Adaptive evolution of conserved non-coding elements in mammals

Su Yeon Kim[a] and

Jonathan K. Pritchard[b1]

[a]Department of Statistics

[b]Department of Human Genetics

The University of Chicago

July 16, 2007

[1]Address for correspondence: Dept of Human Genetics, The University of Chicago, 920 E 58th St–CLSC 507, Chicago IL 60637, USA. Email: skim@galton.uchicago.edu, pritch@uchicago.edu.

Abstract

Conserved non-coding elements (CNCs) are an abundant feature of vertebrate genomes. Some CNCs have been shown to act as *cis*-regulatory modules but the function of most CNCs remains unclear. To study the evolution of CNCs we have developed a statistical method called the "shared rates test" (SRT) to identify CNCs that show significant variation in substitution rates across branches of a phylogenetic tree. We report an application of this method to alignments of 98,910 CNCs from the human, chimpanzee, dog, mouse and rat genomes. We find that $\sim 68\%$ of CNCs evolve according to a null model where, for each CNC, a single parameter models the level of constraint acting throughout the phylogeny linking these five species. The remaining $\sim 32\%$ of CNCs show departures from the basic model including speed-ups and slow-downs on particular branches and occasionally multiple rate-changes on different branches. We find that a subset of the significant CNCs have evolved significantly faster than the local neutral rate on a particular branch, providing strong evidence for adaptive evolution in these CNCs. The distribution of these signals on the phylogeny suggests that adaptive evolution of CNCs occurs in occasional short bursts of evolution. Our analyses suggest a large set of promising targets for future functional studies of adaptation.

Author Summary

Conservation of DNA sequences across evolutionary history is a highly informative signal for identifying regions with important biological functions. In particular, conserved non-coding regions have been shown to be good candidates for containing regulatory elements that have roles in gene regulation. Recent studies have found that there are many thousands of conserved non-coding elements (CNCs) in vertebrate genomes and have suggested possible functions for some of these elements, but the function of most CNCs remains unknown. To study the evolution of CNCs, the authors developed a statistical method to identify CNCs that show changes in evolutionary rates on particular branches of the mammalian phylogenetic tree. Those rate changes may indicate changes in the function of a CNC. The authors have applied their method to CNCs of five mammalian genomes, and found that indeed many CNCs have experienced rate changes during their evolution. They also found a subset of CNCs showing accelerations in evolutionary rate that actually exceed the neutral rates, suggesting that adaptive evolution has shaped the evolution of those elements.

Introduction

Phenotypic evolution proceeds both by changes in protein coding sequences, and by changes in gene expression that determine when, where, and how much genes are expressed [1–3]. Although recent genome-wide studies have begun the process of identifying genes that show signals of adaptive evolution in coding sequences [4], much less is known about adaptation of regulatory sequences. One avenue to studying adaptation of gene regulation is to identify regulatory elements that show rapid evolution at the DNA sequence level [2]. However, a challenge for this approach is that at present we have only limited knowledge of the DNA sequence elements that drive gene expression and regulation.

One possible way forward is to study the evolution of conserved non-coding elements (CNCs) [5–7]. In recent years it has been shown that $\sim 3.5\%$ of non-coding DNA sequence is substantially conserved across diverse mammals [8–10], and that a smaller amount of non-coding sequence is also shared with more distant vertebrates including chicken and even fish [9, 11–13]. Some CNCs show extremely high levels of conservation: for example, Bejerano et al. [9] identified 481 segments longer than 200 bp that are absolutely conserved between the human, rat and mouse genomes. Recent studies of CNCs (using varied definitions) have reported that most CNCs are segments of around 100-300 bp, and that they are widely distributed across the human genome [9, 10, 14–18]. CNCs are not preferentially located near genes [18]. In some cases, clusters of CNCs are found in gene deserts and a subset of these CNCs have been shown to play functional roles as enhancers [19–21].

It has been shown repeatedly that screening for CNCs is an effective method for identifying *cis*-regulatory modules of gene expression [18–25]. CNCs that are shared between humans and distant outgroups such as *Fugu* are heavily overrepresented near developmental regulator genes, and many serve as highly conserved regulators of these functionally conserved genes [13].

That said, there is still considerable uncertainty about the function of most CNCs, and it

has been suggested that some CNCs may serve other kinds of functions, perhaps including roles in chromatin structure or structural connections between chromosomes [26]. In principle, another possibility might be that many CNCs could simply be regions of the genome with low mutation rates. However, two kinds of evidence argue convincingly that the low evolutionary rates of CNCs are indeed due to selective constraint. First, the allele frequency spectrum of human SNPs that lie within CNCs is skewed towards rare variants, consistent with the action of weak purifying selection [27, 28]. Second, the rate of evolutionary change of CNCs is closer to the neutral rate in primates than in rodents [28, 29]. The latter observation is probably due to reduced efficiency of weak purifying selection in primates, which have smaller effective population sizes.

Hence in this study, in view of the likely functional importance of CNCs, we set out to describe the patterns of evolutionary sequence change in these elements. We start with a simple null model in which the evolution of each CNC is characterized by a single substitution rate parameter r that accounts for varying levels of constraint and local mutation rate across CNCs. For each CNC we compare the null model to a hierarchy of alternative models that allow the CNC to have different evolutionary rates in different parts of the phylogeny. In the simplest alternative model the CNC evolves at a single rate across the phylogeny except for one branch, which shows a change in rate (Figure 1). More complex alternative models allow multiple changes in rate. Increases in rate can be interpreted as evidence for positive adaptation or relaxation of functional constraint for the element in question. Decreases in rate are consistent with a tightening of selective constraint.

Two recently published papers [5, 7] have taken similar approaches to identify nongenic regions that show accelerated evolution on the human lineage specifically. Both studies concluded that human-lineage selection signals are enriched near neurological genes. In the study of Pollard et al. [5], the most dramatically accelerated region was found to be part of a novel RNA gene that is expressed during cortical development. Here we expand this kind of approach to look more broadly at evolutionary patterns of CNCs across the mammals.

Results

To scan for functionally interesting CNCs that are shaped by changing selection pressures, we examined regions that are conserved among up to eight vertebrates (human, chimpanzee, dog, mouse, rat, chicken, zebrafish, fugu; see Methods). We started from a publicly available set of aligned regions that were characterized as "most conserved" by the group that maintains the University of California, Santa Cruz (UCSC) genome browser [10, 30]. In short, the "most conserved" regions represent 4.3% of the genome that were identified as conserved by a phylogenetic Hidden Markov Model [10] (Methods). The model used for identifying "most conserved" regions assumes that such regions are conserved across all the species with aligned sequence for the region. However, as we show below, the method was flexible enough to include many regions that show fairly dramatic variation in rates across the vertebrate phylogeny.

We performed extensive filtering of the "most conserved" regions. First, we excluded both translated and untranslated exons, repetitive sequences and sites that are gaps or missing data in any of the five mammalian genome sequences (human, chimpanzee, mouse, rat, dog). We then discarded regions with less than 100 bp of ungapped sequence. The remaining data consisted of 231,285 CNCs spanning $\sim$ 48Mb. The alignments from UCSC make use of global alignment information across species, thus lowering the risk of incorrectly aligning paralogous CNCs as apparent orthologs. However, in order to further reduce the risk of this type of error, we filtered out 98,593 CNCs with human paralogs (see Methods). Since CNCs with different levels of conservation might show differences in their evolutionary patterns, we then subdivided the remaining CNCs into more homogeneous subsets according to conservation levels in chicken and fish (see Methods). Our study examines the properties of the two largest of these subsets, to be denoted as *mammalian CNCs* (conserved within mammals but not found in chicken or fish) and *amniotic CNCs* (conserved in mammals and chicken but not found in fish). For both the mammalian and amniotic CNCs our analysis studied

evolutionary patterns across the history of the 5 mammalian species only.

Our final data set consists of 82,335 mammalian CNCs (for a total of 18.5 Mb) and 16,575 amniotic CNCs (4.6 Mb). The median sizes of CNCs in the two groups are 201 and 240 bp, respectively. We find that overall the amniotic CNCs have a longer length distribution than the mammalian CNCs, consistent with previous results [17]. Further details on the size distribution are in Supplementary Table 1.

Assuming the "Felsenstein 84" substitution model [31], we obtained maximum likelihood estimates of the average numbers of substitutions on each branch of the mammalian tree for each of our CNCs (see Methods and Supplementary Table 3). (All of our analyses assume the phylogenetic tree indicated in Figure 1c [11]). Summing across all branches on the tree, the average number of substitutions per site is 0.16 for amniotic and 0.24 for mammalian CNCs. Notice that, as might be expected, amniotic CNCs show lower overall substitution rates than mammalian CNCs. We estimate that the average substitution rates of our amniotic and mammalian CNCs are $\sim$20% and $\sim$29% of the neutral rate (based on comparison to local unconserved sequences). Overall, our CNCs are more conserved on average than the original set of "most conserved" regions identified by Siepel et al. [10], which averaged $\sim$ 33% of the unconserved rate. This difference indicates that our filtering process preferentially retains more highly conserved elements.

We also examined the location of CNCs with respect to nearby genes. For each CNC, we computed the distance to the nearest gene without considering gene orientation. 37% of the mammalian CNCs are in introns, and the remainder are intergenic. Among intergenic CNCs, 10% are within 10Kb of a gene, 27% are between 10Kb and 100Kb and 26% greater than 100Kb from any gene. The amniotic CNCs have a similar overall distribution in the genome, although they are significantly more clustered (Supplementary Table 2, Supplementary Figure 2). Overall, we find that CNCs are distributed across the genome approximately at random with respect to the locations of nearby genes (Supplementary Table 2), as noted

previously [18].

Analysis of the relationship of CNCs with PANTHER gene ontology categories [32] shows that genes related to developmental processes are significantly enriched near CNCs (1.5-fold enrichment, $p < 10^{-21}$), as seen previously [9, 10, 13, see Methods, Supplementary Tables 4 and 5]. The genes in the 'signal transduction' and 'nucleoside, nucleoside and nucleic acid metabolism' categories are enriched near mammalian and amniotic CNCs, respectively (1.2-fold enrichment, $p < 10^{-11}$; and 1.3-fold enrichment, $p < 10^{-7}$). Olfaction genes are $\sim$ 15-fold underrepresented in our data set, presumably because olfactory genes tend to be highly duplicated and our filtering process removes duplicated CNCs.

The Shared Rates Test. To identify CNCs that have been targets of selection, we introduce a likelihood ratio test that we call the "Shared Rates Test" (SRT). Under the null model, the divergence times of lineages are shared across CNCs, but each CNC may evolve faster or slower according to its local mutation rate and level of evolutionary constraint. For each CNC, we test whether any branches are surprisingly long or short compared to the others, indicating speed-ups or slow-downs of the substitution rate. For example, in Figure 1, the first two trees evolve at different rates, but with the same tree "shape" (i.e., the ratios of branch lengths are the same). In contrast, the third tree has a longer-than-expected branch on the human lineage, suggesting the action of natural selection.

In our model, each branch of the mammalian tree has a branch-length parameter v_b, defined as the average number of substitutions per site on branch b for CNCs evolving under a constant level of constraint. (Here v_b is defined as the average number per site across all CNCs.) In addition, under the null hypothesis, each CNC is associated with a single rate parameter $r_0^{(h)}$ (where h indicates a particular CNC). Then the number of substitutions that occur in CNC h, on branch b has an expectation at each site of $N_{b,h}$, where

$$N_{b,h} = v_b r_0^{(h)}. \tag{1}$$

Under the null model, there are 7 branch length parameters for the tree that we consider, and one additional rate parameter for each CNC. As described in the Methods and Supplementary Methods, we obtain a joint maximum likelihood estimate for all the parameters, assuming the "Felsenstein 84" model of sequence evolution [31].

Our model is designed so that all CNCs have the same expected tree shape (i.e., the ratios of expected branch lengths are the same). However the total size of the tree is allowed to vary according to $r_0^{(h)}$, in order to reflect variation in mutation rates and the level of selective constraint across CNCs. In addition, we place no constraints on the relative values of the v_b, so that lineage-specific variation in mutation rates (such as the higher substitution rate in rodents) is reflected in longer estimates for those branch lengths (Figure 1, Supplementary Figure 1). In summary, the null model allows mutation rates and levels of constraint to vary across CNCs, and it allows for the property that broad-scale mutation rates may vary across lineages.

In addition to the basic null model, we consider a family of alternative models that allow additional rate parameters for particular CNCs. In the simplest alternative, a single branch on the tree evolves at a rate that is different from the background rate shared by the remaining lineages (as for the third tree in Figure 1). In the extreme alternative, each of the seven branches evolves with its own rate $r_i^{(h)}$, giving a total of seven rate parameters for the CNC in question. (For simplicity of notation, we will henceforth drop the notation h on the rate parameters.) In the extreme case, to test the hypotheses $H_0 : r_1 = r_2 = \cdots = r_7 \ (= r_0)$ vs $H_A : r_1 \neq r_2 \neq \cdots \neq r_7$ at a particular CNC, we compute the SRT as

$$SRT = -2 \log \frac{L(\widehat{r_0})}{L(\widehat{r_1}, \cdots, \widehat{r_7})} \ , \tag{2}$$

where L is the likelihood of the sequence data for the five mammalian species, maximized with respect to the rate parameters, and with the fixed estimates of branch lengths param-

eters $(\widehat{v_1}, \cdots, \widehat{v_7})$ and the sequence evolution model. Large values of the SRT indicate a substantially better fit of the alternative than the null model. Another example of alternative model is the case in which branches 2 and 3 have distinct rates r_2 and r_3, while the other branches have a single "background" rate $r_{0,-2,-3}$. In this case, to test the hypotheses $H_0 : r_1 = r_2 = \cdots = r_7 \ (= r_0)$ vs $H_A : r_2 \neq r_3 \neq r_1 = r_4 = \cdots = r_7 \ (= r_{0,-2,-3})$, we can compute the likelihood ratio statistic as

$$SRT = -2 \log \frac{L(\widehat{r_0})}{L(\widehat{r_2}, \widehat{r_3}, \widehat{r_{0,-2,-3}})}. \tag{3}$$

In this paper, we perform two kinds of analyses. One analysis performs model selection using the SRT, while the other tests for individual branches with rate changes. When testing for a rate change on the ith branch only, it is convenient to transform the likelihood ratio statistic as follows. In this case we will use special notation, denoted by SRT_i:

$$SRT_i = \text{sign}(r_i - r_{0,-i}) \times \sqrt{-2 \log \frac{L(\widehat{r_0})}{L(\widehat{r_i}, \widehat{r_{0,-i}})}}, \tag{4}$$

where $\text{sign}(x) = 1$ if $x > 0$ and otherwise $\text{sign}(x) = -1$. Rewriting the SRT in this way provides the convenient property that $SRT_i > 0$ implies that r_i is larger than the background rate $r_{0,-i}$ and hence branch i shows a rate speed-up relative to the rest of the tree; conversely $SRT_i < 0$ implies a slow-down on branch i. As a convention, when we subscript SRT by a character or number, it will represent the signed likelihood ratio statistic testing for rate changes on the indicated branch. Otherwise, the notation SRT without subscripts will be used to indicate use of an unsigned test statistic, in the form of Equations 2 and 3.

Our SRT is a likelihood ratio test and, as such, standard theory suggests that under the null hypothesis the test statistic should asymptotically follow a chi-square distribution with degrees of freedom equal to the difference in the number of estimated parameters between the constrained (null) and less-constrained (alternative) models. Similarly, the signed root of

this statistic for a one-dimensional parameter of interest is asymptotically standard normal. Therefore, when the null hypothesis is true and the number of sites in a CNC is large enough, the unsigned SRT might be expected to follow the chi-square distribution with the degrees of freedom equal to the difference in the number of rate parameters between the two models. For example there are 6 degrees of freedom in the global test (Equation 2) and 2 degrees of freedom in the example in Equation 3. Similarly, under the null, the signed test SRT_i is constructed to have a standard normal distribution as the CNC size goes to infinity. Our simulation studies show that the asymptotic theory is reasonably accurate for both versions of the test statistic, except in the cases in which the lineages tested for selection are relatively short and are expected to accumulate few substitutions (namely, the human and chimpanzee lineages; Supplementary Figure 3). Hence, to reduce computational burden we calculate p-values using the asymptotic chi-square or normal approximations, except for tests on the human and chimpanzee branches for which, except where stated, we compute p-values based on the empirical null distribution in simulated data (see Methods).

An additional consideration is that we do not want the estimated null branch lengths (v_b) to be heavily influenced by outlier CNCs with evidence for selection. To mitigate the impact of such CNCs, we first identify CNCs with clear overall departures from the null model ($SRT > 25$ in the global 6 degree of freedom test, corresponding to p< 0.00034) and then re-estimate the branch lengths after dropping those non-neutral CNCs, which represent 2.8% and 3.8% of the total mammalian and amniotic CNCs, respectively.

In summary then, our analysis performs the following steps:

- Estimate maximum likelihood branch lengths and rates under the null.
- Identify outlier CNCs that have $SRT > 25$ comparing the 7- and 1-parameter models.
- Drop outlier CNCs and recalculate the null branch lengths and rates.
- Compute the shared rates test statistics for each CNC according to a range of alterna-

tive models.

For reasons discussed below, in practice these analyses were performed in a sliding window of 50 consecutive CNCs, as defined by position in the human physical map. All analyses considered the mammalian and amniotic CNCs separately.

Accounting for local variation in tree shape. It is well-established that the extent of divergence between mammalian species varies substantially across large genomic regions [33–38]. For example, Gaffney and Keightley [38] showed that divergence between the mouse and rat genomes varied between and within chromosomes. While the causes and the scales of this type of variation are not completely understood, it has been shown that divergence correlates with various genomic features including GC and CpG content, simple-repeat structures and recombination rate, suggesting that these genomic features drive variation in mutation rates [35, 37].

Variation in mutation rates or levels of CNC conservation across genomic regions should not be problematic for our method, provided that the substitution rate in any given region maintains a constant ratio to the average across the mammalian phylogeny. If a CNC is in a region with a higher, or lower, mutation rate than average, this effect should simply be absorbed into the rate parameter that we estimate for each CNC as part of our null model. However, if mutation rate variation is not stable across the phylogeny, this might produce false signals for our method.

Therefore, we looked at whether the average tree shapes are significantly variable across chromosomes (according to the human physical map) as well as within chromosomes. We found that in fact there is nontrivial variation in tree shape both at the chromosome level, and across genomic regions within chromosomes. For example, within chromosome 2 there is a highly significant auto-correlation in the fraction of the tree occupied by the mouse lineage (Figure 2). This result implies that local variation in large-scale mutation rates is not conserved across evolutionary time: for instance, genomic regions that evolve faster than

average on some lineages may evolve slower than average elsewhere on the tree.

If average tree shapes were constant across the genome, we could use CNCs from across the genome to estimate the tree shape for our null model. However, the observation that tree shape is not constant suggests that instead our model should allow for variation in tree shape across the genome. After some experimentation, we settled on using a sliding window of 50 consecutive CNCs to estimate the tree shape. That is, we test each CNC for significant departures from the tree shape in a 50-CNC window that, in the human physical map, is centered near the CNC in question (see Methods). On average, this window size corresponds to 525 Kb and 1.3 Mb (median) for mammalian and amniotic CNCs, respectively.

Overall, we find that using the sliding window method produces only a modest impact on the rate of significant CNCs but it should improve our inferences by taking account of the local variation in tree shapes (Figure 2, Supplementary Figure 4). An obvious concern about using a sliding window based on the locations of CNCs in humans is that due to chromosomal rearrangements, CNCs that are close together in humans may not be close together in other mammals. Consequently a sliding window based on the human map might not provide a suitable correction. Fortunately, our window size is relatively small compared to the typical size of syntenic blocks [8, 39] and in Figure 3, we show that the results of tests on the human lineage are highly concordant whether we use windows based on the human, or mouse physical maps, and indeed are only modestly different from the results using all CNCs together. Consequently, all subsequent results use 50-CNC windows based on the human map.

Variation in tree shape due to varying constraint. Another plausible concern about our model stems from the prediction that selection against weakly deleterious mutations is more efficient in species with large populations than in small populations. This means that weakly constrained sites in CNCs are likely to evolve more quickly in primates than in rodents (which have larger effective population sizes). This effect has been observed

in a comparison between the evolutionary rates of CNCs and putatively neutral flanking sequences [29]. Hence–in contrast to our null model–one might expect the overall tree shape for a CNC to depend on its level of selective constraint.

To investigate this issue, we classified CNCs into four different levels of conservation, according to their substitution rates on the dog lineage. We then compared the average human to chimpanzee divergence against the average mouse to rat divergence, separately within each of the four conservation levels (Supplementary Table 15). We find that that as the level of constraint increases, the divergence in rodents indeed decreases faster than divergence in hominids, consistent with the results of Keightley et al. [29]. However, we find that the variation across CNCs is relatively small (less than 11% change across different classes of CNCs) and much less than when CNCs are compared to neutral sequences (Supplementary Table 3). As shown below, we do not have power to detect such small variations in tree shape, so we conclude that it is not necessary to control for overall conservation level more carefully for the current study.

Analysis of branch-specific rate changes. For each CNC, we calculated SRT_i for each of the seven branches of the mammalian tree to identify CNCs that have experienced a speed-up or slow-down on a particular branch. Figure 4A shows the histogram of p-values on the mouse lineage (SRT_m) for the mammalian CNCs. The p-values are defined as $P(SRT_i > srt_i)$ where srt_i is the observed value. Hence, p-values near 0 indicate increased rates, and near 1 indicate decreased rates. The histogram is flat for intermediate p-values with peaks at both ends, suggesting that most CNCs fit the null distribution of SRT_m, but with a substantial number of outliers. At the significance level of 0.001, 1027 (1.2%) and 503 (0.6%) mammalian CNCs show speed-ups and slow-downs, respectively. Among amniotic CNCs, 228 (1.4%) and 106 (0.6%) show speed-ups and slow-downs, respectively on the mouse lineage.

Figure 4B plots the expected and observed branch lengths on the mouse lineage for the

CNCs that are significant at $p < 0.001$ in each tail. (Similar plots for other lineages are shown in Supplementary Figure 5.) The red points above the diagonal indicate CNCs with rate speed-ups. For the central 95% of the significantly fast-evolving CNCs, the observed branch lengths are between 0.04 to 0.13 substitutions per site, and are 2-4 fold higher than the expected branch lengths. The blue points below the diagonal are CNCs with reduced branch lengths. Nearly half of these CNCs accumulated no substitutions on the mouse lineage.

The other long lineages show similar p-value histograms though with some variability in the proportion of significant CNCs. The dog lineage is the most enriched for signals, with 2.3% and 1.9% of mammalian CNCs showing speed-ups and slow-downs, respectively, at $p < 0.001$ (in each tail). Even after a stringent Bonferroni correction, 186 and 46 CNCs, respectively, are still significant at p=.001 in the dog lineage. The overall results for amniotic CNCs are similar, but the fraction of significant results is slightly higher on each branch (Supplementary Table 8). For most lineages, our significance threshold (one-sided p-value < 0.001 on each end) corresponds to a genome-wide false discovery rate between 0.05 and 0.1 (Supplementary Table 9).

Since the distribution of SRT_i on the human and chimpanzee lineages does not follow the standard asymptotic distribution, we simulated data under the null over a range of substitution rates that cover the observed range over all 50-CNC windows (see Methods). We account for heterogeneity in the distribution of SRT_i across bins of CNCs with different numbers of expected substitutions on the tested lineage by computing p-values based on the empirical null distribution of SRT_i constructed in each bin (figure not shown). At a significance level of 0.001, 256 mammalian CNCs and 59 amniotic CNCs show rate speed-ups on the human lineage (Supplementary Table 8). Note that there is little power to detect rate reductions on these very short lineages.

To better understand these SRT_i results, we performed power simulations under a range of models. The simulation results, summarized in Supplementary Figure 6, show considerably

greater power to detect speed-ups than slow-downs on all lineages, consistent with the results of Siepel et al. [40]. Thus, the fact that we detect more speed-ups than slow-downs does not necessarily imply that speed-ups are actually more common, and it is likely that many slow-down events are simply not detected by our analysis.

Human accelerated regions. Our human results allow a comparison to the "human accelerated regions" (HARs) identified by Pollard et al. [5] using a similar type of approach, based on regions that were highly conserved (at least 96% identity) across chimpanzee, mouse and rat. Among the top 49 HARs, which include 2 coding regions, 34 overlap with CNCs in our dataset; however generally the HARs are considerably shorter and more conserved and lie within our CNCs. Perhaps not surprisingly, since the HARs are the top genome-wide hits in their data, the signals in our overlapping CNCs tend to be weaker. Among the 34 CNCs, just 5 CNCs are significant in our analysis at a genome-wide false discovery rate less than 0.05. Nonetheless, our CNCs that overlap HARs do show a strong enrichment of modest signals. Our human lineage p-values are < 0.01 for 26 of the 34 CNCs overlapping HARs, and are < 0.1 for 33 of the 34 (Supplementary Table 10).

One of the most significant CNCs on the human lineage in our dataset is a 144 bp amniotic CNC located on human chromosome 21 starting at 33,481,809 (q22.11, NCBI Build 35) that was not detected by Pollard et al. [5] because it fails their filtering threshold for chimpanzee-mouse-rat similarity. As illustrated in Figure 5, the posterior expected number of substitutions (see Methods for details) on the human lineage is 5.2, which is 26-fold higher than the value of 0.2 expected under the null model. The corresponding SRT_h is 4.84. The p-value for this CNC is so small that it is difficult to evaluate by simulation, however the standard normal approximation suggests that $p \approx 6 \times 10^{-7}$ (our simulations indicate that this is conservative). In addition to the five nucleotide substitutions, there is also a 2 bp insertion on the human lineage that was not included in the statistical inference. Since the UCSC genome browser database was recently updated, we were able to inspect an alignment

of 17 vertebrate species for this region. Manual inspection confirmed that all six of these substitutions occurred on the human lineage.

The function of this CNC is unclear but the two nearest genes are C21orf54, 17 Kb upstream, and IFNAR2, 42 Kb downstream of the CNC. Not much is known about C21orf54 but IFNAR2 codes for a type I membrane protein that forms one of the two chains of a receptor for interferons alpha and beta [41]. This CNC is strongly conserved among the other mammalian species and chicken but does not appear to be present in the fugu genome. In addition to the rapid evolution on the human lineage, there is weak evidence for slower evolution of this CNC on the mouse and dog lineages (one-sided p-values =0.011 and 0.023, respectively; see Figure 5B).

Classification of CNCs according to evolutionary patterns. Thus far we have focused on the simplest class of alternative models, in which a CNC changes substitution rate on a single branch only and has a constant background rate elsewhere on the tree. We now extend this approach in order to classify each CNC according to a family of more complicated models of evolutionary patterns.

Our data are connected by a tree containing seven branches. The simplest model (our "null") has a single rate parameter, and the most complicated alternative model has seven different rate parameters. In between, there are 876 ways of partitioning the seven branches into two or more different substitution rate groups. However, considering all of these partitions does not seem biologically meaningful or necessary, and here we focus on a reduced set of 126 alternative candidate models.

The alternative models we consider can be divided into two distinct classes of models. In one class of models, each tree is assumed to have a "background" rate parameter. Then, each CNC may have between 1 and 6 "selected" lineages, and each selected lineage evolves at its own rate. In the other class of models, each tree may be split into subtrees that share a single rate, while the rest of the tree has a single background rate (for full details, see

Supplementary Table 11).

We use a modified AIC (Akaike Information Criterion) procedure to classify each CNC into its best model. In brief, the method attempts to account for multiplicities of alternative models as well as the number of estimable parameters in each model (see Methods). We have performed simulations to test the performance of this method, and we find that it provides suitable control over the rate of "false positives" (i.e., accepting models with more parameters than used to simulate the data). That said, our simulations show that it is often difficult to correctly classify complex models with multiple rate changes (see Methods, Supplementary Figure 7).

The results of our data analysis are summarized in Figure 6 and Supplementary Table 12. We estimate that $\sim$ 68% (54643/81957) of the mammalian CNCs evolve at a single rate. The remaining non-neutral CNCs show rate changes on at least one lineage. The number of CNCs assigned to each model category decreases with increasing model complexity. Among the 32% of CNCs with more than one rate, $\sim$ 75% (20420/27314) exhibit rate changes on a single lineage but not on the remaining lineages and $\sim$ 9% (2419/27314) exhibit rate changes on the primate or the rodent lineage that are inherited across all branches below. For the two-parameter models, the rate change events are easily classified as speed-ups or slow-downs. Counts for both types of event are shown in Figure 6B. For most lineages, there are slightly more speed-up events than slow-downs ($\sim$ 55% vs $\sim$ 45%). However, there are 638 and 530 CNCs that show rate speed-ups on the human and chimpanzee lineages, respectively, far more than the 4 and 8 CNCs, respectively, showing slow-downs. Presumably, these results are due in large part to the greater power to detect speed-ups, as well as differences in power across lineages (Supplementary Figure 6).

It is notable that the dog lineage shows a very large number of rate changes, which may not be fully explained by the long length of this lineage (second longest among the seven). Since there is no strong tendency towards an excess of speed-ups over slow-downs on this lineage,

it is unlikely that this can be explained by occasional CNCs with low-quality dog sequence. Perhaps a hint is that we have observed greater variation in the dog-lineage substitution rates at neutral sites than on other lineages. Perhaps there is greater fine scale variation on the dog lineage that is not well captured by our 50-CNC window method (see Supplementary Methods, Supplementary Figure 8).

Fast evolving CNCs that exceed neutral rates. As discussed above, we have identified many CNCs with significantly accelerated rates on one or more branches. However, it is unclear *a priori* whether these speed-ups reflect positive adaptation or relaxation of functional constraint. In order to address this issue, we estimated substitution rates in unconserved sequences near each CNC to estimate local neutral rates (see Methods). We then determined how many of the CNCs showing rate speed-ups have an accelerated rate that actually exceeds the corresponding lineage-specific neutral rate. If the rate in a CNC actually exceeds the local neutral rate, this is strong evidence for adaptive evolution. However, a negative result here is difficult to interpret, since adaptive evolution in an otherwise slow-evolving sequence may not necessarily bring the total rate above the neutral background rate.

Our results are summarized in Table 1. We observe that most CNCs showing accelerations on the human and chimpanzee branches indeed have rate estimates exceeding the neutral rates; of these, more than half are actually significantly faster than the neutral rate at $p < 0.05$. Meanwhile, the other branches of the mammalian tree all show smaller fractions of CNCs with rates that exceed the neutral rate, and very few of these are significantly faster than the neutral rate. One plausible explanation might be that if there is sufficiently rapid evolution on a long branch, this might cause an otherwise conserved element not to be classified as a "most conserved" region by the HMM model [10]. However, some simple calculations suggest that this is likely to be a modest effect in practice. Moreover, we see the same effect for both the mammalian and amniotic CNCs (Table 1), even though the HMM data for the latter include the relatively long branch to chicken, and should therefore

be much less susceptible to this effect.

Instead, to explain these observations, we hypothesize that the rate speed-ups that we detect may often reflect rapid bursts of adaptation in which a CNC accumulates a series of sequence changes, thus modifying its function. A single burst of adaptation may produce enough sequence changes to exceed the neutral rate on a short branch, but not on a longer branch. In this model, we would have the most power to detect adaptive events on short branches. Our data argue strongly against a model in which a CNC adapts continuously over extended periods of evolutionary time, as such a model should also produce signals on the long branches.

Relationship between fast-evolving CNCs and nearby genes. We have also performed analyses of the locations of CNCs showing branch-specific speed-ups, with respect to nearby genes. A recent report by Drake et al. [27] found that the frequency spectrum in CNCs is most skewed towards rare variants (indicating weak purifying selection) in introns and near genes, and is less skewed in CNCs that are far from genes.

To test whether CNCs showing speed-ups on particular branches occur at higher rates near to, or far from genes, we divided all our CNCs into four classes: intronic, within 10 Kb of a gene, between 10 Kb and 100 Kb, and greater than 100 Kb from any gene. We found that on the mouse and rat lineages, CNCs showing speed-ups ($p < .001$ on the branch-specific test (SRT_i) occur at higher rates in introns and within 10 Kb of genes than among CNCs further from genes. However, this trend was not replicated on the other lineages of the tree (Supplementary Table 13).

We next looked at whether CNCs showing significant rate speed-ups are more likely to be in the proximity of particular kinds of genes [17], using the PANTHER Gene Ontology database [32]. A significant difficulty in this sort of analysis is that even for those CNCs that act as *cis*-regulators, it is unknown *which* of the nearby genes is being regulated. However, as a rather imperfect proxy for this we simply used, for each CNC, the nearest gene (in

either orientation). For each branch of the mammalian tree, we divided the CNCs into those with increased rate on that branch (by AIC) and used CNCs evolving under the null model as "neutral" controls. We looked at whether particular biological process categories were enriched among the nearest genes of the selected CNCs compared to the neutral CNCs.

For mammalian CNCs, there is significant enrichment of the process categories "amino acid activation", and "other coenzyme and prosthetic group metabolism" on the dog and the lineage leading to the common ancestor of mouse and rat (rodent lineage), respectively at $p < 0.05$ after Bonferroni adjustment. We also tested whether any categories show repeated evidence for enrichment on different branches of the tree. For mammalian CNCs, the "sensory perception" category appears in the top ten enriched biological processes for three out of the seven lineages. However in summary, we view these gene ontology associations as rather tentative, since none of them is highly significant or highly repeatable across branches of the tree. Complete results from this analysis are presented in Supplementary Tables 6 and 7.

Discussion

Our paper presents a new approach to studying the evolutionary patterns of CNCs. We find that a large fraction of CNCs ($\sim$ 32%) do not fit a simple model of evolution with a consistent substitution pattern across the mammalian tree. Among those CNCs that do not fit our null model, $\sim$75% show changes in evolutionary rate on a single branch of the mammalian tree, while the remainder have more complex substitution patterns. In many cases–particularly on the short branches of the phylogeny–CNCs with rate accelerations on a particular branch significantly exceed the neutral rate on that branch, suggesting that the changes are driven by adaptive evolution. The less extreme speed-ups may be due to either adaptation or a relaxation of selective constraint; however we suggest that much of our signal on the longer branches may be due to short bursts of adaptation that do not generate enough changes to exceed the total neutral rate on a long branch.

A very recent paper by Galtier and Duret [42] argues that recently reported human accelerated regions (HARs [5]) are most likely the result of biased gene conversion (BGC). One of the main characteristics of BGC is an excess of AT $\rightarrow$ GC transitions. In some of our CNCs showing accelerations on the human lineage, we also observe this transition bias, which seems to be larger with increased acceleration signals. However, for most fast-evolving CNCs, the numbers of AT $\rightarrow$ GC changes roughly match the distribution expected based on the overall distribution across random CNCs (Supplementary Figure 9). In summary, these data suggest that some of the fast-evolving CNCs may in fact be due to BGC, however that most fast-evolving CNCs do not show the signal expected for BGC.

Overall, our results imply that either the levels of functional constraint or the functional roles of CNCs are reasonably changeable across the timespan of mammalian evolution. Though it lies beyond the scope of this paper, it will be of interest to use experimental approaches to probe the functional significance of the many CNCs that we have identified as having had bursts of rapid evolution [5, 25].

Of course, in this type of study, there are inevitably features of the real data that are not fully accounted for in the models. We believe that our results should be reasonably robust to these issues, however, as follows. One natural concern is that our CNC alignments might occasionally align paralogs. This is a serious concern in principle, however we have aimed to aggressively filter out CNCs with related paralogs to minimize this effect, in addition to making use of global alignments. Other model departures might inflate the variance of branch-specific substitution rates. These include the possibility of fine-scale, branch-specific changes in mutation rate, as well as variation in the branch lengths of the human and chimpanzee branches due to coalescent time variation [43]. On the whole these effects are likely to be fairly modest since the observed rate changes are usually not significant unless they are quite dramatic (significant rate changes are usually $\sim$2-4 fold on the mouse lineage, and larger on the shorter branches). For this reason, the analysis that uses a single global tree shape produces fairly similar overall results to the window-based analysis, despite evidence that the window-based analysis fits the data better (Supplementary Figure 4). A related concern is that due to variation across lineages in effective population size, the evolutionary rates of CNCs with different levels of constraint might not scale linearly across the trees [29]. However, our data show that this is a modest effect relative to the size of change needed to produce a significant rate change in a CNC (Supplementary Table 15).

In this study, we aimed to classify CNCs according to their evolutionary patterns. To do so, we used a modified version of the AIC to find the model that best describes the pattern of evolution of each CNC. In order to reduce the space of alternative models, we restrict our alternatives in two classes of models. As we obtain genome sequences for increasingly more species, it will be worth revisiting these models, as we will be better able to distinguish among different modes of evolution [40]. In particular, two natural models for rate-changes in a CNC are (1) that the CNC has a one-time change in evolutionary pattern (for example a burst of adaptation to acquire a new function), or (2) that the CNC changes function or evolutionary constraint in a way that is inherited across all branches below. With larger

numbers of taxa, it should be possible to gain better insight into the relevance of these two possible modes of evolution. More broadly, as we obtain increasing information about the functions of CNCs, we will increasingly be able to interpret the biological relevance of the patterns of rate changes detected here.

Materials and methods

Constructing the raw database of CNCs. We downloaded the genome-wide multiple alignment of 8 vertebrate species (human (hg17, May. 2004), chimpanzee (panTro1, Nov. 2003), dog (canFam1, Jul. 2004), mouse (mm5, May 2004), rat (rn3, Jun. 2003), chicken (galGal2, Feb. 2004), fugu (fr1, Aug. 2002), zebrafish (danRev1, Nov. 2003)) from the University of California, Santa Cruz (UCSC) genome browser [30]. We also downloaded the annotation of "most conserved" regions defined on the same multiple alignment by a phylogenetic Hidden Markov Model on June 2005 [10]. The most conserved regions are defined without regard to whether the sequence is coding or non-coding, and cover around 4.3% of the human genome. To define CNCs, we first extracted those conserved regions from the multiple alignment and then processed them by removing coding regions (exons in the 'known gene' annotation, UCSC genome browser), repetitive sequences (marked by lower case letters in the alignment), and sites that are gaps or missing data in any of the five mammalian genome sequences. Conserved regions of less than 100 bp after the processing were discarded. The remaining 231,285 regions out of the initial 1,451,896 most conserved regions comprised our raw dataset of CNCs and spanned $\sim$ 48Mb.

We used BLAT [44] to exclude spuriously aligned CNCs. We restricted our data to unique CNCs in which the human version of a CNC does not find any similar sequence (>50% sequence identity) elsewhere on the human genome. This resulted in discarding 24,234 CNCs ($\sim$ 5.4Mb). Furthermore, we required that each non-human mammalian verion of a CNC find the human version as the best match when it is BLATed against the whole human genome. This resulted in discarding an additional 74,359 CNCs ($\sim$ 10.7Mb).

Our statistical inferences are based on alignments of the 5 mammalian sequences. However, we used the aligned chicken and fugu sequences to classify CNCs into different conservation level groups, of which we analyzed the two largest, denoted as "mammalian" and "amniotic" CNCs. Roughly speaking, a CNC was classified into the mammalian group if it is conserved

across the mammalian genomes but not chicken or fugu and into the amniotic group if it is conserved among the mammals and chicken but not fugu. The classification depended on (1) the presence or absence of aligned chicken and fugu sequences, and (2) the mean identity between mammals and chicken and fugu. The details are given in the Supplementary Methods.

Sliding window analysis. To examine the scale of local variation in tree shape, we estimated tree shapes over a chosen set of window sizes of 10, 30, 50, or 100 consecutive CNCs (ordered according to the human genome position). Since the scale of local variation seems to vary across chromosomes, there is no clear boundary explaining the rate of decay of auto-correlations. Nonetheless, incorporating such variation in our model is important, since otherwise, regions that show a general pattern of evolution that departs from the shared pattern from all CNCs might produce clusters of spurious signals. After several trials, we decided to estimate tree shape using a sliding window of 50 CNCs with an overlap of 34 CNCs between successive windows. With this window size, we obtained enough data to stably estimate tree shapes, but were able to capture much of the local variation. The one third of CNCs located in the center of each window use the estimated tree shape from that window. In order to reduce the effect of outliers (defined as having $SRT > 25$ with degrees of freedom of 6), we estimate branch lengths in each window, drop outliers, and then re-estimate the divergence times after dropping those non-neutral CNCs. Through this procedure, we expect that our estimates are robust in the presence of outlier CNCs. However, when rate changes are spatially clustered, our locally estimated tree shapes may absorb some of the signal of variable rates, hence potentially reducing power.

Chimpanzee sequence quality control. Our preliminary analysis of classifying CNCs using the modified AIC showed that the number of CNCs with signals on the chimpanzee lineage was 48% larger than on the human lineage. Closer examination indicated that

often CNCs with low quality chimpanzee sequence (PanTro1) produced a large signal of rate changes on the chimpanzee lineage since miscalled bases would appear as mutations. Therefore, we dropped any CNC that is classified by AIC into the group showing rate changes on the chimpanzee lineage but that has low quality chimpanzee sequence.

To identify those CNCs, the chimpanzee sequence in each CNC was BLATed to the chimpanzee genome (PanTro1). The best match position (according to the BLAT score) was found when it was available. Then, in the target region, we counted the number of sites that have low quality score (≤ 20). If this count was larger than 15, we considered the CNC to have low quality chimpanzee data. A total of 378 mammalian and 89 amniotic CNCs that were significant on the chimpanzee lineage were dropped for this reason.

The impact of occasional sequence errors is likely to be much smaller for the other species. The human genome sequence has very high accuracy (the estimated error rate is 1 site per 100 Kb, much lower than the human polymorphism rate [45]). Meanwhile, occasional sequence errors in the other species should have only a small effect due to the much longer branches leading to those taxa.

Likelihood computation and parameter estimation. We estimated branch lengths for an alignment using the "Felsenstein 84" sequence evolution model and using the empirical base frequencies. To make computation feasible, the 'peeling' algorithm [46] was used with the assumption that sites evolve independently and that given their common ancestor, branches evolve independently. Details of the evolution model and the 'peeling' algorithm were described by Felsenstein and Churchill [31]. Note that there are many more general evolutionary models, but the Felsenstein 84 model, which is essentially the same as the HKY85 model [47], seems to be sufficient for our purposes of study [48].

Under the null, our parameters are a set of seven branch lengths shared by all CNCs, and one additional local substitution rate for each CNC. Under an alternative, our parameters are a set of lineage-specific rates that explain a specific scenario for each CNC. Rather than

maximizing the likelihood directly, we developed an EM algorithm that efficiently maximizes many parameters jointly under the null model. The details are given in the Supplementary Methods, but essentially, in our EM, each branch length is updated sequentially by computing the posterior number of substitutions on each branch and updating the related parameters accordingly. We find that our EM algorithm is stable to choices of initial starting estimates. The estimates that we obtain for simple models match well with those computed by Phylip [49] and PAML [50].

Classification of evolutionary models using a modified AIC procedure. There are many possible models of CNC evolution, ranging from the simplest case where there is a single rate across the entire tree, to the most extreme case where each lineage evolves with its own rate. Here we address how to classify CNCs according to their evolutionary patterns.

Each of the possible alternative models corresponds to a partition of the seven lineages into two or more blocks of substitution rates. There are 876 ways of partitioning the seven branches into two or more different substitution rate groups. However to reduce the space of possible models, we restrict ourselves to a subset of 127 candidate models that seem biologically most natural. Our main class of alternative models consists of the models where there are k selected branches ($1 \leq k \leq 6$), each with its own rate parameter, while the remaining branches share a single background rate parameter. Such models have $k+1$ parameters, and there are $\frac{7!}{(k!)(7-k)!}$ such models for $k = 1, \cdots, 5$ and one additional model for $k = 6$. This accounts for 121 candidate models.

In addition, we also consider a further set of six models that seem biologically natural, that split the branches in an unrooted tree into two or three rate groups using an internal branch (connecting two internal nodes). Thus for example, we might hypothesize a single rate-changing event in the ancestor of mouse and rat that leads to a single altered rate on both the mouse and rat branches. To reduce the model space complexity, we assume that such rate change events occur at internal nodes on the tree. These six models are summarized

in Supplementary Table 11.

Since there are many possible models, correct classification of the CNCs is likely to be difficult. Here, we view the classification as a multiple testing problem rather than a model selection problem, where our first goal is to control the rate of over-estimating the number of model parameters. The scheme below, though *ad hoc*, provides a reasonable compromise in providing fairly good model choice while not having excessive rates of "false positives".

For each CNC, we select the model that, among the 127 candidate models produces the highest value of the penalized likelihood, which is $log(L) - (k+1) + log\{(7-k-1)!\}$ for our main class of alternative models, where L is the maximum likelihood and k is the number of selected lineages. The first penalization term $(k+1)$ penalizes for the number of estimated parameters and is introduced for the same reasoning as in the standard AIC. The last term $(log\{(7-k-1)!\})$ aims to account for the multiplicity of different models within each level. This latter term was suggested previously as a prior weight for Bayesian classification in an analogous setting [51]. This term is motivated by thinking of each model as corresponding to a partition of the seven branches into one or more blocks of substitution rate groups; as a natural choice of partition distribution we use the Ewens sampling distribution [52] with concentration parameter λ of 1 (see Supplementary Methods).

To evaluate the performance of the classification, we simulated data under the full range of null and alternative models. We used these simulations to compare among three possible choices of penalty functions: (1) using only the number of parameters (AIC), (2) using only the Ewens prior (Ewens), and (3) using both the number of parameters and Ewens prior (AIC+Ewens), as detailed above. The penalty function computed for each model is summarized in Supplementary Table 14 and the power simulation results are shown in Supplementary Figure 7. The AIC+Ewens penalization provides the best control against over-fitting and that is what we use for our data analysis.

Estimation of lineage-specific neutral rates. There are 6037 mammalian and 1497 amniotic CNCs that show branch-specific accelerations on at least one lineage at significance level of 0.001 (based on the asymptotic distribution of SRT_i). To see if these CNCs actually exceed the neutral rate on a particular branch showing speed-ups, we estimate the local "neutral" tree near each of those CNCs. Specifically, we take the surrounding 10 Kb with each such CNC at the center, then exclude "most conserved" regions as well as exons to construct putatively neutral local regions. The genome-wide average of each branch length on trees estimated from those regions (shown in Supplementary Table 3) is very similar to that from CFTR non-exonic DNA region in Cooper et al. [15]. It is notable that the variation in neutral branch lengths of the dog lineage is considerably larger than that on other lineages (Supplementary Figure 8).

To test whether the accelerated rate exceeds the neutral rate on a particular branch i, we compute a likelihood ratio statistic testing the hypothesis $H_0 : r_{0,-i},\ r_i = r_{i,\text{neutral}}$ vs $H_A : r_{0,-i},\ r_i > r_{i,\text{neutral}}$ only for those CNCs that have a substitution rate on the tested lineage r_i that exceeds the neutral rate $r_{i,\text{neutral}}$ in surrounding region. Based on the chi-square distribution with 1 degree of freedom, we compute p-values and reject the null hypothesis if p-value < 0.05. Our results are summarized in Table 1.

Simulations. We performed simulation studies with the "Felsenstein 84" sequence evolution model using *evolver13*, implemented in PAML [50], to assess the distributions of the SRT and SRT_i statistics under the null model and to evaluate p-values when the distribution is not well approximated by the asymptotic theory.

We simulated two sets of 1 million CNCs under the null, one set for the mammalian and one for the amniotic CNCs, matching the distribution of base frequencies, the CNC size, the variation in the local substitution rates and the overall tree shape. Specifically, each of the 1 million simulated CNCs was based on the characteristics of a randomly sampled CNC in our data set (sampling with replacement). We specified the branch lengths by rescaling the

global tree estimated from all CNCs by the local substitution rate of the chosen CNC, and generated an alignment of five sequences of the corresponding size simulated on the specified tree.

For each set of simulated CNCs, we obtained the empirical null distributions of statistics testing for various alternative scenarios (the simplest and most extreme cases are shown in Supplementary Figure 3). As mentioned earlier, the asymptotic theory works reasonably well, except when testing for rate changes on short branches (i.e., the human and chimpanzee lineages). We also grouped sets of CNCs into bins with similar CNC sizes and local substitution rates and examined empirical distributions for each bin separately (data not shown). For each test statistic, the empirical distributions across bins were homogeneous, except for the cases in which the asymptotic theory does not work because of the small number of accumulated substitutions. For these cases, since the inhomogeneity across bins was mainly explained by differences in the expected number of substitutions on the tested lineage, we reconstructed bins of CNCs according to the number of expected substitutions and evaluated p-values within each bin separately.

We also simulated a number of CNC sets under various alternative scenarios to examine the performance of the modified AIC method. Specifically, for each k selected lineages ($k = 0, \ldots, 6$), we simulated 100,000 CNCs in which each CNC has k branches that evolve with their own rates. These rates were higher or lower than the background rate with 50% probability each. To incorporate the variation in strength of signals in real data, the rate of each selected branch was simulated by multiplying or dividing the background rate by a scale factor that is drawn from $1 + \Gamma(\alpha, \beta)$ distribution with a scale parameter $\beta = 1$ and a shape parameter $\alpha = 1$ (weak signals) or $\alpha = 2$ (stronger signals).

Gene Ontology analysis. We downloaded a reference assembly (seq_gene.md.gz) that corresponds to the human genome build (NCBI build 35) from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.35.1/mapview. We proceeded

by extracting only reference genes (release 3 [45]) in autosomes only, and obtained 25,249 genes. We extracted the nearest gene for each CNC without considering gene orientation. Here the distance from a CNC to a gene is the minimum of distances from the middle of the CNC to either end of the gene.

The PANTHER Gene Ontology database was downloaded from http://www.pantherdb.org/panther/prowler.jsp in March 2006. For each conservation group, we examined what kinds of biological process categories are enriched for being: (1) near CNCs in general; and (2) near CNCs showing lineage-specific rate increases compared to near CNCs evolving under the null model. The nearest genes of CNCs were used for this analysis.

For (1), we compiled the list of all genes in the reference assembly and the list of genes near mammalian (or amniotic) CNCs. For each biological process category, we counted the number of genes in each list and compared them with a chi-square test. Within each list, genes are counted only once. For (2), CNCs were first classified by the AIC (in order to obtain disjoint categories of CNCs showing signals of speedups on each lineage). For each category, we counted genes near CNCs under the null and near CNCs in each of the seven selection groups that show rate speed-ups on a single lineage. In this case, however, individual genes were counted repeatedly each time they were the nearest neighbor of a relevant CNC. The reason for this is that multiple CNCs often have the same nearest neighbor. This effect is more pronounced in the null CNC group than in the selection groups. Consequently, if we count genes only once, then any biological functional category that is enriched near CNCs, in general, may be under-represented in the null but over-represented in each selection group. Since the numbers of selected CNCs are small for many gene categories, p-values were computed using Fisher's exact test. To account for multiple testing, the p-values were multiplied by the number of biological processes that were jointly tested.

Data availability

We will prepare a datafile that contains the list of all CNCs and summarizes our analysis results. It will includes genomic properties, test statistics as well as the best evolutionary pattern of each CNC. It will be downloadable from http://pritch.bsd.uchicago.edu/data.html.

Acknowledgements

We thank Peter McCullaugh, Dan Nicolae, Molly Przeworski, Adam Siepel and members of the Pritchard lab as well as an anonymous reviewer for many helpful discussions or comments. SYK was supported in part by NSF Grant No. DMS0305009 to P. McCullaugh, and JKP by a grant from the Packard Foundation.

Kim and Pritchard-Supplementary Information

Accession numbers

The Entrez gene accession numbers for the gene C21orf54 and IFNAR2 are GeneID:339629 and GeneID:3455, respectively.

References

[1] King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–116.

[2] Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, et al. (2005) Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol 3:e387.

[3] Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440:242–245.

[4] Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963.

[5] Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443:167–172.

[6] Pollard KS, Salama SR, King B, Kern AD, Dreszer T, et al. (2006) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2:e168.

[7] Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314:786.

[8] Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

[9] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304:1321–1325.

[10] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050.

[11] Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.

[12] International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.

[13] Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7.

[14] Margulies EH, Blanchette M, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13:2507–2518.

[15] Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res 13:813–820.

[16] Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913.

[17] Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, et al. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res 16:855–863.

[18] Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, et al. (2004) Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog

genomes shows that their selective constraint is independent of their genic environment. Genome Res 14:852–859.

[19] Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302:413.

[20] Uchikawa M, Takemoto T, Kamachi Y, Kondoh H (2004) Efficient identification of regulatory sequences in the chicken genome by a powerful combination of embryo electroporation and genome comparison. Mech Dev 121:1145–1158.

[21] de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, et al. (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res 15:1061–1072.

[22] Ahituv N, Rubin EM, Nobrega MA (2004) Exploiting human–fish genome comparisons for deciphering gene regulation. Hum Mol Genet 13 Spec No 2:261–266.

[23] Ahituv N, Prabhakar S, Poulin F, Rubin EM, Couronne O (2005) Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum Mol Genet 14:3057–3063.

[24] McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16:451–465.

[25] Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502.

[26] Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet 6:151–157.

[27] Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223–227.

[28] Kryukov GV, Schmidt S, Sunyaev S (2005) Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet 14:2221–2229.

[29] Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005) Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 15:1373–1378.

[30] Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31:51–54.

[31] Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104.

[32] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141.

[33] Smith NGC, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12:1350–1356.

[34] Ellegren H, Smith NGC, Webster MT (2003) Mutation rate variation in the mammalian genome. Curr Opin Genet Dev 13:562–568.

[35] Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res 13:13–26.

[36] Webster MT, Smith NGC, Lercher MJ, Ellegren H (2004) Gene expression, synteny, and local similarity in human noncoding mutation rates. Mol Biol Evol 21:1820–1830.

[37] Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, et al. (2005) Why do human diversity levels vary at a megabase scale? Genome Res 15:1222–1231.

[38] Gaffney DJ, Keightley PD (2005) The scale of mutational variation in the murid genome. Genome Res 15:1086–1094.

[39] Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

[40] Siepel A HD Pollard KS (2006) New methods for detecting lineage-specific selection. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology 3909:190–205.

[41] Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33:54–58.

[42] Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23:273–277.

[43] Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.

[44] Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664.

[45] International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

[46] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376.

[47] Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174.

[48] Yap VB, Pachter L (2004) Identification of evolutionary hotspots in the rodent genomes. Genome Res 14:574–579.

[49] Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author Department of Genome Sciences, University of Washington, Seattle .

[50] Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556.

[51] Gopalan R, Berry DA (1998) Bayesian multiple comparisons using Dirichlet process priors. Journal of the American Statistical Association 93:1130–1139.

[52] Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112.

[53] Durbin J, Watson GS (1951) Testing for serial correlation in least squares regression. II. Biometrika 38:159–178.

[54] Racine J HR (2002) Using R to teach econometrics. Journal of Applied Econometrics 17:149–174.

[55] Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440–9445.

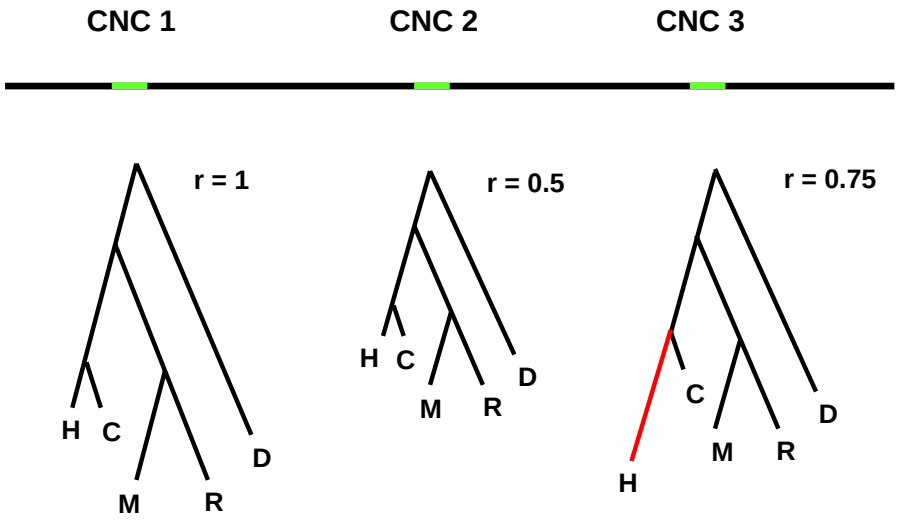


Figure 1: Schematic illustration of our method applied to three CNCs. The tree beneath each CNC shows hypothetical branch lengths for the phylogeny connecting human, chimpanzee, mouse, rat and dog (H, C, M, R, and D, respectively). Each CNC is associated with a single rate parameter r that accounts for variation in the local mutation rate and level of conservation; however under the null model the relative branch lengths are all the same. CNC 3 has an unusually long human branch suggesting positive adaptation on the human lineage.

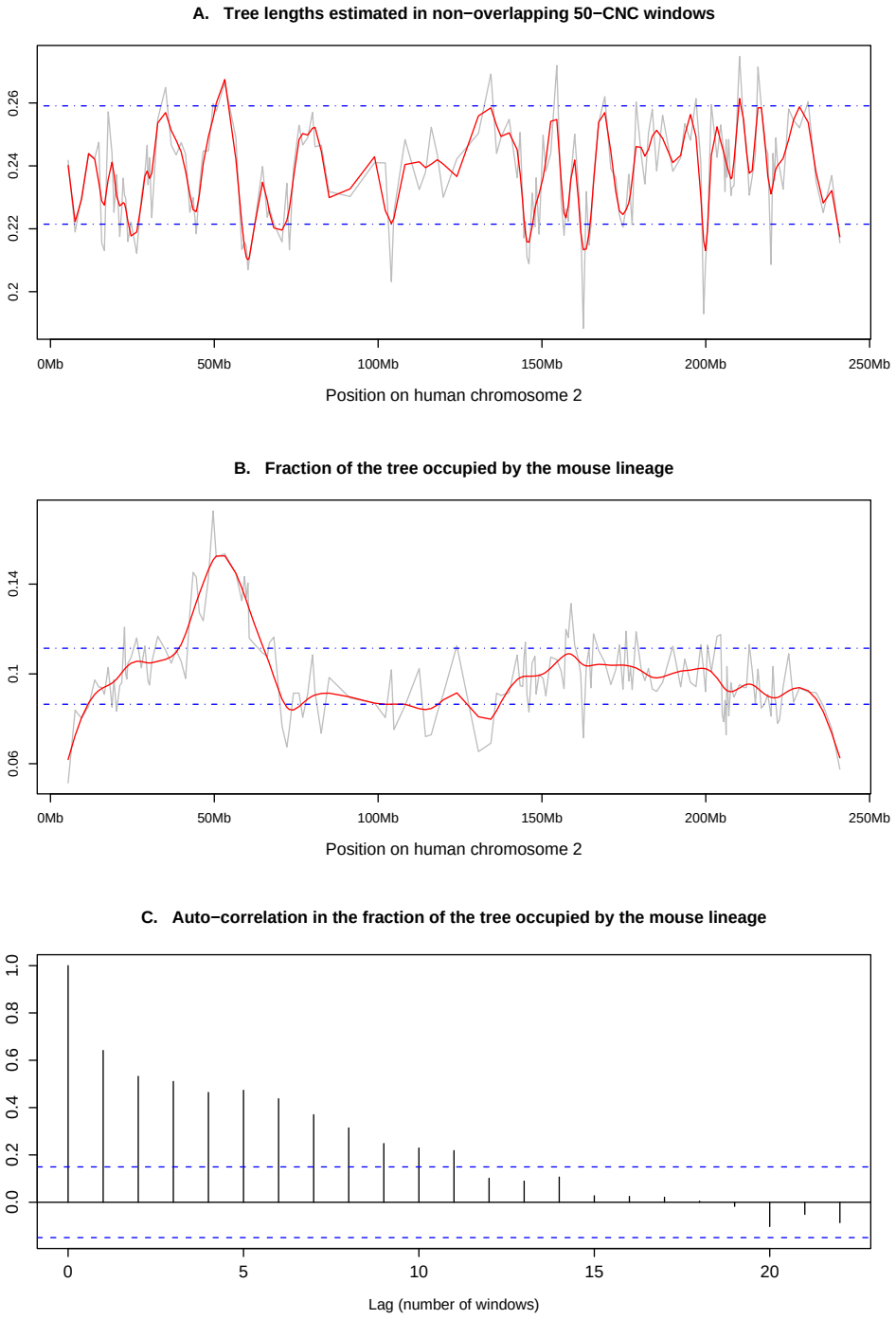


Figure 2: Local variation in the tree shape in mammalian CNCs located on human chromosome 2. Values for the upper two plots are calculated in non-overlapping windows of 50 consecutive CNCs. **A.** Local variation in total tree lengths (average number of substitutions per site). Both the raw data (gray) and the smoothed data (red) are shown. The dashed blue horizontal lines indicate the 2.5% and 97.5% quantiles of the (unsmoothed) distribution expected if there were no spatial heterogeneity (estimated by randomly shuffling the location of chromosome 2 CNCs). **B.** Fraction of the tree occupied by the mouse lineage. **C.** Long range dependence in the fraction of the tree occupied by the mouse lineage. The plot shows the correlation between windows separated by a gap of i other windows; values outside the dotted lines are significant at the 5% level. The Durbin-Watson test for autocorrelation is significant at $p < 10^{-15}$ [53, 54].

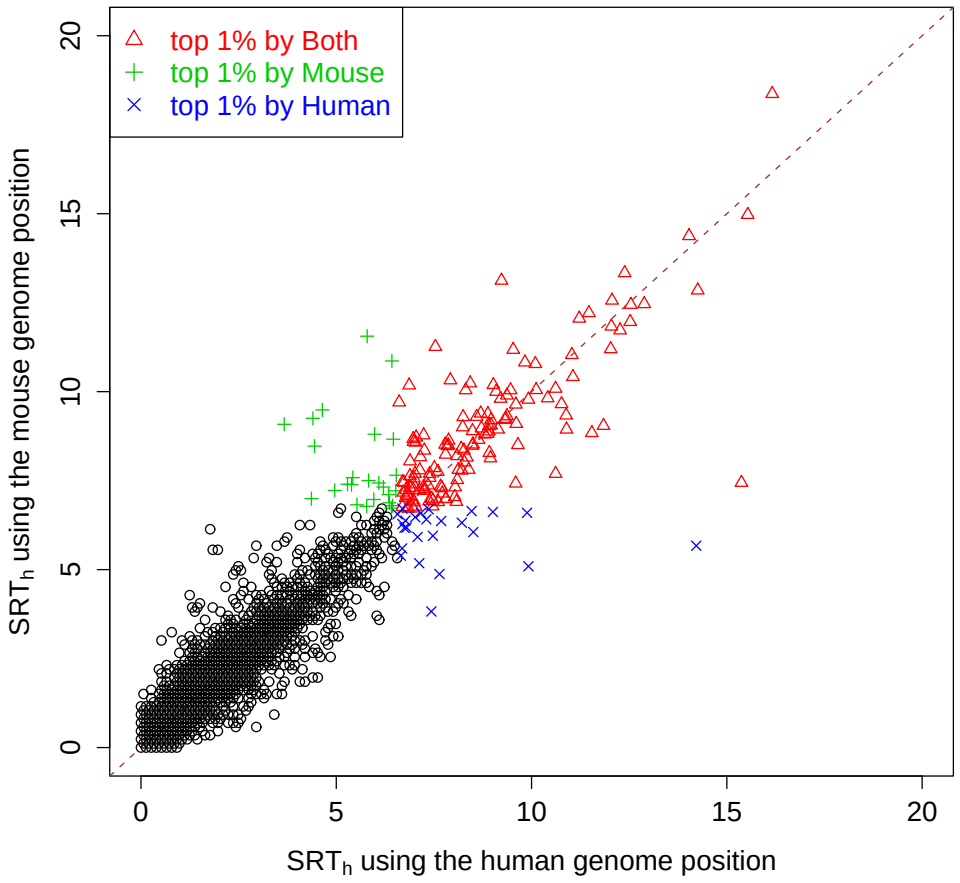
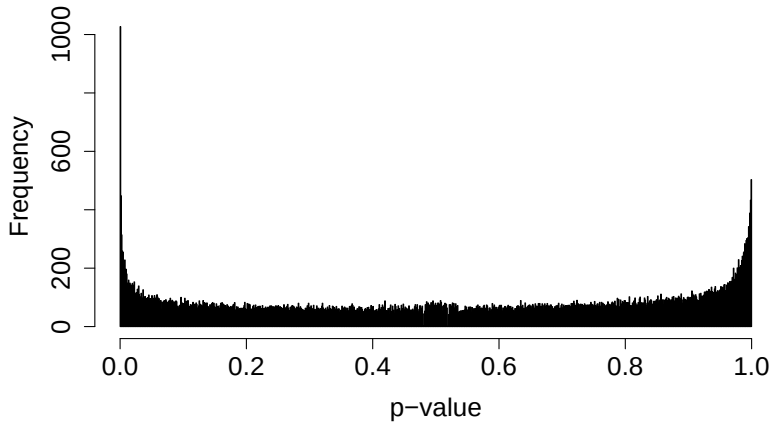


Figure 3: Robustness of the SRT test (SRT_h) to the definition of windows used for estimating the null tree shape. The SRT_h values for amniotic CNCs were computed separately based on windows defined using the human genome position (x-axis) and the mouse genome position (y-axis). Signals that are in the top 1% by both window definitions are shown in red. Eight outliers above 20 were removed from the plot. The overall concordance between the two data sets implies that changes in synteny between human and mouse do not greatly disrupt our sliding window estimation procedure.

A



B

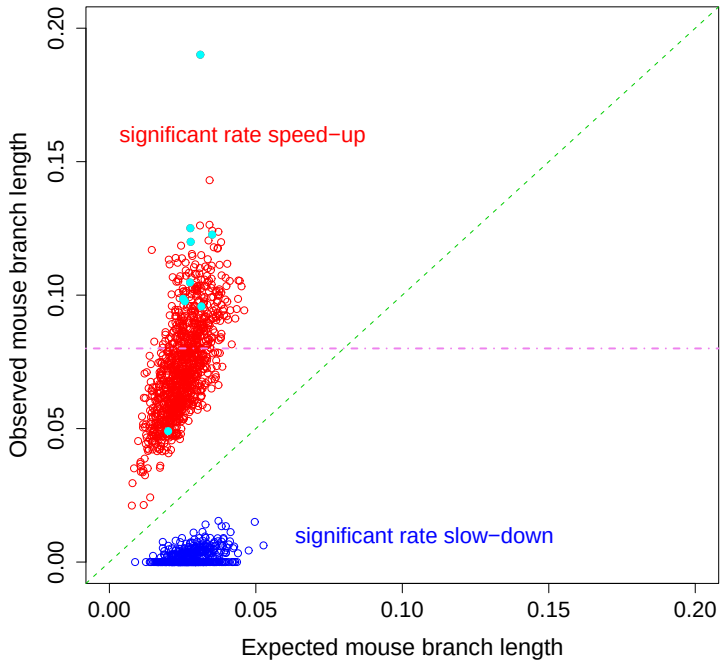


Figure 4: Excess of significant mammalian CNCs on the mouse lineage. **A.** Histogram of p-values of SRT_m. The peaks on the left and the right indicate an excess of CNCs that are fast- and slow-evolving in mice, respectively. **B.** Observed and expected branch lengths (per site) of mammalian CNCs that are significant on the mouse lineage at $p < 0.001$. Fast- and slow-evolving CNCs are indicated in red and blue, respectively. The violet dashed horizontal line shows the genome-wide average substitution rate on the mouse lineage for unconstrained regions near the fast-evolving CNCs (see text). Nine CNCs that have evolved significantly faster than their local neutral rates on the mouse lineage ($p < 0.05$) are indicated by light blue dots.

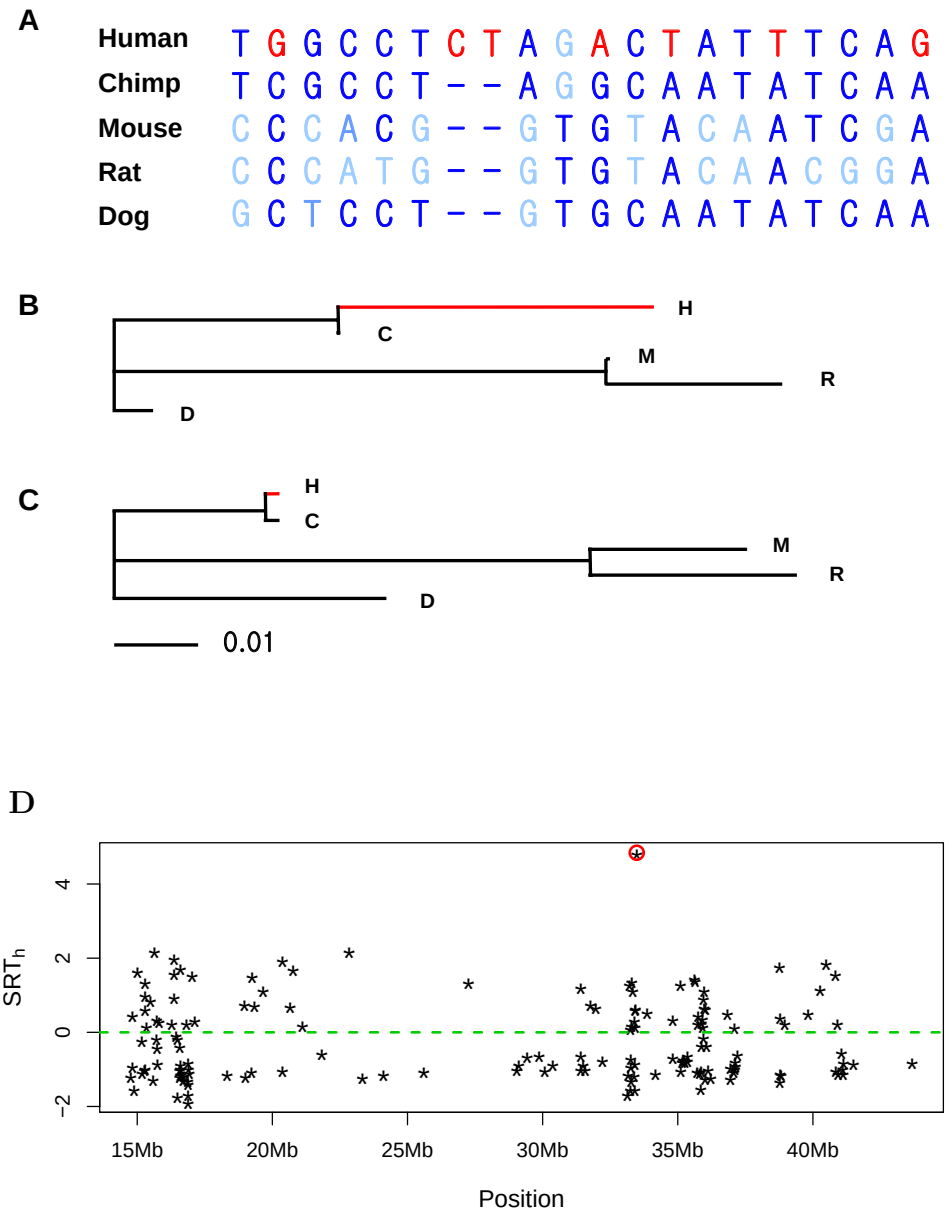


Figure 5: Example of a 144 bp CNC on chromosome 21 (q22.11) with a dramatic accumulation of changes on the human lineage (see text). **A.** Data at the 20 sites that are variable among these five species, with human-specific changes in red. **B.** Estimated tree for this CNC. **C.** Estimated neutral tree based on neighboring CNCs. The scale bar indicates the expected number of substitutions per site, per unit branch length. **D.** SRT_h values for amniotic CNCs located on human chromosome 21. The red circle indicates the CNC illustrated above.

A

Number of parameters	*#CNCs*
1	54,643
2	22,839
3	3,989
4	451
5	35
6	0
7	0
Total	81957

B

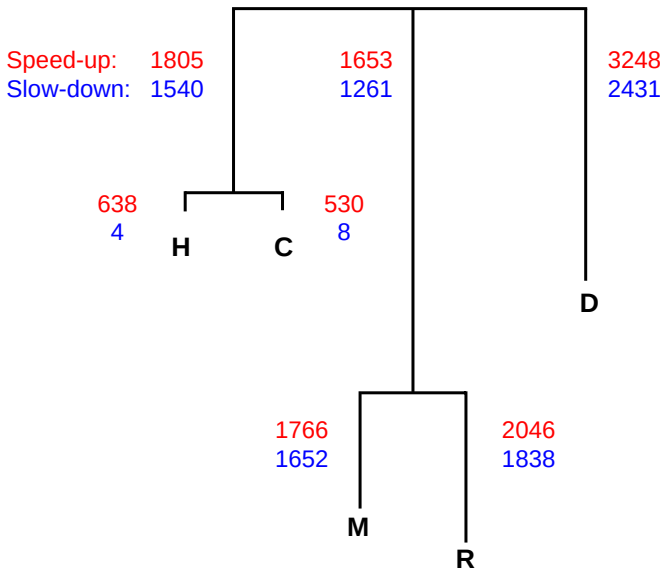


Figure 6: Patterns of evolution in mammalian CNCs. **A.** Classification of evolutionary patterns in mammalian CNCs according to our modified AIC. The left-hand column indicates the number of model parameters, where '1' indicates that there is a single substitution rate on the entire tree, and where '7' indicates a separate rate on every branch. **B.** The tree shows how many of the CNCs that are best fit by the two-parameter model have altered rates on each branch. Rate increases are printed in red (upper text) and rate decreases in blue (lower text). The classification of the remaining 2419 CNCs with 2 rate parameters that fall into our compound models is summarized in Supplementary Table 12.

Table 1: Evidence for adaptive evolution in fast-evolving CNCs. The table shows, for mammalian and amniotic CNCs, (1) the numbers of elements that show significant accelerations on each branch ($p < 0.001$ by SRT_i) (Total); (2) the numbers for which the maximum likelihood rate estimate exceeds the local neutral rate for that branch (Exceed); and (2) the numbers for which the rate on that branch significantly exceeds the local neutral rate ($p < 0.05$) (SigExceed). See Methods for further details. Each branch is labeled using the species that it leads to. The "primate" and "rodent" lineages indicate the lineages leading to the common ancestors of human and chimpanzee, and mouse and rat, respectively. Note that on the long branches, even the fast-evolving CNCs are generally slower than the neutral rate, which is why the fraction significantly faster than neutral is $< 5\%$.

		Human	Chimp	Mouse	Rat	Primate	Rodent	Dog
Mammalian	Total	211	181	1027	1271	855	723	1903
Mammalian	Exceed	207	180	339	455	137	0	13
Mammalian	SigExceed	144	128	9	9	0	0	1
Amniotic	Total	44	37	228	305	200	198	511
Amniotic	Exceed	42	35	62	90	14	0	2
Amniotic	SigExceed	18	15	6	8	1	0	0