test does not dilute signals of positive selection when the query region is between 150 bp and 1 kb, and given that most ncHAEs are relatively short (67% are shorter than 300 bp), we normalized query length by capturing sequence up to 300 bp centered on each ncHAE. At a significance level of 0.05, we confirmed 25.8% of previously reported ncHAEs using *adaptiPhy* (Fig. S4A), while confirming 59% using *phyloP* (Fig. S4B). More broadly, the distribution of *P*-values for previously reported ncHAEs is skewed toward 0; the peak at the lower end of the distribution is pronounced (Fig. S4C). As expected, this distribution is more skewed towards 0 using *phyloP* and our set of neutral references (Fig. S4D). Indeed, these results are consistent with rapid acceleration in the human lineage among ncHAEs when using our proxy neutral with both *phyloP* and *adaptiPhy*.

Interestingly, there is a substantially higher degree of overlap between the regions we replicated in the consolidated set of ncHAEs and those in each of the published studies than between any two of them (Fig. S6). For instance, we validated nearly 55% of the ncHAEs identified by Bush and Lahn [18], a significantly higher fraction than any of the other studies were able to validate (Fisher's Exact Test, Two-sided, $P = 0.0007$). Of the other methods that were used to identify human accelerated elements, Bush and Lahn's is the most methodologically similar to ours, although they used ancient repeats within 750 kb surrounding each ncHAE as the neutral proxy. Only four loci were identified as ncHAEs in all four prior studies, and all four were replicated here (Table S1 and Fig. S7). One of these is near *BNC2*, a gene that may be responsible for skin pigmentation differences between humans and other primates [42, 43].

We also scanned a region of 100 kb containing one of the genes that was significant for positive selection in the published studies [15, 17–19] and the present study using a sliding window. We found additional signals of positive selection around the *NEIL3* locus that were as strong as the single ncHAE element that was originally identified (Fig. 6). The striking clustering of signals of positive selection around this locus suggests that transcriptional regulation of *NEIL3* changed extensively during human origins and highlights the ability of *adaptiPhy* to identify signatures of positive selection that other methods may miss.

Discussion

Rigorous testing for selection depends upon accurate identification of reference regions that represent neutral rates of evolution. Prior studies have used four-fold degenerate (4D) sites from coding sequences [15], surrounding non-coding sequences [27], concatenated conserved LINE elements (ancient repeats; ARs) [28], and non-first intron sequences [21] as neutral references. However, regulatory elements are often located in regions that are dense with functional annotations and thus potentially constrained or near sites evolving under weak purifying selection (Fig. 2a). Consequently, identifying sufficient non-functional regions to use as a neutral proxy in the vicinity of most open chromatin sites is simply not possible in primate genomes and this problem is even more acute in organisms with higher gene

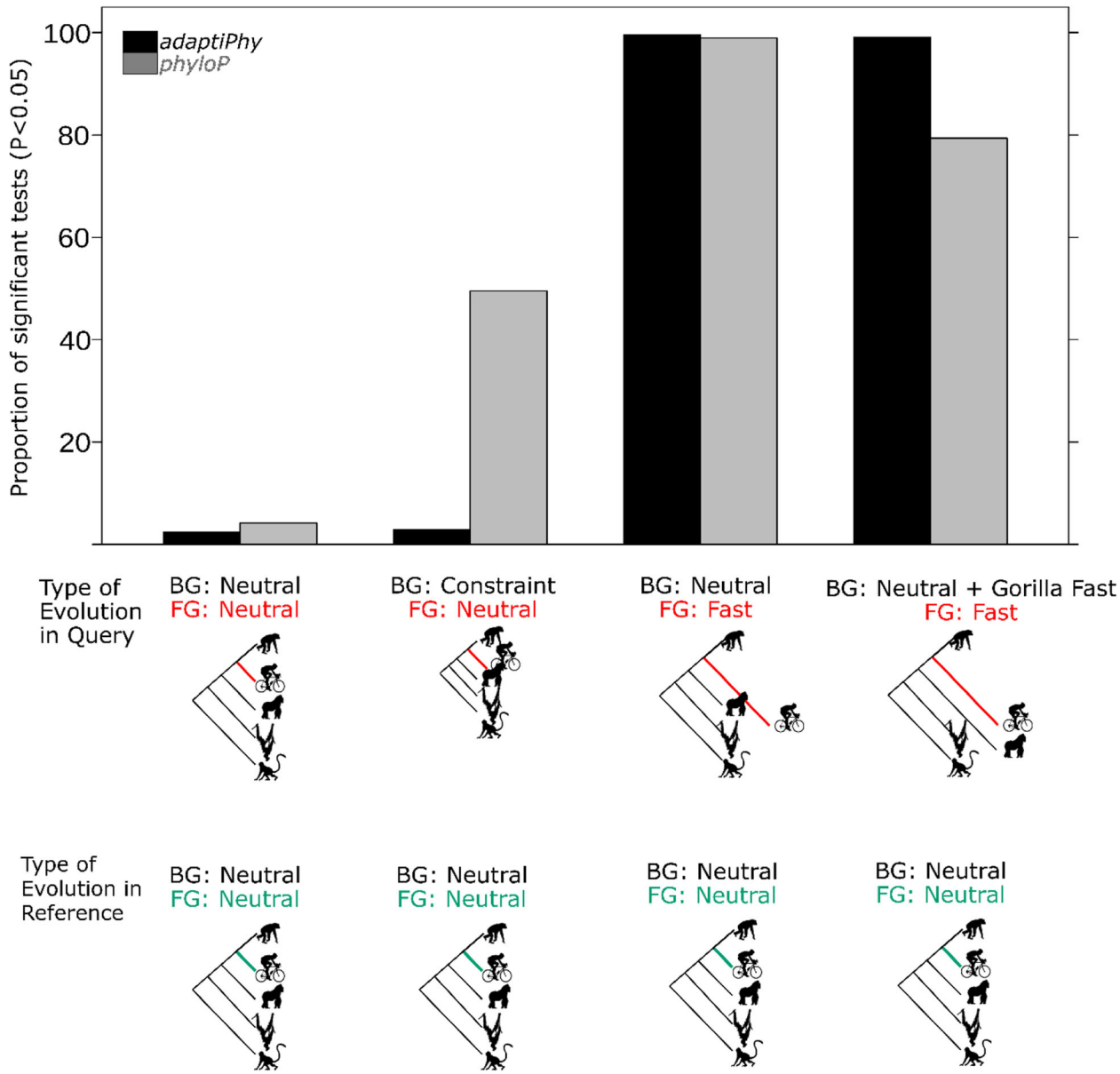

Fig. 5 Specificity test of *adaptiPhy* and *phyloP* for three types of evolution. All trees and alignments of each category were simulated to be evolving by neutral evolution (left), relaxation of the constraint (center-left), positive selection (center-right), and positive selection in the foreground and the gorilla branch with positive selection (right). The left and center simulations give an idea of the amount of type I error, while the right simulation provides an estimation of the type II errors

density. To address this limitation, we used the dense functional annotation of the human genome to identify a set of putatively non-functional regions (based on absence of any functional annotation) that can be used as an appropriate neutral proxy (see Methods).

We then filtered this set of putatively non-functional regions (NFRs) to address the challenges of rate heterogeneity and lineage sorting. For the entire set of DHS elements in our sample, we found a wide distribution of substitution rates among concatenated local samples including local NFRs, prefiltered global NFRs and ARs (Fig. 2b). This means that there are many such regions in which the substitution rate on the human branch is substantially slower or faster than average and thus potentially confounding when used to test for positive selection. These observations reflect generally less availability of non-functional elements in regions where DHSs and ncHAEs are more common and slower rates of evolution nearby each functional element, but also the possibility that some ancient LINE elements in the

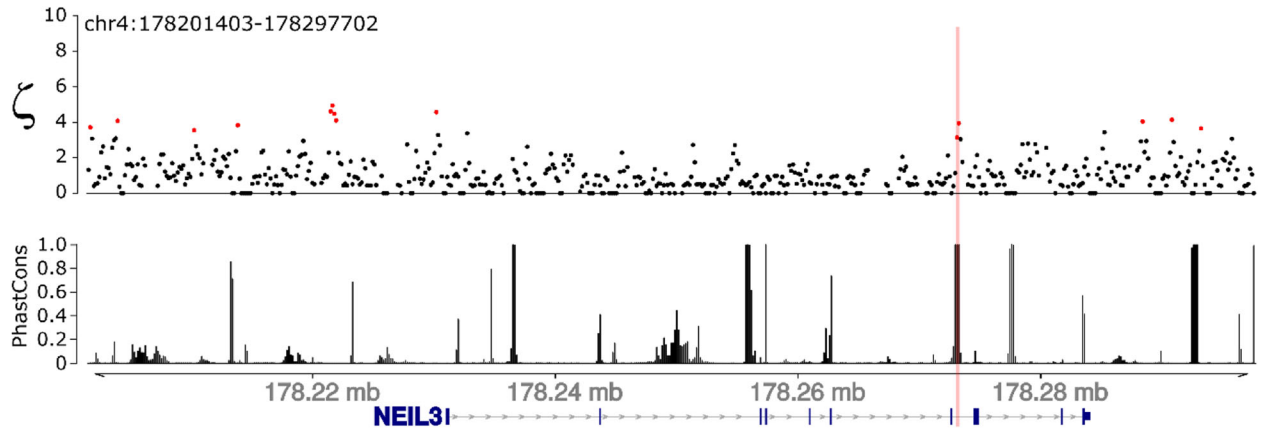

Fig. 6 Sliding windows analysis of the rate of evolution along a genomic region containing a ncHAE. Distribution of positive selection as ζ (top panel), and vertebrate conservation scores (bottom panel) along a 100 kb region surrounding the gene NEIL3. This locus is located nearby one of the human accelerated elements identified in our test and across all four of the cited studies (Fig. S6). The location of this ncHAE is highlighted in pink, and the red dots represent windows of 300 bp where ζ scored significant for positive selection ($P < 0.05$)

vicinity of DHSs are under functional constraint or evolving under high mutation rates. Indeed, many studies have shown that ARs can become co-opted as local regulatory elements e.g., [44–47]. Consistent with this interpretation, we found that the proportion of reference ARs with at least one multitissue eQTL or brain specific eQTL is almost three times as large as in a set of global non-functional sequences (Fig. S3). The relative absence of local non-functional elements may also increase the effect of incomplete lineage sorting in the reference sequence, producing an increase of false positives and negatives for query elements if the reference foreground sequence is evolving at slower or faster rates than the background.

Together, these results suggest that using local non-functional elements and ARs can both contribute to misestimation of selection. At the same time, ARs have the tendency to overestimate the rates of evolution of query alignments compared to a set of neutral references built from global NFRs (Fig. S2). Our findings also suggest that the local tests have a tendency to overestimate positive selection, probably because they contain unknown functional sites and thus underestimate the neutral rate. As expected, most putative non-functional elements appear neutral in tests for selection, with only one testing positive for selection using both local and global references (Fig. 3a). Overall, our results suggest that global proxies are capable of testing more genomic regions, with an important gain in accuracy as we sample putatively neutral elements based on their distribution pattern relative to the tree rates. As a consequence, every query should be tested against a reference region in which the rate on the foreground branch is approximately equal to the rate on the background branches. Local reference regions can yield higher sensitivity to positive selection but also more false positives due to unrecognized purifying selection on parts of the reference region. Since incomplete lineage sorting or variation in the time to coalescence can be a potential confounding factor for both local and global neutral proxies [48], our filtering of global sequences based on relative branch lengths and concatenation of non-functional elements provides a useful strategy to control these issues [49–51]. Moreover, coalescence times for human polymorphisms are generally in the 10^3–10^5 year range, which is shorter than the time of the between-species branches. Cases of lineage sorting between ape species certainly occur but are rare [52]. Nevertheless, we suggest caution when analyzing query regions scoring high for positive selection; this is particularly important for genomes with relatively short branch lengths. Despite these recommendations, there are situations where local references may be useful alone or as a complement to global references. In particular, local references may be useful in regions of the genome with a lower density of functional elements, because they may more accurately reflect the local substitution rate. Indeed, the implementation of *adaptiPhy* allows the user to choose which reference to use (or both).

We also observed that among DHSs and nonfunctional DNA elements, the distribution of *P*-values tends towards the upper limit ($P = 1$) (Fig. S4C). This means that most open-chromatin regions are probably evolving neutrally or under purifying selection because the maximum likelihood estimate of the null and the alternative are exactly the same when $P = 1$ (Fig. S4C). In contrast, most tests on ncHAEs show a distribution of *P*-values strongly skewed towards zero with a small but important density peak at 1, suggesting that at least 92 out 2416 regions originally identified as ncHAEs appear to be evolving

neutrally or by purifying selection in the human branch when using global proxy (Fig. S4C). Some of these may reflect errors in the early reference assemblies that were used in prior studies. Overall, however, our results are generally consistent with earlier findings [15, 17–19], identifying many of the same regions as being under positive selection. Moreover, our results suggest that the distribution of *P*-values is strongly dependent on the genomic partition: DHSs and our set of putatively non-functional regions scored nearly 2.5 and 0.02% of sites under positive selection respectively, confirming the expectation that DHSs in general are more often subject to positive selection than putatively nonfunctional DNA regions. This fraction is higher than previously reported [27, 28]. The likely reason is that these prior studies underestimated the proportion of sites under selection because the AR elements used as neutral proxies are evolving under higher substitution rates than the rest of the genome (Fig. S2).

By testing reference alignments of different lengths, we show a diminishing added return as these neutral proxies get longer. We recommend using reference alignments between 3 and 9 kb (Fig. 4a), as longer alignments require more computing power while providing only minimal additional sensitivity. Reference alignments shorter than 3 kb introduce more variation in estimation of ζ and thus increase the risk of false positives and negatives. We also tested the effect of query region length and found little difference in sensitivity between regions of 150 bp and 1000 bp, which means our approach can be applied to the vast majority of putative regulatory elements. We recommend caution in extending query region length indefinitely because this risks combining multiple functional elements with neutral and non-neutral substitution rates.

Our framework can be used to detect sequence outliers under a high substitution regime while controlling for relaxation of constraint. We propose that most signals that are detected reflect true instances of positive selection. Locally elevated mutation rate is a potential confound [53], however. Some evidence suggests that mutation rate correlates with substitution rates between species within long segments of the genome (~ 50 kb) [54], perhaps due to mutation pressure. However, the situation remains unclear, as other studies found that mutation rate does not explain variation in divergence between species [55, 56]. In addition, regulatory elements are much shorter (generally 100–350 bp), and our sliding window scans reveal highly localized peaks of elevated substitution rather than the broad regions analyzed in the studies just mentioned. To investigate whether local differences in mutation or recombination rate might produce a false positive, we recommend investigating the amount of common and rare variation present in a query region by consulting dbSNP or 1000 genomes databases. Rare variants are often used as a proxy for mutation rate [56–58], making this a straightforward way to flag queries that score high for positive selection but are under a higher or lower mutational regime. Moreover, finding an excess of common variants under high-to-intermediate frequencies in the Site-Frequency Spectrum (SFS) in a given region that scores high for positive selection is potentially a useful way to identify previously unknown regions under persistent balancing selection [59]. Indeed, we found that the most variable element under positive selection is located in the MHC region (Supplementary Data), a region known to be under balancing selection.

Another potential confound of our approach is variation in recombination rate [48], which can cause discordance in the phylogenetic inference of substitution rates. Again, however, our query regions are very short (100 s of bp) while values for recombination rates in the genome are calculated over much longer scales (~ 100 kb) and are still rather imprecise [60, 61]. Furthermore, the effects of a selective sweep are more evident at the within- than between species scale. Some tests for selection designed to work within species measure reduced polymorphism around the selected site and later a signal of low-frequency derived mutations in linkage disequilibrium (LD) with it. These signals only persist for tens of thousands of years in human populations, since over time recombination erodes LD and drift eliminates most low-frequency alleles [62]. As a result, it is not possible to test for selection across deeper evolutionary time scales using signals like the SFS or LD. Divergence times between great apes range from ~ 7–16 million years, much too long for the LD effects of a selective sweep to be evident and leaving only a signature of elevated substitution.

If selective sweeps did influence tests for positive selection at the between-species level, we would expect the same query regions to be flagged by tests at the within- and between-species levels. This is not what we observe. As a specific example, we analyzed the region that includes SNPs responsible for lactase persistence in some human populations. The region around these SNPs shows one of the strongest selective sweeps known from the entire human genome, making it a good test case. Of 1847 windows tested around the LCT locus, only 9 (0.48%) scored significant for positive selection using *adaptiPhy* (Fig. S7), which is about the percentage among non-functional regions distributed randomly across the genome. Further, the specific area around the lactase persistence SNP that recently swept in humans is not positive for selection by our test for selection between species.

A final potential confound is the idea that reference genome sequences do not capture all the variation that

is present at the population level [63]. In this study, we used the human (hg19), chimpanzee (panTro4), gorilla (gorGor3), orangutan (ponAbe2), and macaque (rheMac3) reference assemblies. None of these reflect the ancestral state for their respective species, and are instead based on observed sequences from at least one diploid individual. As a result, some regions may contain concentrations of rare or common derived alleles simply by chance. This will artificially increase the apparent interspecies substitution rate. To control for this potential confound, we suggest re-running tests for selection for any query regions of particular interest using local alignments that represent the reconstructed ancestral state in order to correct for intraspecific variation.

Conclusion

Together, the improvements introduced here increase the ability to identify proxy neutral regions and to modulate the sensitivity and conservativeness of branch-specific tests for positive selection in noncoding regions. The test is sensitive across nearly the entire range of annotated functional regulatory elements, dropping only for elements < 150 bp in length. It is possible to apply these tests to nearly any noncoding region of the genome, even those in functionally dense locations.

Methods

Testing for branch-specific positive selection using our approach requires at least one reference alignment in which all branches of the tree are evolving at putatively neutral rates, and one query alignment, which can be obtained from any genomic region of interest, such as putative open chromatin regions or segments of a GWAS peak (Fig. 1b). First, we downloaded the 100-way multiple alignment from the University of California Santa Cruz (UCSC) website (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/) and several annotations of functional DNA elements from the ENCODE Project at UCSC, including: 5′ and 3′ UTRs, total human mRNA, lincRNAs, microRNAs, sncRNAs, short repeats, CpG islands, etc. (supplementary Table 2). We also enriched this list of functional elements with a set of HoneyBadger2-intersect promoters and enhancers from reg2map annotated by the Broad Institute and Epigenomics Roadmap project. Then, using *maf_parse* implemented in PHAST [22], we transformed the 100-way alignment into a smaller 5-way genome-wide alignment in MAF format that included only our focal species human (hg19), chimpanzee (panTro4), gorilla (gorGor3), orangutan (ponAbe2), and rhesus macaque (rheMac3). To draw reference alignments and non-functional regions, we generated a masked 5-way MAF alignment with a BED file containing all known functional DNA regions using *maf_parse* with the optional command *--mask-features*. To mask the genome, we used a merged BED file that included 5′ and 3′ exons, all coding and non-coding RNAs, vista enhancers, roadmap and ENCODE regulatory elements and promoters, CpG repeats, microsatellite sequences and simple repeats. Interestingly, the remaining non-functional fraction of the genome covered only ~ 20.5% of the genome. We used *msa_split* [22] to draw alignments from non-coding human accelerated elements, ncHAEs [15, 17–19, 64], a random subset of non-functional regions (as defined below), and DNA Hypersensitive Sites (DHSs) from 125 human cell types and tissues [33].

To select non-functional regions (NFRs) for our analyses of positive selection, we randomly chose around two million (1,893,795) non-overlapping segments of 300 bp from the non-functional fraction of the genome described above, collectively amounting to 18.3% of the genome. Subsequently, we excluded all the alignments containing any masked regions. Next, we computed branch lengths of each of the tree branches using the tool *phyloFit* available in PHAST [22]. *phyloFit* computes a tree with branch lengths and a substitution rate matrix by fitting a tree model to a multiple sequence alignment using maximum likelihood for a total of 92,160 regions. We noted a large peak of substitution rates near 0, and to avoid this bias we removed sites with a relative substitution rate < 0.001 for any of the sites. Within the remaining sample of 52,879 alignments from the last filtration step, we excluded any alignments in which the human branch was evolving too quickly or too slowly with respect to the total tree by using a custom R script; in this step, we omitted all trees with a relative branch length within the top and bottom 25% of this distribution (Fig. S1). This step reduced the pool of global NFRs to 26,426 FASTA alignments in which the relative human branch length ranged between the lower and the upper quartiles. The genome locations of these NFRs are available in the supplement (NFR.sorted.bed.csv).

To run our tests of selection, we fitted the null and alternative models for each DNA-element alignment using the batch scripts written by Haygood and collaborators [21] that run under the program HyPhy [29]. After all tests were completed, we extracted the best maximum likelihood estimates from twenty fittings to allow for stochasticity. Then, we obtained *P*-values from each likelihood ratio test (LRT) using the Chi^2 distribution tool (*pchisq*) with one degree of freedom, implemented in R [65]. Consistent with previous findings [66], we observed that all *P*-value distributions were non-uniform and highly skewed to 1, therefore we considered our test to be conservative (Fig. 3b). Consequently, we decided to use nominal *P*-values smaller than 0.05 to name regions scoring high for positive selection, instead of correcting for multiple testing to avoid violating the assumption of uniform *P*-value distribution underlying the False Discovery Rate

methodology [67, 68, 69]. We recognize that each query is independent of the rest, and thus the problem of multiple testing is real. We thus recommend that any query region where the nominal *P*-value indicates positive selection should be re-tested against different reference regions. In this case, we recommend adjusting *P*-values when the same query region is tested multiple times or when testing enrichment between multiple groups. All data generated during this study are included in the supplementary information files as comma delimeted files (NFR.data, DHS.data and ncHAE.data).

Evaluation of the effect of query and reference length on sensitivity using empirical data

To examine the sensitivity of our framework to detect positive selection under varying lengths of the query and reference, we obtained seven sets of 1000 queries of DHS alignments from the center of the peak position up to a total of 150 bp, 250 bp, 300 bp, 400 bp, 500 bp, 600 bp and 1000 bp on both sides. For each of these alignments, we also generated 10 reference alignments in order to account for the stochasticity of the evolutionary processes by concatenating 10 alignments from our set of 26 K non-functional 300 bp elements. Likewise, to identify the effects of variation in reference length in our queries, we also ran each query alignment of 300 bp against each of ten reference alignments varying from 300 bp to 30,000 bp. We also investigated the effects of including five species in our alignments rather than the minimum of three species considered by Haygood and collaborators [21]. There are now many genomewide assemblies that are available for many species, so we decided to employ additional species to test the effect of removing two branches in the tree. To do this, we extracted a random pool of 1000 queries with both three (human, chimp and macaque) and five species (human, chimp, gorilla, orangutan and macaque) from the DHSs prepared by Thurman and collaborators [33].

Evaluation of the effect of evolutionary state on specificity and sensitivity using simulated data

One typical concern of any test of directional selection is the fact that it may deliver a signal of positive selection if reference sequences are under unrecognized purifying selection or if mutation rates are increased in the query region [66]. To test for these effects, we used the distribution of both relative and absolute branch lengths among non-functional sequences to simulate the distribution of trees under different classes of evolution (i.e. relaxation of constraint, increased mutation rate, positive selection or neutrality). To do this, we assumed that the alignments in the second quartile were putatively neutral, while those in the lower quartile were constrained, and those beyond the highest value of the upper quartile were under positive selection. Subsequently, we used these 'neutral' distributions to generate random sets of simulated trees, which were executed in the program *seq-gen* [70] to simulate sequence alignments in FASTA format. Consequently, we generated four sets of 1000 query alignments in which 100% of the alignments were evolving by different classes of evolution: *i*, neutral in the foreground and in the background; *ii*, neutral in the foreground but constrained in the background by scaling down the substitution rates in the background by a factor $\alpha = 0.1$; *iii,* scaling up the neutral rate in the foreground by a factor $\alpha = 7$; and iv, scaling up the neutral rate in the foreground by a factor of $\alpha = 7$ and scaling up the neutral rate in one of the background species (i.e. gorilla) by a factor of $\alpha = 10$. Next, we used these simulated alignments to explore the power of *adaptiPhy* to distinguish positive selection from relaxation of constraint and other types of evolution at the phylogenetic scale. Finally, we compared the results obtained from *adaptiPhy* with *phyloP*. This computational tool estimates a null distribution of the number of substitutions from the reference model, and computes the number of actual substitutions that occur in the query alignment, then it estimates *P*-values in the branch of interest given the total tree using a LRT method [16]. Here, for each query region used before, we tested the human subtree using the SPH model against the putatively neutral reference alignment obtained from *phyloFit*. Then, we parsed the *P*-value of acceleration in subtree given total tree using custom bash scripts.

Evaluation of sensitivity using known non-coding human accelerated elements

To test for selection in regions of the genome that have been previously investigated, we obtained a consolidated set of positive control queries of 2649 ncHAEs studied by Capra (2013) [41]. To compare the fraction of positive selection among ncHAEs with other sets of known regulatory elements, we used a subset of DHSs previously published by Thurman and collaborators (2012) [33]. To sample the DHSs in our test, we defined the "Ubiquity Score" as the fraction of cell types a given open chromatin site was open among the total number of tissues or cell types surveyed, thereby identifying a different set of queries to run our LRT method on known regulatory elements. To do this, we selected 4216 DHSs that were open in at least 124 of 125 cell types, which we term 'widespread sites' (Ubiquity score > 0.98), and we selected 7433 DHSs that were open in at most 2 of 125 cell types, which we term 'specific sites' (Ubiquity score < 0.02). Although these numbers seem arbitrary, they are a consequence of losing sequence alignments due to missing sequences from any of the branches or high frequency of gaps.

Adapting the method to sliding windows

To test if our method is suitable for sliding windows, we applied *adaptiPhy* to test for positive selection within a 100 kb region around the *NEIL3* locus, with the purpose of investigating additional signals of selection in the vicinity of a ncHAE. Here, we split the entire region in windows of 300 bp with a 150 bp step size, and each query was run against a reference made from the filtered NFRs that were extracted from the masked genome alignment.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6752-4.

Additional file 1: Supplementary information.
Additional file 2: NFR.sorted.bed.csv.
Additional file 3: NFR.data.
Additional file 4: DHS.data.
Additional file 5: ncHAE.data.

Abbreviations

NFR: Non-Functional Region; **DHS**: DNA Hypersensitive Site; **HAR**: Human Accelerated Region; **ncHAE**: Noncoding Human Accelerated Element; **4D**: Fourfold Degenerate sites; **ChIP-seq**: Chromatin Immunoprecipitation sequencing; **ATAC-seq**: Assay of Transposase-Accessible Chromatin using sequencing; **AR**: Ancient Repeat; **LRT**: Likelihood Ratio Test; **SPH**: Siepel, Pollard and Haussler; **GERP**: Genomic Evolutionary Rate Profiling; **LD**: Linkage Disequilibrium; **SFS**: Site-Frequency Spectrum

Acknowledgements

We thank Dr. Andrew Allen and members in the Wray lab who provided feedback during the writing of this manuscript.

Authors contributions

A.B and G.A.W designed the study and wrote the manuscript. G.A.W and R. H. oversaw the project and analysis. A.B., R. H. and G.A.W analyzed and discussed. A.B., R. H., and G.A.W helped revise the manuscript. The author(s) read and approved the final manuscript.

Funding

Alejandro Berrio was supported by a Hargitt Fellowship from Duke University and NIH grant 1R01-TW010870 during the design of the study, data collection, analysis, interpretation of data, and during the writing of this manuscript.

Availability of data and materials

Genomewide aligment of the 100 vertebrate alignment used in this study is available at the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/). Functional genomic data used in this manuscript to mask functional regions of the genome are available at UCSC genome data browser (https://genome.ucsc.edu/cgi-bin/hgTables), ENCODE (https://www.encodeproject.org/data/annotations/) and HoneyBadger2 (https://personal.broadinstitute.org/meuleman/reg2map/), see Supplementary Table 2 for direct web links. Scripts and software for running tests of selection using adatiPhy are available in Github (https://github.com/wodanaz/adaptiPhy). This GitHub repository also contains a Docker file with the minimal requirements to run *adaptiPhy* in a sample dataset. Datasets containing results of selection scores and neutral sequence regions used and/or analyzed during the current study are available from the supplementary information. Other datasets and R scripts used in the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not Applicable.

Competing interests

All the authors declare no conflict of interest.

Author details

[1]Department of Biology, Duke University, Biological Sciences Building, 124 Science Drive, Durham, NC 27708, USA. [2]Ronin Institute for Independent Scholarship, 127 Haddon Pl., Montclair, NJ 07043, USA.

Received: 22 October 2019 Accepted: 21 April 2020
Published online: 13 May 2020

References

1. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15:496–503. https://doi.org/10.1016/S0169-5347(00)01994-7.
2. Wayne ML, Simonsen KL. Statistical tests of neutrality in the age of weak selection. Trends Ecol Evol. 1998;13:236–40. https://doi.org/10.1016/S0169-5347(98)01360-3.
3. Nadeau NJ, Jiggins CD. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet. 2010;26:484–92. https://doi.org/10.1016/j.tig.2010.08.004.
4. Pardo-Diaz C, Salazar C, Jiggins CD. Towards the identification of the loci of adaptive evolution. Methods Ecol Evol. 2015;6:445–64. https://doi.org/10.1111/2041-210X.12324.
5. Reilly SK, Noonan JP. Evolution of gene regulation in humans. Annu Rev Genomics Hum Genet. 2016;17:45–67. https://doi.org/10.1146/annurev-genom-090314-045935.
6. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 1985;2:150–74. https://doi.org/10.1093/oxfordjournals.molbev.a040343.
7. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3:418–26. https://doi.org/10.1093/oxfordjournals.molbev.a040410.
8. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, et al. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. Science (80-). 2003;302:1960–3. doi:https://doi.org/10.1126/science.1088821.
9. Iacobuzio-Donahue CA, Ashfaq R, Maitra A, Adsay NV, Shen-Ong GL, Berg K, et al. Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies. Cancer Res. 2003;63:8614–22. https://doi.org/10.1126/science.1058040.
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. https://doi.org/10.1038/35057062.
11. Rhesus Macaque Genome Sequencing and Analysis Consortium RA, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science. 2007;316:222–34. doi:https://doi.org/10.1126/science.1139247.
12. Hedrick PW, McDonald JF. Regulatory gene adaptation: an evolutionary model. Heredity (Edinb). 1980;45:83–97. https://doi.org/10.1038/hdy.1980.52.
13. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh S-D, et al. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature. 2006;440:1050–3. https://doi.org/10.1038/nature04597.
14. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:206–16. https://doi.org/10.1038/nrg2063.
15. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2006;2:e168. https://doi.org/10.1371/journal.pgen.0020168.
16. Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg; 2006. p. 190–205. doi:https://doi.org/10.1007/11732990_17.

17. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, et al. Fast-evolving noncoding sequences in the human genome. Genome Biol. 2007; 8:R118. https://doi.org/10.1186/gb-2007-8-6-r118.
18. Bush EC, Lahn BT. A genome-wide screen for noncoding elements important in primate evolution. BMC Evol Biol. 2008;8:17. https://doi.org/10.1186/1471-2148-8-17.
19. Prabhakar S, Visel A. Akiyama J a, Shoukry M, Lewis KD, Holt a, et al. human-specific gain of function in a developmental enhancer. Science. 2008;321: 1346–50. https://doi.org/10.1126/science.1159974.
20. Wong W, Nielsen R. Detecting selection in noncoding regions of nucleotide sequences. Genetics. 2004;167:949–58.
21. Haygood R, Fedrigo O, Hanson B, Yokohama K, Wray G. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet. 2007;39:1140–044. https://doi.org/10.1038/ng2104.
22. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12:41–51. https://doi.org/10.1093/bib/bbq072.
23. Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006;443:167–72. https://doi.org/10.1038/nature05113.
24. Rao CR. Score Test: Historical Review and Recent Developments. Adv Rank Sel Mult Comp Reliab. 2005;:3–20. doi:https://doi.org/10.1007/0–8176-4422-9_1.
25. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13. https://doi.org/10.1101/gr.3577405.
26. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.
27. Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, et al. Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. 2015;25:1245–55. https://doi.org/10.1101/gr.192591.115.
28. Dong X, Wang X, Zhang F, Tian W. Genome-wide identification of regulatory sequences undergoing accelerated evolution in the human genome. Mol Biol Evol. 2016;33:2565–75. https://doi.org/10.1093/molbev/msw128.
29. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9. https://doi.org/10.1093/bioinformatics/bti079.
30. Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011;9:e1001046. https://doi.org/10.1371/journal.pbio.1001046.
31. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74. https://doi.org/10.1038/nature11247.
32. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518: 317–30. https://doi.org/10.1038/nature14248.
33. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489:75–82. https://doi.org/10.1038/nature11232.
34. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
35. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, et al. ENCODE whole-genome data in the UCSC genome browser: update 2012. Nucleic Acids Res. 2011. https://doi.org/10.1093/nar/gkr1012.
36. Crawford GE. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2005;16:123–31.
37. Birney E. Stamatoyannopoulos J a, Dutta a, Guigó R, Gingeras TR, Margulies EH, et al. identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799–816. https://doi.org/10.1038/nature05874.
38. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010:pdb.prot5384. doi:https://doi.org/10.1101/pdb.prot5384.
39. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012;8:e1002789. https://doi.org/10.1371/journal.pgen.1002789.
40. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation Of Chromatin Accessibility In Prefrontal Cortex Of Schizophrenia Cases And Controls. doi.org. 2017;:141986. doi:https://doi.org/10.1101/141986.
41. Capra J a, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. Philos Trans R Soc B Biol Sci. 2013;368:20130025. doi:https://doi.org/10.1098/rstb.2013.0025.
42. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507:354–7. https://doi.org/10.1038/nature12961.
43. Vernot B, Akey JM. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science (80-). 2014;343:1017–21. doi:https://doi.org/10.1126/science.1245938.
44. Greene E, Entezam A, Kumari D, Usdin K. Ancient repeated DNA elements and the regulation of the human frataxin promoter. Genomics. 2005;85:221–30. https://doi.org/10.1016/j.ygeno.2004.10.013.
45. Kamal M, Xie X, Lander ES. A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci. 2006;103: 2740–5. https://doi.org/10.1073/pnas.0511238103.
46. Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat Commun. 2014;5:4104. https://doi.org/10.1038/ncomms5104.
47. Zeng L, Pederson SM, Cao D, Qu Z, Hu Z, Adelson DL, et al. Genome-wide analysis of the Association of Transposable Elements with gene regulation suggests that Alu elements have the largest overall regulatory impact. J Comput Biol. 2018;25:551–62. https://doi.org/10.1089/cmb.2017.0228.
48. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 2007;56:17–24.
49. Rokas A, Williams BI, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003;425: 798–804.
50. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013;497:327–31.
51. Tonini J, Moore A, Stern D, Shcheglovitova M, Ortí G. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. PLoS Curr. 2015;7 TREEOFLIFE. doi:https://doi.org/10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be.
52. Mailund T, Munch K, Schierup MH. Lineage sorting in apes. Annu Rev Genet. 2014;48:519–35.
53. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14:262–74. https://doi.org/10.1038/nrg3425.
54. Castellano D, Eyre-Walker A, Munch K. Impact of mutation rate and selection at linked sites on fine-scale DNA variation across the homininae genome. bioRxiv. 2018;:452201. doi:https://doi.org/10.1101/452201.
55. Smith TCA, Arndt PF, Eyre-Walker A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLoS Genet. 2018;14.
56. Terekhanova N V., Seplyarskiy VB, Soldatov RA, Bazykin GA. Evolution of local mutation rate and its determinants. Mol Biol Evol. 2017;34:msx060. https://doi.org/10.1093/molbev/msx060.
57. Nei M. Estimation of mutation rate from rare protein variants. Am J Hum Genet. 1977;29:225–32 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1685315/. .
58. Slatkin M. Rare Alleles as Indicators of Gene Flow. Evolution (N Y). 1985;39:53. https://doi.org/10.2307/2408516.
59. Bitarello BD, De Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al. Signatures of long-term balancing selection in human genomes. Genome Biol Evol. 2018;10:939–55.
60. Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. 2001;17:481–5. https://doi.org/10.1016/S0168-9525(01)02409-X.
61. Nachman MW. Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev. 2002;12:657–63. https://doi.org/10.1016/S0959-437X(02)00358-1.
62. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science (80-). 2006;312:1614–20.
63. Eisfeldt J, Mårtensson G, Ameur A, Nilsson D, Lindstrand A. Discovery of novel sequences in 1,000 Swedish genomes. Mol Biol Evol. 2019. https://doi.org/10.1093/molbev/msz176.
64. Franchini LF, Pollard KS. Human evolution: the non-coding revolution. BMC Biol. 2017;15:89. https://doi.org/10.1186/s12915-017-0428-9.

65. Team RC. R: a language and environment for statistical computing. 2015. https://www.r-project.org/.
66. 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. 2005;22:2472–9.
67. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995;57:289–300. https://doi.org/10.2307/2346101.
68. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Inference. 1999;82:171–96. https://doi.org/10.1016/S0378-3758(99)00041-5.
69. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100:9440–5. https://doi.org/10.1073/pnas.1530509100.
70. Rambaut A, Grassly NC. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997;13:235–8 http://www.ncbi.nlm.nih.gov/pubmed/9183526.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions