Robust inference of positive selection on regulatory sequences in human brain

Jialin Liu[1,2,3]*, Marc Robinson-Rechavi[1,2]*

1. Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
2. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
3. Current address: Biozentrum, University of Basel, Basel, Switzerland

* corresponding authors: jialin.liu@unil.ch, marc.robinson-rechavi@unil.ch

Abstract

A long standing hypothesis is that divergence between humans and chimpanzees might have been driven more by regulatory level adaptions than by protein sequence adaptations. This has especially been suggested for regulatory adaptions in the evolution of the human brain. There is some support for this hypothesis, but it has been limited by the lack of a reliable and powerful way to detect positive selection on regulatory sequences. We present a new method to detect positive selection on transcription factor binding sites, based on measuring predicted affinity change with a machine learning model of binding. Unlike other methods, this requires neither defining a priori neutral sites, nor detecting accelerated evolution, thus removing major sources of bias. The method is validated in flies, mice, and primates, by a McDonald-Kreitman-like measure of polymorphism vs. divergence, by experimental binding site gains and losses, and by changes in expression levels. We scanned the signals of positive selection for CTCF binding sites in 29 human and 11 mouse tissues or cell types. We found that human brain related cell types have the highest proportion of positive selection. This is consistent with the importance of adaptive evolution on gene regulation in the evolution of the human brain.

Introduction

It has long been suggested that changes in gene regulation have played an important role in human evolution, and especially in the evolution of the human brain and behavior (King and Wilson 1975; Anon 2005). Many human and chimpanzee divergent traits (Varki and Altheide 2005) cannot be explained by protein sequence adaptations. For example, there is little evidence to link protein sequence adaptations to traits related to cognitive abilities (Franchini and Pollard 2015). Conversely there is some evidence of brain-specific gene expression divergence in humans (Enard et al. 2002), which is consistent with a role of regulatory evolution. Yet a central question remains open: which regulatory changes were adaptive, if any? A major limitation in answering this is the lack of a robust model of neutral vs. adaptive evolution for regulatory elements.

One approach to detect adaptive evolution on regulatory elements is to detect noncoding regions with lineage-specific accelerated evolutionary rates (Pollard et al. 2006; Prabhakar et al. 2006; Gittelman et al. 2015). For example, Gittelman et al. (2015) found human accelerated regions close to genes annotated to terms such as brain or neuron development. A major caveat is that such acceleration may result from neutral mechanisms such as biased gene conversion (Galtier and Duret 2007) rather than from selection. A second approach is to use a MK test framework (McDonald and Kreitman 1991; Ludwig and Kreitman 1995; Andolfatto 2005; Arbiza et al. 2013; Gronau et al. 2013). This approach has two limitations. First, an expected neutral divergence to polymorphism ratio needs to be defined, whereas defining neutral sites for regulatory elements is difficult and can bias results (Zhen and Andolfatto 2012). And second it lacks power on individual elements, since many regulatory elements are short and present very few variable sites (Andolfatto 2005).

We have developed a new method to detect adaptive evolution of transcription factor binding sites (TFBSs), based on predicted binding affinity changes. As a proof-of-principle, we first applied this method to well-known transcription factors, such as CEBPA and CTCF, in species triples focused on human, mouse, or fly. We validated it with three independent lines of evidence: our evidence of positive selection is associated to higher empirical binding affinity, higher substitution to polymorphism ratio in sequence, and lower variance in expression of neighbouring genes. Then, we used this method to detect positive selection of CTCF binding sites in 29 human tissues or cell types. We found the highest positive selection in brain samples,

followed by male reproductive system. The same analysis in mouse found the highest positive selection in the lung, with no special signal in the brain. Thus, we provide clear evidence for adaptive evolution of gene regulation in the human brain.

Results

Detecting positive selection on transcription factor biding sites

We propose a computational model to detect positive selection on transcription factor binding sites (TFBSs), or any other elements for which we have experimental evidence similar to ChIP-seq (Figure 1, and Methods). Briefly, a gapped k-mer support vector machine (gkm-SVM) classifier is trained on ChIP-seq peaks (here, TFBSs). This allows to compute SVM weights of all possible 10-mers, which are predictions of their contribution to transcription factor binding affinity (Lee et al. 2015). We can then predict the binding affinity impact of substitutions by calculating deltaSVM, the difference of sum weights between two homologous sequences. We compare each empirical TFBS to an ancestral sequence inferred from alignment with a sister species and an outgroup.

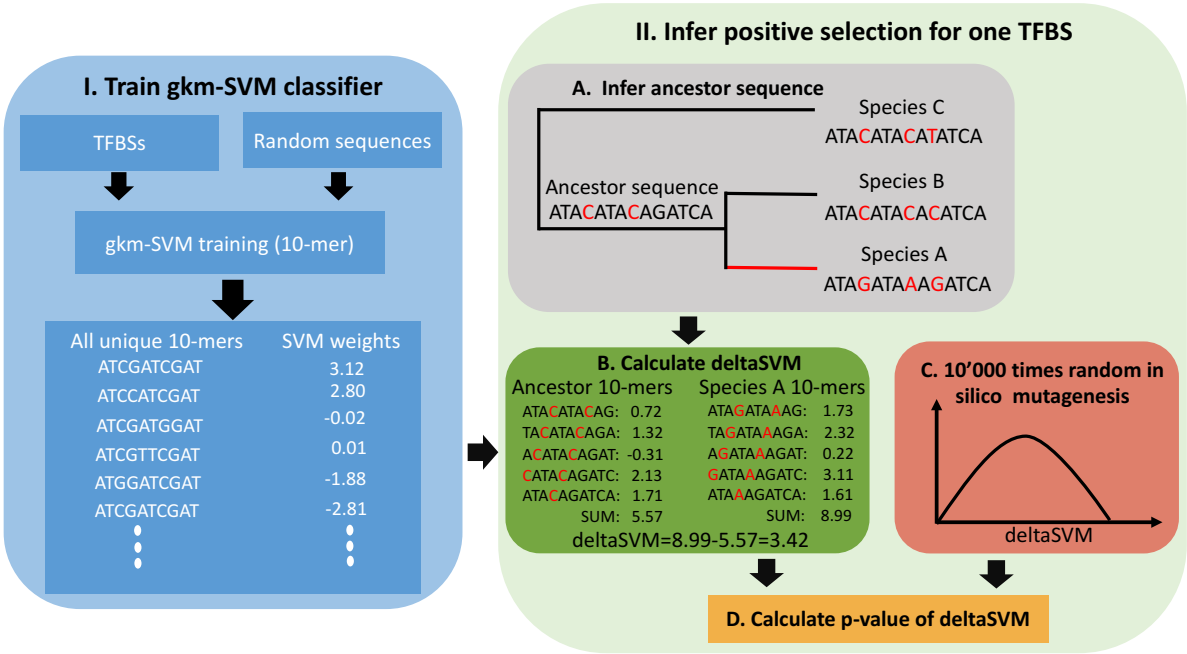

Figure 1: Illustration of the procedure for inferring positive selection

The method includes two parts. Part I (left) is the gapped k-mer support vector machine (gkm-SVM) model training. The gkm-SVM classifier was trained by using TFBSs as a positive training set and randomly sampled sequences from the genome as a negative training set. Then, SVM weights of all possible 10-mers, the contributions of prediction transcription factor binding affinity, were generated from the gkm-SVM. Part II (right) part is the positive selection

inference. The ancestor sequence was inferred from sequence alignment with a sister species (species B) and an outgroup (species C). Then, the binding affinity change (deltaSVM) of the two substitutions accumulated in the red branch leading to species A were calculated based on the weight list. The significance of the observed deltaSVM was evaluated by comparing it with a null distribution of deltaSVM, constructed by scoring the same number of random substitutions 10000 times.

Adaptive evolution on TFBSs is expected to push them from a sub-optimal towards an optimal binding strength, or from an old optimum to a new one (e.g. in response to changing environment). Thus TFBSs evolving adaptively are expected to accumulate substitutions which consistently change the phenotype to stronger or to weaker binding, whereas TFBSs evolving under purifying selection are expected to accumulate substitutions which increase or diminish binding in approximately equal measure, around a constant optimum. This reasoning follows the principle of a sign test of phenotypes (Coyne 1996; Orr 1998), although it uses the actual values and not just the sign. In practice, this should lead to a large absolute value of deltaSVM under adaptive selection. We estimate by randomization a p-value specific to each individual TFBS and to its number of substitutions (see Methods). Thus, our method can infer the action of natural selection pushing a TFBS to a new fitness peak of either higher (positive deltaSVM) or lower (negative deltaSVM) binding affinity than its ancestral state.

Detecting positive selection on liver TFBSs in *Mus musculus*

We first applied our method to a large set of TFBSs in the liver of three mouse species (*Mus musculus domesticus* C57BL/6J, *Mus musculus castaneus* CAST/EiJ, and *Mus spretus* SPRET/EiJ), identified by ChIP-seq for three liver-specific transcription factors, CEBPA, FOXA1, and HNF4A (Stefflova et al. 2013). We inferred positive selection on the lineage leading to C57BL/6J after divergence from CAST/EiJ (Figure 2A). For the sake of simplicity, we only present the results of CEBPA in the main text; results are consistent for FOXA1 and HNF4A (Supplementary materials). We first trained a gkm-SVM on 41945 CEBPA binding sites in C57BL/6J (see Methods). The gkm-SVM very accurately separates CEBPA binding sites and random sequences (Figure 2B). Based on the experimental ChIP-seq peaks in the three species, using SPRET/EiJ as an outgroup, we identified three categories of CEBPA binding sites: conserved in all three species ("conserved", 24280 sites), lineage specific gain in C57BL/6J ("gain", 6304 sites), and lineage specific loss in C57BL/6J ("loss", 6692 sites).

Based on whole genome pairwise alignments of C57BL/6J to CAST/EiJ and to SPRET/EiJ, we derived the substitutions accumulated on the C57BL/6J lineage for each CEBPA binding site (see Methods). We only kept binding sites with at least two substitutions, leading to 5114, 1445, and 1497 TFBSs for conserved, gain, and loss categories respectively. For each binding site, we calculated a deltaSVM value, and inferred its significance by random in silico mutagenesis (see Methods).

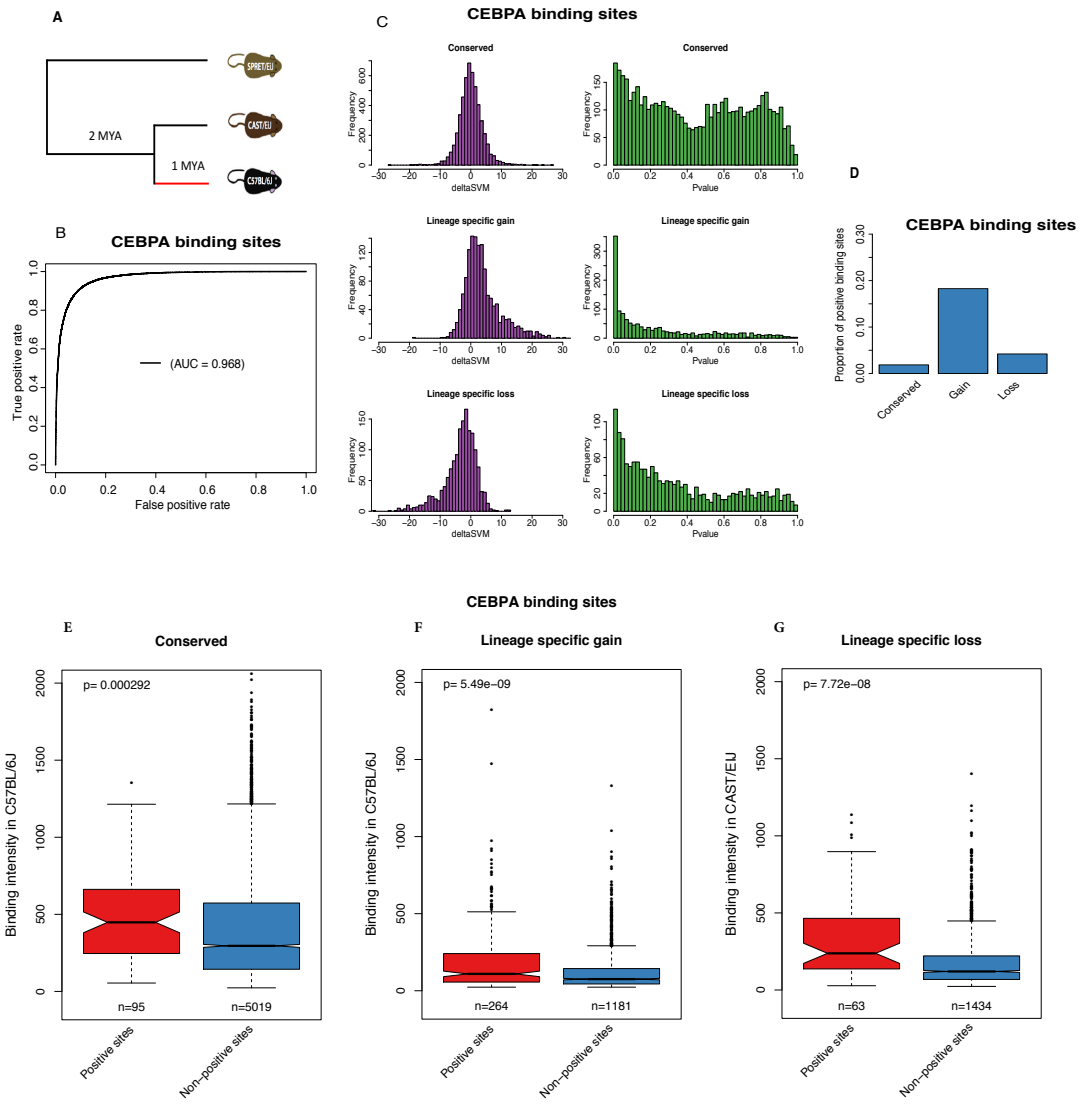

Figure 2: Mouse CEBPA biding sites study

A. Topological illustration of the phylogenetic relationships between the three mouse species used to detect positive selection on CEBPA binding sites. We want to detect positive selection which occurred on the lineage of C57BL/6J after divergence from CAST/EiJ, as indicated by the red branch. SPRET/EiJ is the outgroup used to infer binding site sequence in the ancestor of C57BL/6J and CAST/EiJ.

B. Receiver operating characteristic (ROC) curve for gkm-SVM classification performance on CEBPA binding sites (5-fold cross validation). The AUC value represents the area under the ROC curve and provides an overall measure of predictive power.

C. The left hand graphs are the distributions of deltaSVM for conserved, gain, and loss binding sites. The right hand graphs are the distributions of deltaSVM p-values (test for positive selection) for conserved, gain, and loss binding sites.

D. The proportion of CEBPA binding sites with evidence of positive selection in conserved, gain, and loss binding sites.

E-G. Comparison of biding intensity between positive sites and non-positive sites for mouse CEBPA. The number of binding sites in each category is indicated below each box. The p-values from a Wilcoxon test comparing categories are reported above boxes. The lower and upper intervals indicated by the dashed lines ("whiskers") represent 1.5 times the interquartile range, or the maximum (respectively minimum) if no points are beyond 1.5 IQR (default behavior of the R function boxplot). Positive sites are binding sites with evidence of positive selection (deltaSVM p-value < 0.01), non-positive sites are binding sites without evidence of positive selection.

E. Conserved binding sites.

F. Lineage specific gain binding sites.

G. Lineage specific loss binding sites. We compare the binding intensity from CAST/EiJ, as an approximation for ancestral binding intensity, between positive loss binding sites and non-positive loss binding sites.

We plot the distributions of deltaSVMs and of their corresponding p-values for each binding site evolutionary category (Figure 2C). As expected, the distribution of deltaSVMs is symmetric for conserved, has a skew towards positive deltaSVMs for gain, and a skew towards negative deltaSVMs for loss. These results confirm that the gkm-SVM based approach can accurately predict the effect of substitutions on transcription factor binding affinity change. For the distribution of p-values, in all binding site categories, there is a skew of p-values near zero, indicating some signal of positive selection. Gain has the most skewed distribution of p-values towards zero. Hereafter we will use 0.01 as a significant threshold to define positive selection, but results are robust to different thresholds (see **Validation based on ChIP-seq binding intensity**). This identifies almost 20% of gain having evolved under positive selection (Figure 2D), relative to 4% of loss, and 2% of conserved. Random substitutions tend to decrease the

binding affinity rather than increase it (Figure S1), because it's easier to break a function than to improve it. Thus our method could be biased towards reporting as positive sites with more left shifted null distributions. However, this is not the case (Figure S2).

In summary, we found widespread positive selection driving the gain of CEBPA binding sites. We also found some evidence of positive selection driving loss, or increase of binding affinity in some conserved sites. For the other two transcription factors (FOXA1 and HNF4A), we found very consistent patterns (Figure S3, S4).

Validation based on ChIP-seq binding intensity

We expect that conserved or gained sites which evolved under positive selection with positive deltaSVM should have increased binding affinity. Thus the positive binding sites (PBSs) should have higher biding affinity than non-positive selection binding sites (non-PBSs) in the focal species C57BL/6J. This is indeed the case (Figure 2E and 2F). In addition, conserved TFBSs have higher activity than recently evolved ones ('gain'). For loss, however, the PBSs have a strong decrease in binding affinity, so we expect higher binding affinity of PBSs in the ancestor. Using CAST/EiJ as an approximation for ancestor binding affinity, this is indeed the case (Figure 2G). Results are also consistent with different p-value thresholds (Figure S5). We performed the same validations in FOXA1 and HNF4A, with consistent results (Figure S6).

Validating the inference of positive selection with human liver TFBSs

To further validate our method, we took advantage of the abundant population genomics transcriptomics data in humans. We inferred positive selection of CEBPA binding sites in the human lineage after divergence from chimpanzee, with gorilla as outgroup (Figure 3A). As in mouse, the gkm-SVM trained from 15806 CEBPA binding sites in human can very accurately separate TFBSs and random sequences (Figure 3B). The distribution of deltaSVMs is slightly asymmetric, with a higher proportion of positive values (Figure 3C). This is because these binding sites contain both conserved and gain, but no loss (since we detect only in the focal species). Based on the distribution of p-values, 7.5% of CEBPA binding sites are predicted to have evolved adaptively in the human lineage.

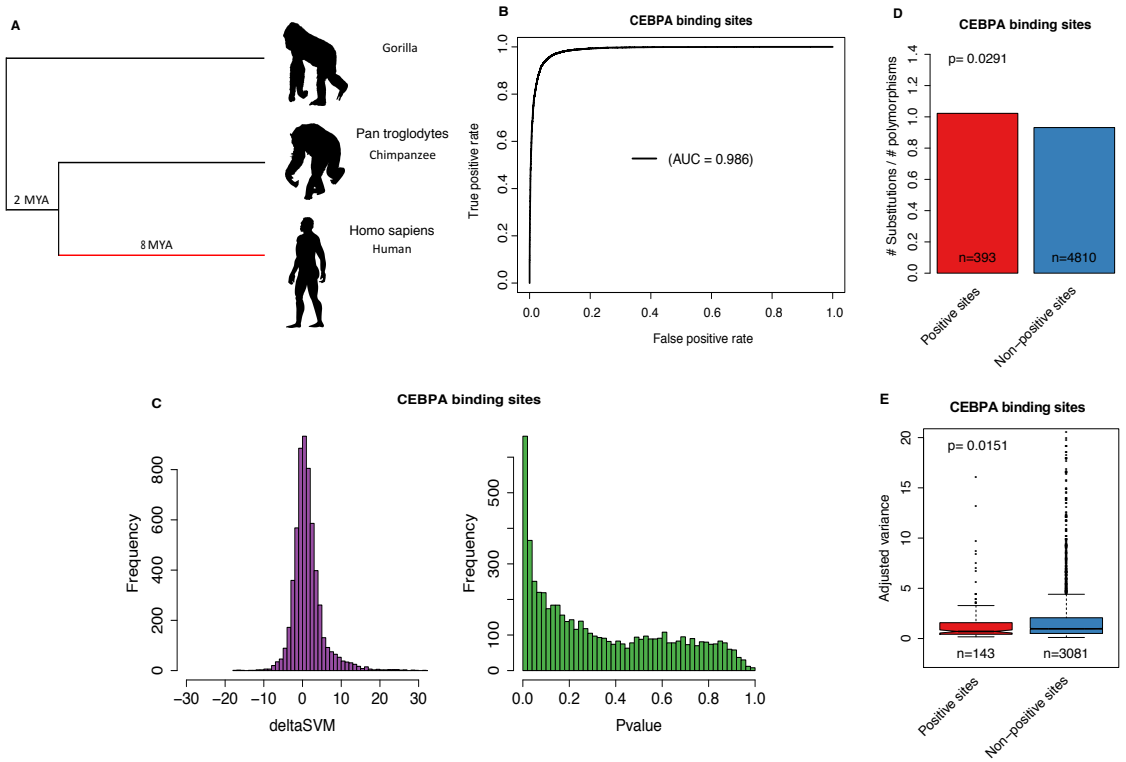

Figure 3: Human CEBPA biding sites study

A. Topological illustration of the phylogenetic relationships between human, chimpanzee and gorilla. We detected positive selection which occurred on the lineage of human after divergence from chimpanzee, as indicated by the red branch. Gorilla is the outgroup used to infer binding site sequence in the ancestor of human and chimpanzee.

B. Receiver operating characteristic (ROC) curve for gkm-SVM classification performance on CEBPA binding sites (5-fold cross validation). The AUC value represents the area under the ROC curve and provides an overall measure of predictive power.

C. The left graph is the distribution of deltaSVM. The right graph is the distribution of deltaSVM p-values (test for positive selection).

D. The ratio between the number of substitutions and the number of polymorphisms (SNPs) for CEBPA binding sites. Positive sites are binding sites with evidence of positive selection (deltaSVM p-value < 0.01), non-positive sites are binding sites without evidence of positive selection. The p-value from Fisher's exact test is reported above the bars.

E. Comparison of expression variance (adjusted variance) of putative target genes (closest gene to a TFBS) between positive sites and non-positive sites. The number of binding sites in each category is indicated below each box. The p-values from a Wilcoxon test comparing categories are reported above boxes. The lower and upper intervals indicated by the

dashed lines ("whiskers") represent 1.5 times the interquartile range, or the maximum (respectively minimum) if no points are beyond 1.5 IQR (default behavior of the R function boxplot). Positive sites are binding sites with evidence of positive selection (deltaSVM p-value < 0.01), non-positive sites are binding sites without evidence of positive selection.

Using the MK framework (McDonald and Kreitman 1991), we predict that PBSs should have higher substitutions to polymorphisms ratios than non-PBSs. Note that we do not need to define neutral sites a priori. As expected, we found that the PBSs have a significantly higher ratio of fixed nucleotide changes between human and chimpanzee to polymorphic sites in human than non PBSs (Figure 3D). This is an external validation that our method detects positive selection, as the input did not contain any information about polymorphism.

Besides a higher substitutions to polymorphisms ratio, we also expect that the expression of PBSs putative target genes (see Methods) should be more conserved among human populations. If the expression of PBSs target genes is an adaptive trait in humans, further changes in expression will reduce fitness. Moreover, recent adaptive sweeps are expected to have reduced variability for the regulation of these genes. As expected, we found that PBSs target genes have significantly lower expression variance (adjusted variance, controlling for the dependency between mean and variance, see Methods) across human populations than non-PBSs target genes (Figure 3E).

Thus, results from different sources of information support the expectations of our PBSs predictions. We performed the same analyses in HNF4A, and results are consistent (Figure S7). These results strongly suggest that our method is detecting real adaptive evolution signals.

Detecting positive selection of TFBSs in *Drosophila melanogaster*

By using a MK test framework (McDonald and Kreitman 1991), Ni et al. (2012) detected signatures of adaptive evolution on CTCF binding sites in *D. melanogaster*. They reported that positive selection has shaped CTCF binding evolution, and that newly gained binding sites show a stronger signal of positive selection than conserved sites. We applied our method to the same data as used in Ni et al. (2012). We detected positive selection in the *D. melanogaster* lineage after divergence from *D. simulans* (Figure S8A and S8B). Consistent with the findings

of Ni et al. (2012), we observed widespread positive selection for both conserved and gain (Figure S8C). In addition, the gain has a higher proportion of positive selection than conserved (Figure S8D). As Ni et al. (2012) did not report specific sites, we cannot compare results more precisely. For lineage specific loss binding sites, however, we did not detect any signal of positive selection (Figure S8C). Interestingly, the proportion of positive selection in *D. melanogaster* is much higher than in *Mus musculus*. For example, we find almost 40% of gain under positive selection in *D. melanogaster*, twice the proportion in *Mus musculus*. It should be noted that different transcription factors and tissues were used, which complicates direct comparison.

Adaptive evolution of CTCF binding sites across tissues in human

To test whether there is stronger adaptive evolution of gene regulation in some human tissues, we applied our method to 80074 CTCF binding sites across 29 adult tissues or primary cell types (hereafter "cell types"; see Table S2). We chose CTCF because it was the factor with the largest number of tissues or primary cell types studied in a consistent manner, by the ENCODE consortium (The ENCODE Project Consortium 2012; Davis et al. 2018). CTCF is well known as a transcriptional repressor, but it also involved in transcriptional insulation and chromatin architecture remodeling (Phillips and Corces 2009). The gkm-SVM model trained from one cell type can accurately predict the biding sites in another cell type, and the model trained with all CTCF binding sites has better performance than the model trained with cell type specific binding sites (Figure S9). Thus we used a general gkm-SVM rather than different models for different cell types.

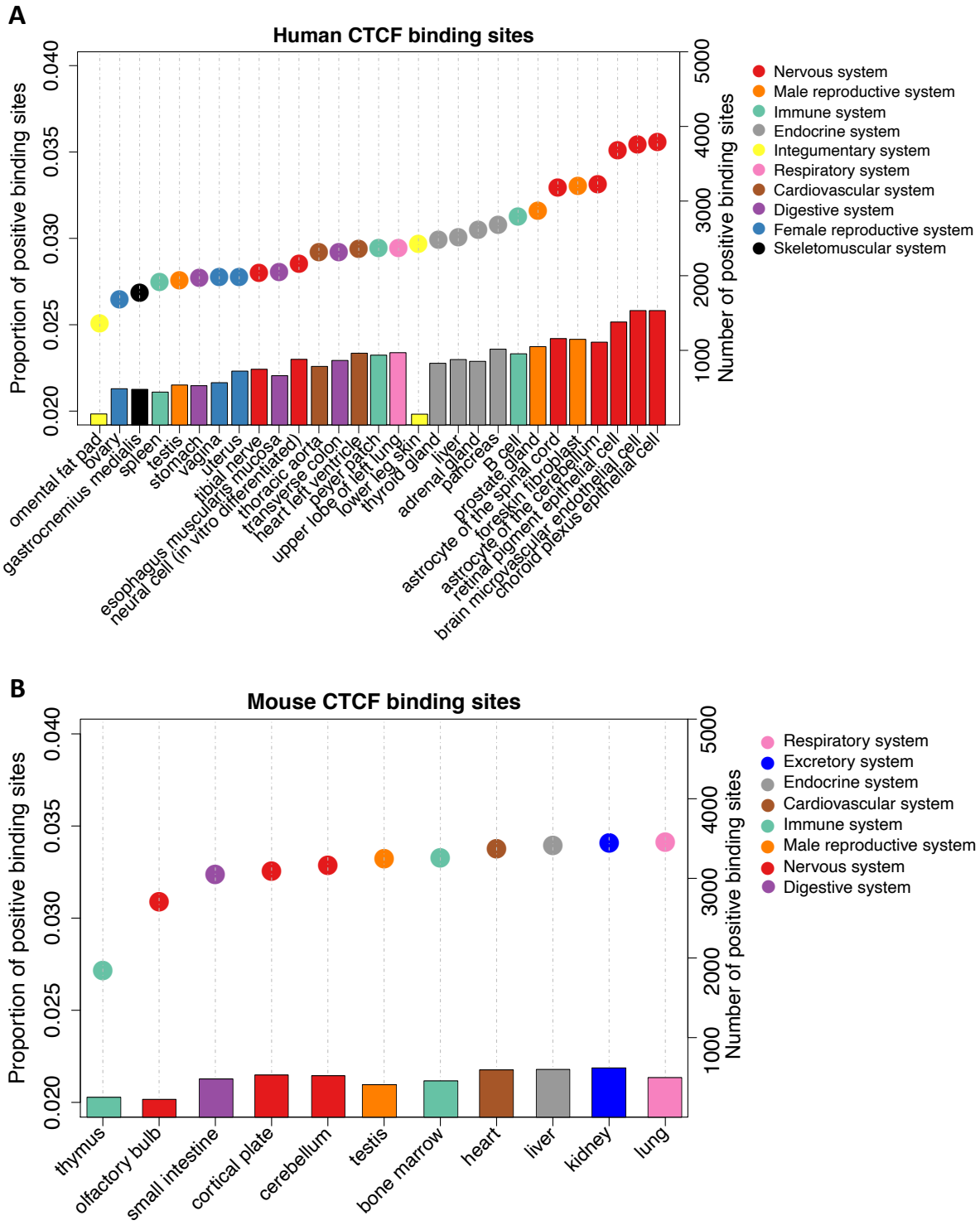

***Figure 4*: Proportion of positive CTCF binding sites in different tissues or cell types.**

Positive binding sites are binding sites with evidence of positive selection (deltaSVM p-value < 0.01). Colors correspond to broad anatomical systems.

A. CTCF binding sites in 29 human tissues or cell types.

B. CTCF binding sites in 11 mouse tissues.

We detected 3.52% of positive binding sites (PBSs) for adaptation on the human lineage (Figure S10A). We found that PBSs have higher substitutions to polymorphisms ratio than non-PBSs (Figure S11). In addition, PBSs are associated with a lower number of active cell types (Figure S12A) than non-PBSs, consistent with the prediction that pleiotropy can limit adaptive evolution (Wagner and Zhang 2011). We ranked cell types according to the proportion of binding sites that exhibit statistically significant evidence of positive selection. Brain related cell types have a higher proportion of positive selection than other cell types (Figure 4A). This pattern is consistent if we only use tissue specific CTCF binding sites (Figure S13A). Choroid plexus epithelial cell, brain microvascular endothelial cell, and retinal pigment epithelial cell have notably high PBSs frequencies. Non brain related nervous system cell types do not share this high positive selection. Nor does in vitro differentiated neural cell, which may reflect that they do not preserve the signal of specific in vivo differentiated cells. Notably, these brain related cell types also have a higher fraction of substitutions fixed by positive selection (see Methods) than other cell types, except lower leg skin (Figure S14).

To check whether our test could be too liberal or conservative for some sites, we first analyzed the substitution rate of all possible substitutions and their corresponding affinity change (delatSVM) in human CTCF binding sites. We split all substitutions into two categories: substitutions on CpG, and substitutions not on CpG. Within each category, we found as expected that the transition rate is much higher than the transversion rate, but we didn't find a trend for specific substitution types to strengthen or weaken binding affinity (Figure S15A and Figure S16). Between categories, we found that there is generally higher substitution rate on CpG, again as expected. Substitutions on CpG tend to weaken binding affinity (Figure S15A and Figure S16), indicating that our test could be conservative for sites with CpG substitutions. Second, we checked whether neighboring substitutions (dinucleotide substitutions) have a general tendency to change affinity in the same direction. Indeed, this is the case (Figure S15B), suggesting that our test could be too liberal or too conservative for dinucleotide substitutions, depending on the direction of affinity change.

To check whether these biases (substitutions on CpG and dinucleotide substitutions) affect the pattern we found, first, we split all CTCF biding sites into two categories: sites with neither

CpG substitutions nor dinucleotide substitutions, and sites with either CpG substitutions or dinucleotide substitutions. For both categories, the proportion of positive selection binding sites (PBSs) detected is highly correlated with the original pattern (Figure S17A and B). In addition, as expected, there is a higher proportion of PBSs for sites without substitions on CpG, comfirming that our test is conservative for sites with CpG substitutions. Second, we both excluded all CpG sequences and dinucleotide substitution sequences from all binding sites, and we integrated the transition and transversion rate (4:1, estimated from Figure S15A) into our null model. Patterns of results were very robust to these changes (Figure S17C).

To test whether the high regulatory adaptive evolution in brain is general to mammals, we performed the same analysis on CTCF binding data from 11 mouse adult tissues (Table S2; Figure S18). We investigated adaptive evolution in the *M. musculus* branch after divergence from *M. spretus*, a similar evolutionary divergence as that between human and chimpanzee (Enard et al. 2002). Similarly to human, we detected 3.54% binding sites which evolved under positive selection (Figure S10B) and found PBSs associated with a lower number of active cell types (Figure S12B). However, no tissue type had especially high adaptive evolution, and brain related tissues were among the lowest (Figure 4B). When restricting to tissue specific CTCF binding sites, lung has notably high adaptive evolution (Figure S13B).

Discussion

A robust test for positive selection on regulatory elements

Detecting positive selection on regulatory sequences has long been a difficult problem (Zhen and Andolfatto 2012). Nearby non-coding regions are often used as a neutral reference (Andolfatto 2005; Haygood et al. 2007; Arbiza et al. 2013), but such neutrality is difficult to establish. Our approach does not require defining a priori neutral sites, but instead considers the effects of variation on activity (Berg et al. 2004; Moses 2009; Smith et al. 2013). Moreover, positive selection on a background of negative selection might not elevate the evolutionary rate above the neutral expectation, yet consistent changes in binding affinity can still be detectable. Indeed, the TFBSs of cell types detected under selection do not necessarily evolve faster (Figure S19). In principle, our method can also be applied to other genomic regions for which experimental peaks are available, such as open chromatin regions or histone modification regions.

Because positive selection on regulatory sequences is difficult to determine, it is important to validate our predictions with independent evidence. The most important validation is that predictions made independently of population data verify the expectations of higher substitution to polymorphism ratio (Figure 3D). Both this and the lower expression variance of neighboring genes (Figure 3E) are consistent with the prediction that positive selection will increase divergence but remove polymorphism (McDonald and Kreitman 1991), and that recently selected phenotypes will be under stronger purifying selection. Moreover, binding affinity change occurs in the direction predicted by our model (Figure 2E-G), and we can verify the prediction that pleiotropy limits adaptation (Wagner and Zhang 2011) (Figure S12).

Despite its advantages, our method can still be improved. For example, in the null model of sequence evolution, we assume independent mutation patterns at each base-pair site and a uniform mutation rate over all sites. But both mutation rate and pattern can depend on neighboring nucleotides (Krawczak et al. 1998). These limitations of our null model might explain why the observed p-values do not quite follow the expected uniform distribution for high values.

The importance of regulatory adaptation on human brain evolution

Our results support the long proposed importance of adaptive regulatory changes in human brain evolution (King and Wilson 1975). They are remarkably consistent with accelerated gene expression evolution in the human brain, but neither in human blood or liver, nor in rodents, from Enard et al. (2002). Previous studies on human regulatory sequence evolution reported acceleration in brain related functions, but could not demonstrate adaptive evolution nor direct activity in the brain (Enard et al. 2002; Pollard et al. 2006; Prabhakar et al. 2006; Haygood et al. 2007; Gittelman et al. 2015). The reported link between human accelerated regions and function was very indirect, depending both on the attribution of a region to the closest gene, and on the functional annotation of that gene.

The brain related cell types for which we detect a high proportion of positive selection are functionally related with cognitive abilities. For example, for astrocyte, abnormal astrocytic signaling can cause synaptic and network imbalances, leading to cognitive impairment (Santello et al. 2019). In addition, for choroid plexus epithelial cell, its atrophy has been reported to be related with Alzheimer disease (Kaur et al. 2016).

While we did not find a similar pattern by applying the same analysis to mouse, it isn't possible yet to conclude to a human or primate specific pattern. Indeed, the mouse analyses have two potential caveats. First, for the olfactory bulb and cortical plate in the mouse analyses, there are no corresponding anatomical structures in the human analyses. It is an open question whether the human olfactory bulb and cortical plate also have high adaptation. Second, the human analyses were based on ChIP-seq at cell type level but the mouse analyses were based on ChIP-seq at tissue level. In mouse, the astrocyte in cerebellum may also have high adaptation like the astrocyte in human, but the signal might be diluted by other cell types in cerebellum.

Regulatory adaptation differs between tissues

Outside of brain cell types, we found that male reproduction system (prostate and foreskin) has higher adaptive regulatory evolution than female reproduction system (ovary, uterus and vagina). This is consistent with the observation of high adaptive sequence evolution in human male reproduction (Wyckoff et al. 2000; Nielsen et al. 2005), and probably caused by sexual selection related selective pressures, such as sperm competition. However testis has a relatively low proportion of adaptive evolution, similar to ovary. This suggests that the high expression divergence in testis (Brawand et al. 2011) is mainly caused by relaxed purifying selection, maybe due to the role of transcription in testis for 'transcriptional scanning' (Xia et al. 2020). Outside of the brain, the top adaptive regulatory evolution systems seem to be the same as found for adaptive protein evolution, i.e. male reproduction, immune and endocrine systems (Clark et al. 2003; Bustamante et al. 2005; Nielsen et al. 2005; Daub et al. 2017). The high fraction of substitutions fixed by positive selection in skin is interesting (Figure S14), since skin is both involved in defense against pathogens, and has evolved specifically in the human branch with loss of fur (Brettmann and de Guzman Strong 2018). The lack of adaptive protein sequence evolution despite high adaptive regulatory evolution might be related to selective pressure on proteins in the brain (Drummond and Wilke 2008; Roux et al. 2017).

Materials and Methods

Code and data availability

Data files and analysis scripts are available on GitHub:

https://github.com/ljljolinq1010/A-robust-method-for-detecting-positive-selection-on-regulatory-sequences

All data analyzed during this study are available from public databases listed under each dataset in the relevant Materials and Methods section.

Mutagenesis for positive selection

1. Training of the gapped k-mer support vector machine (gkm-SVM)

gkm-SVM is a method for regulatory DNA sequence prediction by using *k*-mer frequencies (Ghandi et al. 2014) . For the gkm-SVM training, we followed the approach of Lee et al. (2015). Firstly, we defined a positive training set and its corresponding negative training set. The positive training set is ChIP-seq narrow peaks of transcription factors. The negative training set is an equal number of sequences which randomly sampled from the genome with matched the length, GC content and repeat fraction of the positive training set. This negative training set was generated by using "genNullSeqs", a function of gkm-SVM R package (Ghandi et al. 2016). Then, we trained a gkm-SVM with default parameters except -l=10 (meaning we use 10-mer as feature to distinguish positive and negative training sets). The classification performance of the trained gkm-SVM was measured by using receiver operating characteristic (ROC) curves with fivefold cross-validation. The gkm-SVM training and cross-validation were achieved by using the "gkmtrain" function of "LS-GKM : a new gkm-SVM software for large-scale datasets" (Lee 2016). For details, please check https://github.com/Dongwon-Lee/lsgkm.

2. Generate SVM weights of all possible 10-mers

The SVM weights of all possible 10-mers were generated by using the "gkmpredict" function of "LS-GKM". The positive value means increasing binding affinity, the negative value means decreasing binding affinity, the value close to 0 means functionally neutral.

3. Infer ancestor sequence

The ancestor sequence was inferred from sequence alignment with a sister species and an outgroup.

4. Calculate deltaSVM

We calculated the sum of weights of all 10-mers for ancestor sequence and focal sequence respectively. The deltaSVM is the sum weights of focal sequence minus the sum weights of

ancestor sequence. The positive deltaSVM indicating substitutions increased the binding affinity in the focal sequence, vice versus.

5. Generate Empirical Null Distribution of deltaSVM

Firstly, we counted the number of substitutions between the ancestor sequence and the focal sequence. Then, we generated a random pseudo-focal sequence by randomly introducing the same number of substitutions to the ancestor sequence. Finally, we calculated the deltaSVM between the pseudo-focal sequence and the ancestor sequence. We repeated the above processes 10000 times to get 10000 expected deltaSVMs.

6. Calculate p-value of deltaSVM

For lineage specific gain TFBSs, the p-value was calculated as the probability that the expected deltaSVM is higher than the observed deltaSVM (higher-tail test). For lineage specific lose TFBSs, the p-value was calculated as the probability that the expected deltaSVM is lower than the observed deltaSVM (lower-tail test). For conserved TFBSs, we primarily focused on selection to increase binding affinity, and thus we performed higher-tail test. The motivation for this is that when we have ChIP-seq data in only one species, which is the most common case, the observed peaks are a mix of conserved and gained sites, and thus very little signal of decrease of binding is expected. The p-value can be interpreted as the probability that the observed deltaSVM could arise by chance.

Mouse validation analysis

1. ChIP-seq data

The narrow ChIP-seq peaks and their corresponding intensity (normalized read count) datasets of three liver specific transcription factors (CEBPA, FOXA1 and HNF4A) in three mouse species (C57BL/6J, CAST/EiJ, SPRET/EiJ) were downloaded from https://www.ebi.ac.uk/research/flicek/publications/FOG09 (Stefflova et al. 2013, accessed in May, 2018). Peaks were called with SWEMBL (http://www.ebi.ac.uk/~swilder/SWEMBL). To account for both technical and biological variabilities of peak calling, Stefflova et al. (2013) carried out the following approaches. For each transcription factor in each species, they first called three sets of peaks: one for each replicate (replicate peek), and one for a pooled dataset of both replicates (pooled peak). Then, the peaks detected from the pooled dataset were used as a reference to search for overlaps in the two other replicates. When a pooled peak overlapped with both replicate peeks (at least one base pair overlap), it was kept for downstream analyses. For the number of peaks and average peak length, please check Table S1.

2. Peak coordinates transfer

Based on pairwise genome alignments between C57BL/6J and CAST/EiJ or SPRET/EiJ, Stefflova et al. (2013) transferred the coordinates of ChIP-seq peaks in both CAST/EiJ and SPRET/EiJ to its corresponding coordinates in C57BL/6J.

3. Sequence alignment files

The sequence alignment files between C57BL/6J and CAST/EiJ or SPRET/EiJ were downloaded from https://www.ebi.ac.uk/research/flicek/publications/FOG09 (Stefflova et al. 2013, accessed in May, 2018).

4. Define different types of binding sites

1) Conserved binding sites

The conserved binding sites were defined as peaks in C57BL/6J which have overlapping peaks (at least one base pair overlap) in the other two species by genome alignment.

2) Lineage specific gain binding sites

The lineage specific gain binding sites defined as peaks in C57BL/6J with no overlapping peaks (at least one base pair overlap) in the other two species.

3) Lineage specific loss binding sites

The lineage specific loss binding sites defined as peaks in CAST/EiJ which overlapping peaks in SPRET/EiJ but not in C57BL/6J.

Human validation analysis

1. ChIP-seq data

The narrow ChIP-seq peaks datasets of two liver specific transcription factors (CEBPA and HNF4A) in human were downloaded from https://www.ebi.ac.uk/research/flicek/publications/FOG01 (Schmidt et al. 2010, accessed in October, 2018). Peaks were called with SWEMBL (http://www.ebi.ac.uk/~swilder/SWEMBL). Negligible variation was observed between the individuals in terms of peak calling, so Schmidt et al. (2010) pooled replicates into one dataset for peak calling.

2. Sequence alignment files

The pairwise whole genome alignments between human and chimpanzee or gorilla were downloaded from http://hgdownload.soe.ucsc.edu/downloads.html (accessed in December, 2018).

3. Single nucleotide polymorphism (SNP) data

Over 36 million SNPs for 1,092 individuals sampled from 14 populations worldwide were downloaded from phase I of the 1000 Genomes Project ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ (1000 Genomes Project Consortium 2012, accesed in December, 2018). As suggested by Luisi et al. (2015), we only used SNPs of a subset of 270 individuals from YRI, CEU, and CHB populations.

4. Liver expression data

The library site normalized expression data of 175 livers was downloaded from downloaded from The Genotype Tissue Expression (GTEx) project https://gtexportal.org/home/ (GTEx Consortium 2017, Release V7, accessed in December, 2018). We further transformed it with $\log_2$.

5. Putative target genes of TFBSs

We assigned the nearest gene to each TFBS as its putative target gene.

6. Adjusted variance

There is a very strong dependency between mean and variance for gene expression (Figure S20A). To remove this dependency, as previously proposed (Barroso et al. 2018; Liu et al. 2019), we calculated the adjusted variance. Specifically, we fitted a polynomial model to predict the variance from the mean in the log space. We increased the degrees of the model until there was no more significant improvement (tested with ANOVA, $p<0.05$ as a significant improvement). The adjusted variance is the ratio of the observed variance over the variance component predicted by the mean expression level. After this adjustment, there is no correlation between mean and variance (Figure S20B).

Fly validation analysis

1. ChIP-seq data

The narrow ChIP-seq peaks of transcription factor CTCF in three drosophila species (*D. melanogaster, D. simulans* and *D. yakuba*) were downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24449 (Ni et al. 2012, accessed in January, 2019). Peaks were called with QuEST(Valouev et al. 2008) at a False Discovery Rate (FDR) <1%. We obtained 2182, 2197 and 2993 peaks with average length of 243bp, 240bp and 201bp for *D. melanogaster, D. simulans* and *D. yakuba* respectively.

2. Peak coordinates transfer

The peaks identified in *D. simulans* and *D. yakuba* were translated onto *D. melanogaster* coordinates by using pslMap (Zhu et al. 2007).

3. Sequence alignment files

The pairwise whole genome alignments between *D. melanogaster* and *D. simulans* or *D. yakuba* were downloaded from http://hgdownload.soe.ucsc.edu/downloads.html (accessed in January, 2019).

4. Define different types of binding sites. These were defined as in mouse, using *D. melanogaster* vs. *D. simulans* and *D. yakuba*.

Human CTCF analysis

1. ChIP-seq data

The narrow ChIP-seq peaks of transcription factor CTCF across 29 tissues or cell types (Table S2) were downloaded from ENCODE (The ENCODE Project Consortium 2012). We did not use ChIP-seq datasets from cell lines, and only kept ChIP-seq datasets from tissues and primary cells. Briefly, peaks were called with MACS (Zhang et al. 2008) separately for each replicate. Irreproducible Discovery Rate (IDR) analysis was then performed (Li et al. 2011). Final peaks are the set of peak calls that pass IDR at a threshold of 2%. Peaks identified in different tissues or cell types were integrated by intersecting all peaks across datasets, with at least one base pair overlap used as the merge criteria. Overall we obtained 118970 merged peaks.

2. Sequence alignment files

The pairwise whole genome alignments between human and chimpanzee or gorilla were downloaded from http://hgdownload.soe.ucsc.edu/downloads.html (accessed in December, 2018).

3. Proportion of substitutions fixed by positive selection

We calculated the Proportion of substitutions fixed by positive selection, a measure of effect size of selection, under the MK test framework (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002):

$$\alpha = 1 - \frac{DnpPp}{DpPnp}$$

Dnp is the substitution number in non-PBSs; *Pp* is the polymorphism number in PBSs; *Dp* is the substitution number in PBSs; *Pnp* is the polymorphism number in non-PBSs.

4. Estimate substitution rate

The substitution rate, for example C -> T, was estimated as the number of C -> T divided by the number of nucleotide C in the ancestor sequence.

Mouse CTCF analysis

1. ChIP-seq data

The narrow ChIP-seq peaks of transcription factor CTCF across 11 tissues (Table S2) were downloaded from ENCODE (The ENCODE Project Consortium 2012). Briefly, peaks were called with MACS (Zhang et al. 2008) separately for each replicate. Irreproducible Discovery Rate (IDR)) analysis was then performed. Final peaks are the set of peak calls that pass IDR at a threshold of 2%. Peak identified in different tissues/cell types were integrated by intersecting all peaks across data sets, with at least 1 base pair overlap used as the merge criteria. Overall we obtained 112657merged peaks.

2. Sequence alignment files

The sequence alignment file between C57BL/6J and SPRET/EiJ, please check "Mouse validation analysis" part of Materials and Methods. The sequence alignment file between C57BL/6J and Caroli/EiJ were downloaded from https://www.ebi.ac.uk/research/flicek/publications/FOG09 (Stefflova et al. 2013, accessed in May, 2018).

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–1152.

Anon. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A. 2013. Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. 45:723–729.

Barroso GV, Puzovic N, Dutheil JY. 2018. The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. Genetics 208:173–189.

Berg J, Willmann S, Lässig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4:42.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Brettmann EA, de Guzman Strong C. 2018. Recent evolution of the human skin barrier. Exp. Dermatol. 27:859–866.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, et al. 2003. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. Science 302:1960–1963.

Coyne JA. 1996. Genetics of a difference in male cuticular hydrocarbons between two sibling species, Drosophila simulans and D. sechellia. Genetics 143:1689–1698.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46:D794–D801.

Drummond DA, Wilke CO. 2008. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell 134:341–352.

Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, et al. 2002. Intra- and interspecific variation in primate gene expression patterns. Science 296:340–343.

Franchini LF, Pollard KS. 2015. Can a few non-coding mutations make a human brain? Bioessays 37:1054–1061.

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

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput. Biol. 10:e1003711.

Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 32:2205–2207.

Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, Hawkins RD, Akey JM. 2015. Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res. 25:1245–1255.

Gronau I, Arbiza L, Mohammed J, Siepel A. 2013. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. Mol. Biol. Evol. 30:1159–1171.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. Nature 550:204–213.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat. Genet. 39:1140–1144.

Kaur C, Rathnasamy G, Ling E-A. 2016. The Choroid Plexus in Healthy and Diseased Brain. J. Neuropathol. Exp. Neurol. 75:198–213.

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

Krawczak M, Ball E V., Cooper DN. 1998. Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes. Am. J. Hum. Genet. 63:474–488.

Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics 32:2196–2198.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. 47:955–961.

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 5:1752–1779.

Liu J, Frochaux M, Gardeux V, Deplancke B, Robinson-Rechavi M. 2019. Selection against expression noise explains the origin of the hourglass pattern of Evo-Devo. bioRxiv: https://doi.org/10.1101/700997.

Ludwig MZ, Kreitman M. 1995. Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Mol. Biol. Evol. 12:1002–1011.

Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H. 2015. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol. Evol. 7:1141–1154.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Moses AM. 2009. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. BMC Evol. Biol. 9:286.

Ni X, Zhang YE, Nègre N, Chen S, Long M. 2012. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome. PLoS Biol 10:1001420.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol. 3:e170.

Orr HA. 1998. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. Genetics 149:2099–2104.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. Cell 137:1194–1211.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006. Forces Shaping the Fastest Evolving Regions in the Human Genome. PLoS Genet. 2:e168.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated Evolution of Conserved Noncoding Sequences in Humans. Science 314:786–786.

Roux J, Liu J, Robinson-Rechavi M. 2017. Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. Mol. Biol. Evol. 34:2773–2791.

Santello M, Toni N, Volterra A. 2019. Astrocyte function from information processing to cognition and cognitive impairment. Nat. Neurosci. 22:154–166.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S,

Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328:1036–1040.

Smith JD, McManus KF, Fraser HB. 2013. A Novel Test for Selection on cis-Regulatory Elements Reveals Positive and Negative Selection Acting on Mammalian Transcriptional Enhancers. Mol. Biol. Evol. 30:2509–2518.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature 415:1022–1024.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell 154:530–540.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods 5:829–834.

Varki A, Altheide TK. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res. 15:1746–1758.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat. Rev. Genet. 12:204–213.

Wyckoff GJ, Wang W, Wu C-I. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309.

Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, Kim SY, Keefe DL, Alukal JP, Boeke JD, et al. 2020. Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. Cell 180:248-262.e21.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9:R137.

Zhen Y, Andolfatto P. 2012. Methods to detect selection on noncoding DNA. Methods Mol. Biol. 856:141–159.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage. PLoS Comput. Biol. 3:e247.

Acknowledgements

We thank Jérôme Goudet, Gunter Wagner, David Garfield, Laurent Duret, and members of the Robinson-Rechavi lab for helpful discussions. Part of the computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. JL and MRR are supported by Swiss National Science Foundation grants 31003A_153341 / 1 and 31003A_173048.

Conflict of Interests

The authors declare that they have no conflict of interest.

Author contributions

JL designed the work with input from MRR. JL performed data collection and computational analyses. JL and MRR interpreted the results. JL wrote the first draft of the paper. JL and MRR finalized the paper.