Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor REST.

Research Article

Rory Johnson[1]*, John Samuel[2], Calista Keow Leng Ng[3], Ralf Jauch[3], Lawrence W. Stanton[1] and Ian C. Wood[2]*.

1. Stem Cell and Developmental Biology Group, Genome Institute of Singapore, Singapore 138672.
2. Institute of Membrane and Systems Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK.
3. Laboratory of Structural Biochemistry, Genome Institute of Singapore, Singapore 138672.

* Corresponding authors.

Running Head: Evolution of the REST Regulatory Network.

Keywords: REST, NRSF, RE1, evolution, transcription factor binding, motif, gene regulation, neural gene, primate, network, primate-specific, human-specific, lineage-specific.

Abstract

Specific wiring of gene regulatory networks is likely to underlie much of the phenotypic difference between species, but the extent of lineage-specific regulatory architecture remains poorly understood.. The essential vertebrate transcriptional repressor REST (RE1-Silencing Transcription Factor) targets many neural genes during development of the preimplantation embryo and the central nervous system, through its cognate DNA motif, the RE1 (Repressor Element 1). Here we present a comparative genomic analyses of REST recruitment in multiple species by integrating both sequence and experimental data. We use an accurate, experimentally-validated Position-Specific Scoring Matrix (PSSM) method to identify REST binding sites in multiply-aligned vertebrate genomes, allowing us to infer the evolutionary origin of each of 1298 human RE1 elements. We validate these findings using experimental data of REST binding across the whole genomes of human and mouse. We show that one-third of human RE1s are unique to primates: these sites recruit REST *in vivo*, target neural genes, and are under purifying evolutionary selection. We observe a consistent and significant trend for more ancient RE1s to have higher affinity for REST than lineage-specific sites and to be more proximal to target genes. Our results lead us to propose a model where new transcription factor binding sites are constantly generated throughout the genome; thereafter, refinement of their sequence and location consolidates this remodelling of networks governing neural gene regulation.

Abbreviations: ChIP – Chromatin Immunoprecipitation; PSSM – Position Specific Scoring Matrix; TSS – Transcription Start Site.

Introduction

What is the genomic basis for phenotypic variation between species? To date, most studies on genome evolution have focused on the evolution of protein-coding DNA regions, where mutations can give rise to protein products with altered physicochemical properties. Instances of primate- and human-specific innovation in protein sequences have been described, often in genes with highly-suggestive roles in brain size and language [1-3], as well as immunity [4,5], hair [6] and reproduction [5]. However, the majority of genomic DNA does not encode protein, but contains numerous elements which direct organismal phenotype by regulating gene expression [7]. Sequence evolution of such non-coding DNA can thus lead to alterations in gene regulation, and thence to reorganization of gene regulatory networks [8]. With the recent publication and alignment of high-quality genome sequences and, more recently, comparative transcription factor binding datasets from multiple organisms, it is now becoming possible to interrogate non-coding DNA evolution on a comprehensive genome-wide basis and discover the non-conserved regulatory elements responsible for lineage-specific phenotypes.

Prior to the era of comparative genomics it was proposed that evolution in gene regulatory networks is a major underlying cause of phenotypic differences between species [9,10]. While the concept of regulatory evolution is broadly accepted and has been explored in detail [8,11] little evidence exists due to the technical difficulty in making comparative genomic measurements of transcriptional regulation, particularly in metazoans. Consequently, debate continues as to the relative contributions to evolution from regulatory versus coding DNA mutation [12]. Nevertheless the data available do point to widespread intra-specific variation in gene regulatory systems: these include well-documented instances of regulatory evolution underlying phenotypic change in yeast [13], Drosophila [14] and stickleback [15], divergent transcription factor targeting in equivalent human and mouse tissues [16,17], and distinct gene expression profiles in the brains of human compared to closely-related primates [18,19]. It is anticipated that mutations in *cis* regulatory DNA sequences such as transcription factor binding sites are the major cause of such divergence [20], but until the recent availability of relevant whole-genome datasets, this hypothesis has been impossible to test in a comprehensive manner.

Combinatorial recruitment of activating and repressive transcription factors directs precise spatial and temporal gene expression patterns, particularly during development, and such regulation is frequently disrupted in disease states. The specific recruitment of transcription factors to their target genes is mediated by non-coding DNA sequence elements known as transcription factor binding sites (TFBS). Ranging in length from around 4 – 21 bp, such elements recruit transcription factors in a sequence-dependent manner, where the degree of recruitment *in vivo* correlates with the similarity of the element to an ideal motif [21,22]. It is likely that various mechanisms contribute to the evolutionary generation and loss of new regulatory motifs, including insertions, deletions and substitution of DNA bases, as well as genomic rearrangements such as DNA duplication [8]. Intriguingly, growing evidence implicates transposable elements in copying and inserting pre-existing transcriptional regulatory motifs into novel target genes [23-26]. Notably, this process seems to be particularly prevalent amongst the longest motifs, which are the least likely to appear by random DNA mutation [27].

To date, attempts to comprehensively identify evolving functional non-coding DNA have relied on sequence-evolutionary analysis of multispecies aligned regions to detect regions under significant positive selection [28,29]. Not surprisingly, these studies have concentrated on identifying regions that distinguish the human genome from that of other primates, and have generally implicated neurodevelopmental genes to be amongst those to have experienced particularly accelerated regulatory evolution. Notably, Prabhakar and colleagues found positively-selected non-coding regions are significantly associated with genes mediating neural cell adhesion [28], supporting the notion that modifications to neurodevelopmental gene expression programs underlie human-specific cognitive abilities. While these studies have allowed us to make inferences based on the classification of likely target genes, they tell us little about the mechanism by which evolving regions contribute to new programmes of gene expression. For this we must have some prior understanding of the regulatory motifs in question.

An alternative approach, which we describe in the present study, is to search for functional evolution of non-coding DNA based on the sequence properties of a known motif. The nature of regulatory non-coding DNA, often consisting of short, degenerate motifs at highly variable locations with respect to the genes they target, makes the study of non-coding DNA evolution inherently challenging. One important exception is the vertebrate transcription factor Repressor Element 1-Silencing Transcription Factor (REST), which binds to a relatively long sequence motif called the Repressor Element 1 (RE1). The RE1 serves to represses transcription of many genes expressed in the nervous system by specifically recruiting REST, which serves as a platform for various repressive cofactor complexes [30-32]. This regulatory system is specific to vertebrates, and has been co-opted into diverse regulatory roles including neurodevelopment [33,34], vascular smooth muscle proliferation [35], heart development [36] and embryonic stem cell (ESC) pluripotency [17,37,38]. We previously reported a position-specific scoring matrix (PSSM) method to identify RE1 sequences and target genes in vertebrate genomes [24]. The RE1 PSSM is highly specific: the majority of predicted RE1s bind to REST *in vivo* [39] and thus can be used to predict REST binding sites in multiple available genomes. Using this approach to search the human genome, we previously found evidence for extensive duplication of pre-existing RE1s located within transposable elements, suggesting that the REST regulatory network has been dynamically remodeled by retrotransposition during evolution.

In this present study we have used the RE1 PSSM to identify orthologues of all human RE1s in 16 vertebrate genomes, in order to investigate the evolutionary divergence of the transcriptional network controlled by REST. Due to the unique availability of whole-genome, *in vivo* binding data for REST in multiple species ([39] and [40]), we were able to validate the majority of our predicted lineage-specific RE1s. We show that new REST binding sites have arisen throughout vertebrate evolution, generating species-specific REST regulatory targets. RE1s of various evolutionary ages have characteristic properties of affinity and proximity to target genes: overall, ancient RE1s are more proximal to target genes and recruit REST with higher affinity. Nevertheless, recently-evolved primate-specific RE1s target neural genes and are capable of recruiting REST *in vivo*. Confirming their functional significance, these primate-specific motifs have been under purifying evolutionary selection during primate and human evolution. This study represents the first integrated, genome-wide analysis of a transcriptional regulatory motif across multiple vertebrate species, providing new insights into the extent and mechanism of transcriptional network evolution amongst vertebrates.

Materials and Methods

Identification of RE1s in Vertebrate Genomes. Human (version 36.1, hg18), chicken (galGal3), chimp (panTro2), cow (bosTau4), dog (canFam2), fugu (fr2), mouse (mm6), opossum (monDom4), rat (rn4), macaque (rheMac2) and xenopus (xenTro2) genome sequences were downloaded from the UCSC Genome Browser at (http://genome.ucsc.edu/) and searched for RE1s as described previously using the 21bp RE1 position specific scoring matrix (PSSM) in conjunction with the C program, Seqscan [24]. Searches were carried out using a constant nucleotide ratio background (we used the human nucleotide ratios : A/T – 0.295, C/G, - 0.205) The RE1 count is relatively insensitive to the %GC backgrounds that are found within the genomes tested (Supplementary Figure 1). Seqscan assigns a score from 0 to 1 to each 21-mer on the basis of its similarity to the RE1. Based on experimental data, we previously defined a stringent cutoff score of 0.91, above which sequences were considered to be bona fide RE1s. Genomes were not masked for repeat elements prior to scanning. To estimate the false positive rate of RE1 identification, we generated a randomized 3Gb genome using human nucleotide frequencies and scanned it using the RE1 PSSM. Using human/fugu (A 0.273 C 0.227 G 0.227 T 0.273)/opossum (A 0.312 C 0.188 G 0.188 T 0.312) nucleotide frequencies we identified 32/37/30 RE1s above cutoff – suggesting that our human genome-wide RE1 search has a false-positive rate of the order of 2.5%.

Evolutionary Classification of Human RE1s. All RE1 conservation analysis was carried out in comparison to the 1298 above-cutoff RE1s in the human genome. To classify RE1s based on evolutionary conservation, we performed the following analysis: MultiZ multiple alignments of 16 vertebrate genomes with human (hg18) were downloaded from the UCSC Genome Browser. [41]. These alignments consist of aligned sequence blocks of variable length. Using the locations of human RE1s, a custom Perl script was used to extract the orthologous aligned regions (where available) of other genomes. Gaps were removed from aligned sequences. These aligned sequences were scored with the RE1 PSSM. The highest scoring 21mer was recorded, and those cases having a motif scoring above cutoff were considered to represent an orthologous RE1. Species where no alignment exists, or where aligned sequence contains a highest-scoring motif below the cutoff score, were considered to have no orthologous RE1. Every human RE1 was then classified by the species having aligned RE1s: those with no aligned RE1s were classified as "Human Specific"; those with an aligned RE1 in at least one of chimp (panTro1) or macaque (rheMac2) but in no other species, were classified as "Primate Specific"; similarly, for mouse (mm8), rat (rn4), rabbit (oryCun1), dog (canFam2), cow (bosTau2), armadillo (dasNov1), elephant (loxAfr1), tenrec (echTel1) or opossum (monDom4) – "Mammal Specific"; for chicken (galGal2) – "Reptile Specific"; for xenopus (xenTro1) – "Amphibian Specific"; and for zebrafish (danRer3), tetraodon (tetNig1) or fugu (fr1) – "Deeply Conserved". Lineage-specific classifications of RE1s are available in the Supplementary Data File 2.

We carried out a similar analysis using five control RE1-like matrices, employing the same PSSM cutoff score of 0.91. In each case, the positions of the RE1 PSSM were randomly shuffled; resulting matrices did not have CpG propensity greater than the original RE1 motif. Although between 55 and 120 instances of high-scoring matches were found in the human genome for five shuffled PSSMs, in no case did we discover a single orthologous motif in any aligned species, indicating that the false-positive rate of multispecies conserved RE1 motifs predicted by our method is negligible.

Target Gene Annotation. Target genes were defined as the unique Refseq-annotated gene having its transcriptional start site most proximal to either an RE1, or to the central basepair of an experimentally-defined binding region. Refseq gene coordinates for hg18 and mm9 were downloaded using the Biomart tool from Ensembl (www.ensembl.org/biomart).

Comparison of RE1s to Experimentally Determined REST Binding Sites. High-throughput sequencing-based experimental datasets of REST binding from human (Jurkat cell line, as determined by ChIP-seq) [39] and mouse (E14 embryonic stem cells, as determined by ChIP-PET) [40] were downloaded and converted to BED coordinate format. For ChIP-PET data, the minimal overlap region of each PET cluster was used. BED coordinates were compared to each other using a custom Perl script. Coordinates were converted between genomes using the UCSC LiftOver tool on default settings.

GO Analysis. Gene ontology (GO) analysis was carried out on nonredundant sets of human or mouse Refseq genes, using the PantherDB tool (www.pantherdb.org). Unless indicated, P-values were corrected for multiple hypothesis testing using the Bonferroni method.

Nucleotide divergence in RE1s. BLASTZ pairwise alignments of hg18-rheMac2 and hg18-mm9 were obtained from GALAXY [42] for either the two half-sites of the RE1 (positions 1-9 and 12-17 of RE1 motif), or to two 100bp flanking regions extending from either extremity of the RE1 motif. The baseml program from PAML [43] was used to calculate the numbers of identical and divergent nucleotides. Uncorrected divergence rates were estimated by dividing the number of divergent nucleotides by the total number of nucleotides. Although this method is likely to underestimate nucleotide substitution rates for highly diverged species [44], this should not affect our conclusions in testing substitution rates differences between regions. Statistical significance was estimated by comparing the total numbers of divergent nucleotides to conserved nucleotides between RE1 and flanking regions, using the Pearson Chi-Square test with continuity correction (relevant contingency tables can be found in the Supplementary Figure 3). An equivalent control region was defined 1kb upstream of every primate-specific RE1, and the analysis repeated.

Assessing RE1 conservation with MONKEY. The MONKEY program [45] was run under Cygwin with default settings on BLASTZ pairwise alignments downloaded from GALAXY (hg18-rheMac2, hg18-mm9) [42]. We used the same RE1 PSSM described in [24], or randomly-shuffled equivalents, except that all values were converted to frequencies, and a pseudocount of 0.01 was added to positions containing a zero in the original PSSM. For every scan, the resulting hit with the lowest P-value was recorded. We scanned alignments containing predicted RE1s (including 100bp flanking DNA up- and downstream) with the RE1 PSSM. To estimate background, we rescanned the whole-genome human-macaque or human-mouse alignments for each of 100 shuffled PSSMs. In each case, the median MONKEY-reported P-value across all predicted binding sites in the genome was recorded.

SNP Density. Data from dbSNP129 were downloaded and filtered to remove all features extending for >1bp. Data were analysed essentially as described for nucleotide divergence above.

Production of recombinant REST protein. The DNA binding domain of REST was amplified by PCR from a human cDNA (IMAGE:40146881; NCBI accession: BC132859) using the

following DNA Oligos 5'-
GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAAACCTGTATTTTCAGGGCGCGGAGGACAAAGGCAAGAG, 3'-
GGGGACCACTTTGTACAAGAAAGCTGGGTTTAATTATCAGGCAAGTCAGCCTC.
The pENTR-hREST/DBD plasmid was generated by GATEWAY BP cloning (Invitrogen) encoding residues 147-440 of the full length REST protein. A plasmid suitable for bacterial expression was generated by performing a GATEWAY LR reaction using pDEST-HISMBP as destination vector [46]. BL21(DE3) cells transformed with the expression plasmid were grown in terrific broth (TB) at 37˚C to an optical density of 0.5-0.8, whereupon protein expression was induced by addition of 0.2mM isopropyl β-D-1-thiogalactopyranoside and incubated at 17˚C for 5 hours. Cells were harvested by centrifugation and lysed by sonication. HisMBP-REST/DBD fusion protein was extracted from the bacterial lysate at 4°C using an amylose resin (New England Biolabs) following the manufacturers instructions. The protein was further purified on a HiPrep Superdex 200 16/60 size exclusion column (GE Healthcare) equilibrated with 10mM Tris-HCl pH 8.0, 100mM NaCl. Fractions containing HisMBP-REST/DBD were pooled, aliquoted and stored at -80˚C.

Electrophoretic mobility shift assay (EMSA). EMSA were essentially performed as described [47] with the following modifications. A 30bp Cy5 labeled RE1 element, adapted from the rat *Scn2a2* gene [32] was pre-mixed with a 100-fold excess of unlabelled competitor, then a master mix containing the HisMBP-REST/DBD protein was added. The final reaction, containing 1nM Cy5-labelled DNA, 100nM competitor DNA and 50nM protein in EMSA buffer (2mM β-mercaptoethanol, 10mM Tris-HCl pH 8.0, 100mM KCl, 50μM ZnCl2, 10% Glycerol, 0.1% NP-40 and 0.1mg/mL bovine serum albumin), was incubated at 4˚C for one hour then electrophoresed at 4˚C on a pre-run tris-glycine 5% polyacrylamide gel at 300V for 1 hour.

Results

RE1s Can Be Classified by Evolutionary Lineage

Previous data have suggested that the population of REST binding motifs (RE1s) has grown in concert with the expansion of the mammalian genome during evolution, and that specific duplication mechanisms may have contributed to this process [24,48]. Using a PSSM search for RE1 motifs, we find that the genomic population of RE1s in mammals to be approximately twice that of non-mammalian vertebrates (Supplementary Figure 4). Furthermore, inspection of multispecies alignments suggests that the degree of evolutionary conservation is highly heterogeneous amongst RE1s (Supplementary Figure 4). To investigate the degree of evolutionary conservation of RE1s in more detail, we developed a PSSM pipeline to comparatively identify RE1s across the genomes of multiple vertebrates (Fig 1A and Materials and Methods). Because our method relies on identifying RE1s from their sequence characteristics alone, and does not use sequence conservation as a criterion, it should be capable of isolating instances of newly-evolved RE1s in the human lineage. Furthermore, this method is unlikely to be affected by evolutionary changes in the RE1 motif itself, given the extreme conservation of the REST DNA-binding domain (DBD) amongst vertebrates (Supplementary Figure 5). Using this approach, we sought to determine the degree of evolutionary conservation of the 1298 human RE1 sites. The orthologous regions of every human RE1 was identified from multiple alignments of 16 other vertebrate species [49] and each of these regions was searched for RE1s with the RE1 PSSM. We thus were able to classify all human RE1s by their degree of evolutionary conservation; for instance, motifs which could be identified in the orthologous region of at least one other mammal - but not amongst birds, amphibians or fish – were classified as "mammal-specific". Similarly, binding sites with orthologues in either chimp or macaque or both, but not in any other species, were classified as "primate-specific" (Fig 1A).

This lineage analysis (displayed in the form of a heatmap in Fig 1B) showed that approximately two-thirds of human RE1s (869/1298 = 67%) are conserved in at least one other mammalian species (Fig 1C). Small numbers of ancient RE1s could be distinguished in orthologous regions of chicken (reptile-specific, 13/1298, 1%), xenopus (amphibian-specific, 7/1298, 0.5%) and fish genomes (24/1298, 2%). Most intriguingly, almost one-third of RE1s are specific to either human alone (40/1298, 3%), or to human and at least one other primate (345/1298, 27%). We also asked what proportion of RE1s are conserved between human and mouse, two high-coverage genomes of moderate evolutionary divergence. We found that just over one-third of human RE1s have an identifiable orthologue in mouse (476/1298, 37%) (Fig 1D). Similar proportions of human RE1s have an orthologous region in mouse containing no identifiable RE1 motif (416/1298, 32%), or have no orthologous region at all (406/1298, 31%).

Evolutionary turnover of transcription factor binding sites – that is, the compensatory gain and loss of a motif regulating a particular gene – is frequently observed [50]. We next wished to estimate what proportion of lineage-specific RE1s are truly lineage-specific (ie appeared at a locus that is not regulated by REST in more anciently diverged lineages), compared to those RE1s that evolved to replace a recently lost motif at a pre-existing target locus. This analysis showed that, even with very relaxed definitions of turnover, the majority of lineage-specific RE1

have arisen within loci that previously were not bound by REST (Supplementary Figure 6). This suggests that new RE1s evolve more frequently to regulate a new target gene, rather to regulate a pre-existing target gene.

Some examples of lineage-specific RE1s are shown in Figure 2. The *LHX3* gene (encoding a transcription factor involved in pituitary development [51]) was found to contain an upstream RE1 that is deeply conserved, having an orthologous sequence in zebrafish. The RE1 is conserved both in terms of its individual sequence identity, shown by alignment, as well as its functionality, inferred from the high RE1 PSSM score in these orthologous sequences. In contrast, *AGBL1* contains two recently evolved RE1s – a primate-specific site and a human-specific site. The primate specific site (above) apparently evolved through DNA sequence mutation since aligned DNA sequence exists in non-primate mammalian species. The putative human-specific RE1 of the gene is part of a primate-specific LINE1 insertion; we previously observed this phenomenon of RE1 insertion by retrotransposition [24].

REST target genes tend to have roles in neurodevelopment and neuronal function [24,48,52]. We observed that amongst the most proximal genes to primate-specific RE1s are a number with neural functions, including those encoding the synaptic neuromodulator Cerebellin 1, predicted signaling molecule IQSEC2, the RNA-binding protein LARP6 and the soluble decoy receptor, TNFRSF6B (Table 1). Many lineage-specific RE1s localise to pre-existing target genes that also have a more ancient motif: of the 313 predicted targets of primate-specific RE1s, 48 also have an older, mammalian-specific RE1 (this represents a 6-fold enrichment over that expected by chance: P=5E-23, Hypergeometric test). This suggests that at least some novel RE1 generation serves to refine regulation of existing REST target genes, rather than to acquire new ones.

Validation of Lineage Predictions *In Vitro*

To determine the affinity of predicted RE1s biochemically, we performed electrophoretic mobility shift assays (EMSA), to test for binding of the REST protein to a representative selection of lineage specific RE1s *in vitro*. Purified recombinant REST DNA binding domain (DBD) was mixed with a labeled probe representing the high-affinity RE1 of the mouse *Scn2a2* gene [32] (Fig 3A). Unlabelled oligonucleotides representing all available orthologous regions were tested for their ability to compete the DBD:probe interaction (Fig 3B and 3C).

First, we tested all orthologues of the deeply conserved RE1 from the promoter of *LHX3* (Fig 2A and 3C). We found that with the exception of the chicken, all sequences were capable of binding REST with high affinity, including orthologous sequences from *Tetraodon* and *Fugu* (Figs 3B and 3C). In the case of chicken, the orthologous sequence is drastically different from other vertebrates, likely resulting from either a misalignment, or from loss of this RE1 element in the chicken lineage (Fig 2A and 3A). Closer inspection of the chicken gene revealed that it contains an RE1 residing in chicken-specific sequence within the second intron. It is possible that this RE1 is the result of a genomic rearrangement within *LHX3* which is specific to birds.

We also tested a predicted primate-specific RE1 (derived from a LINE2 within an exon of *LARP6*, encoding a putative RNA-binding protein). We observed elevated binding of REST to human and macaque orthologues of the *LARP6* RE1, compared to rat, mouse and dog (Fig 3A

and 3C). Finally we tested the binding of three predicted human-specific RE1s from the *SFRS8*, *TSNARE1* and *AGBL1* genes (Fig 3C and 2B). Each of the predicted RE1s from the human genes were able to bind REST. Striking human-specific binding was observed for the human SFRS8 RE1, whose orthologous regions from chimp, macaque and dog showed no detectable affinity for REST. The chimp RE1 of *TSNARE* was bound by REST with significantly higher affinity than macaque, suggesting that the activity of this element has increased in the great apes. Finally, the predicted human specific RE1 from the *AGBL1* had markedly elevated affinity in human, although moderate binding was also observed in other primates (Fig 2B and 3C). In general, the affinity differences observed in the EMSA experiments can be attributed to sequence variations affecting key residues of the RE1 element: for example, a single substitution from a preferred cytosine to adenosine at Position 8 in the chimp *SFRS8* RE1 confers a large decrease in affinity compared to humans, while substitution of a preferred guanosine at position 13 to an adenosine or cytosine in macaque or dog respectively is the most likely reason for decreased affinity for the *SFRS8* RE1s in these species (Fig 3A). Together, *in vitro* binding studies provide experimental support for the existence of lineage specific RE1s that can be accurately predicted by our PSSM.

Differential *In Vivo* Recruitment to Lineage Specific RE1s

To comprehensively test the accuracy of RE1 evolutionary classifications, we validated the *in vivo* binding of lineage-specific RE1 motifs with reference to recently published whole-genome maps of REST binding in human [39] and mouse [40] determined by chromatin immunoprecipitation (ChIP). Overall, at least 70% (909/1298) of PSSM-predicted human RE1s recruit REST *in vivo* for the human dataset, compared to 68% (671/985) of RE1s predicted in mouse. Therefore the PSSM is consistent and selective in detecting sites which are bound by REST. Indeed this is likely to be an underestimate of PSSM accuracy as REST does not occupy every functional binding site in each cell type [53,54].

We divided human RE1s into categories based upon their conservation in the mouse genome, as in Figure 1D, where the mouse genome contains either orthologous RE1-containing sequence ("RE1"), orthologous non-RE1-containing sequence ("No RE1"), or no orthologous sequence ("No alignment", Fig 4A). This clearly showed that, while all human RE1s have similar capacity to recruit REST in human ES cells (ranging from 60–80%), only orthologous RE1s effectively recruit REST in mouse ES cells. Thus our sequence-based method is capable of correctly identifying lineage-specific sites, of which the majority recruit REST in vivo. Furthermore, despite our stringent PSSM cutoff score there is not an excessive number of false-negative orthologous RE1s in mouse – evidenced by the low number in the "No RE1" category that have evidence for binding in mouse.

Human RE1s from each lineage category, or their orthologous region in mouse, were compared to experimentally-determined REST binding locations (Fig 4B). One would expect that if an RE1 is specific to primates, then binding will be observed in human but not at the orthologous region in mouse, whereas ancient RE1s will be bound in both. Consistent with this, mammal-specific RE1s are bound by REST in both human and mouse cells, while primate- and human-specific RE1s are only observed to be bound in human cells. Interestingly, the latter two categories have a lower rate of REST binding (Human-specific: 23%; Mouse-specific: 44%) than RE1s as a whole (70%), suggesting that more recently evolved RE1s have less optimal characteristics for the

recruitment of REST in vivo. Thus our PSSM pipeline accurately predicts the functional conservation and ability to recruit REST to specific genes across multiple organisms.

Characteristic Properties of Lineage-Specific RE1s

The above data suggest that recently-evolved RE1s (human- and primate-specific) are less effective at recruiting REST than more ancient sites, even in human cells. Both human and mouse whole-genome REST maps are based on ChIP coupled to high-throughput sequencing: REST-bound DNA fragments are sequenced from a pool of immunoprecipitated DNA and mapped to the reference genome. A region of DNA that shows high affinity for REST will be precipitated efficiently and result in a greater number of tags than a region only weakly associated with REST. Thus, one expects that the number of ChIP-Seq tags should be correlated with the degree of REST recruitment *in vivo*. We quantified the mean ChIP-Seq tag count for human RE1s grouped by lineage (Fig 4C): mammal-specific RE1s have significantly higher numbers of overlapping sequences tags compared to primate-specific sites, suggesting that more ancient RE1s more effectively recruit REST in vivo. Ancient RE1s tend to have higher similarity to the canonical RE1 motif than more recent RE1s, as judged by their higher PSSM score (Fig 5A). Analysis of human and mouse ChIP datasets showed that , the probability of an RE1 being bound *in vivo* increases dramatically with increasing PSSM score between 0.91 to 0.95 though further increases in motif score have little effect (Supplementary Figure 7). However the extent of REST recruitment *in vivo* shows only a weak positive relationship possibly because other factors such as chromatin environment influence the level of REST recruitment(Supplementary Figure 8). In summary, the ability of an RE1 to recruit REST in the nucleus is positively correlated to its evolutionary age, and this is a result, at least in part, of more ancient RE1s having better quality RE1 sequence motifs.

We have recently found that the proximity of an RE1 to the transcriptional start site (TSS) of a gene is a strong determinant of the resultant transcriptional repression [40]. Interestingly, we found that proximity also correlates with evolutionary age: by calculating the distance of all RE1s to the nearest gene's TSS, we found that more ancient RE1s tend to reside proximal to gene TSS, while recent human- and primate- specific genes tend to reside distal from genes (Fig 5B).

DNA Sequence Evolution Drives Species-Specific Transcription Factor Recruitment

Comparison of sequencing-based maps of REST recruitment for human [39] and mouse [40] showed weak conservation of binding in orthologous regions: just 34% of human loci which recruit REST also do so at their orthologous loci in mouse (Fig 6A). Does DNA sequence underlie this poor conservation of transcription factor recruitment? If so, we might expect the genomic sequence of human-specific REST binding sites to have higher similarity to the RE1 motif compared to the orthologous region in mouse where REST is not found. To test this, we compared the RE1 PSSM scores of all regions underlying experimentally-validated binding sites from both human and mouse (Fig 6B and 6C). Those regions which are bound in both human and mouse have high PSSM scores in both genomes (green spots). Strikingly, sites bound in human only tend to have elevated PSSM scores in the human genome compared to that of mouse (blue spots) while sites bound in mouse only have elevated PSSM scores in the mouse genome (orange

spots) (Fig 6B). Additionally, using this functionally obtained set of REST binding sites, we compiled the distance of human-only and human-mouse conserved RE1s to the TSS of nearest genes in the human genome, and vice versa for mouse (Fig 6D). Consistent with our previous findings of predicted RE1 sites (Fig 5B), non-conserved REST binding sites tend to be more distal from target genes than conserved sites. Furthermore this is true for the mouse as well as the human genome; mouse specific sites show lower PSSM score and are further from the TSS of genes than conserved RE1 sequences in the mouse genome, and is likely to be applicable to all vertebrate genomes.

Evidence that Primate-Specific RE1s have been under Purifying Selection Since Human-Macaque Divergence

It is possible that primate-specific RE1s are simply neutrally-evolving, non-functional sequence elements that do not contribute to phenotype. One prediction of this hypothesis is that these non-functional elements would not be under evolutionary selection. To test this, we compared the rate of nucleotide substitution of primate-specific and mammal-specific RE1s to a region of flanking DNA that we assume to be neutrally evolving (Figs 7A and 7B). We carried out this analysis with alignments of RE1s to another primate genome, macaque, as well as to a non-primate mammal, mouse. Since divergence of human and macaque, we observe a statistically significant reduction in DNA substitution in the nucleotides of both primate-specific (P=2.2E-16) and mammal-specific RE1s (P=2.2E-16), compared to their immediate flanking DNA (Fig 7A). Consistent with primate-specific evolution, primate-specific RE1s show no evidence of negative selection in human-mouse alignments (P=2.2E-16) (Fig 7B).

A more sophisticated method for identifying evolutionarily-conserved DNA motifs in aligned DNA is represented by the program MONKEY [45]. Provided with a PSSM and a multiple sequence alignment, MONKEY identifies instances of a transcriptional regulatory motif and, using various substitution models, assigns statistical significance to their evolutionary conservation amongst the aligned sequences. One challenge in carrying out this analysis on recently diverged genomes, such as human and macaque, is the difficulty in distinguishing sequence elements that are truly conserved, from those that are neutrally evolving but have not yet had time to diverge. Therefore, we compared values of statistical significance of primate- and mammal-specific RE1s assigned by MONKEY, to the distribution yielded by a set of 100 randomly-shuffled RE1 PSSM searches which we expect to cover neutrally-evolving DNA. In pairwise alignments of human-macaque, both primate- and mammal-specific RE1s lay outside the shuffled motif distribution, indicating that they have significant signatures of evolutionary conservation compared to background (Fig 7C). This evidence for selection is almost completely lost when the primate-specific RE1 sites are compared to orthologous mouse DNA, consistent with their coming under negative selection much later after human-mouse divergence (Fig 7D). In summary, primate-specific RE1s have experienced reduced nucleotide substitution since divergence of human and primates, an observation which is consistent with their being functional and under purifying evolutionary selection since that time.

Reduced Diversity of RE1 Sequence within Humans

Evidence for recent negative selection can be inferred from reduced nucleotide diversity within human populations [55]. To test whether this is the case for RE1s, we examined the density of small nucleotide polymorphisms (SNPs) within RE1s. For human RE1s as a whole, we observed a significantly lower SNP density across the nucleotides that mediate REST binding, compared to 200bp of immediate flanking region (P=3.2E-4) (Fig 8A). We see a similar effect for primate-specific RE1s (P=5.0E-4) and mammal-specific RE1s (P=2.4E-12) (Fig 8B). These data are consistent with primate-specific RE1s being under negative selection during recent human evolution.

Discussion

Here we provide the first systematic analysis of a transcriptional regulatory motif across ~450 million years of vertebrate evolution, providing a genomic view of the evolution of a neural regulatory network. Our approach compared and integrated both bioinformatic motif identification and experimental chromatin immunoprecipitation (ChIP) data. This has been made possible by focusing on the neural gene regulation network commanded by REST – its cognate RE1 binding element is unusually long and can be identified with high confidence using probabilistic bioinformatic approaches [24,48,56{Johnson, 2008 #213] and importantly, REST is one of the first factors for which whole genome, sequencing-based ChIP datasets are available for comparison in both human and mouse [39,40]. As a result, we propose a model where new transcription factor binding sites are created by duplication/insertion followed by refinement of sequence and position, ultimately generating high affinity sites proximal to their target genes.

Emerging evidence, including that presented in this manuscript, point to highly divergent transcription factor recruitment between mammalian species [16,17]. What is the basis for this divergence? Many transcription factors bind short degenerate sequences which can be readily created by single base pair mutations of a similar sequence [27]. However, this is unlikely to be the case for transcription factors with long recognition elements, such as REST, p53 [57] or CTCF [58]: for simple probabilistic reasons, long periods of time must pass before long regulatory motifs can arise through DNA mutation in a given stretch of random sequence [27]. What processes can explain the genomic remodelling of transcriptional regulatory networks observed in vertebrates? Using the PSSM-predicted RE1s, or the experimentally-discovered REST binding sites, we observed simple yet significant differences between those binding sites which are conserved between species (more ancient) and those sites which arose more recently. A consistent trend emerges for more ancient sites to be closer to an ideal RE1 (as judged by PSSM score) and more proximal to their target gene. We previously showed that, in humans, expansion of RE1s into target genes has taken place through duplication by Alu, LINE1 and LINE2 retrotransposons [24]. This leads us to propose a model for the genomic basis of REST network evolution (Fig 9). New RE1 elements arise throughout the genome, in part mediated by retrotransposition. The majority of such novel RE1s have only low in vivo affinity for REST. Furthermore, due to the stochastic nature of insertions they are likely to be located distal from gene transcriptional start sites, (distal insertion in the primate lineage may be enhanced by the fact that LINE elements favour integration in AT-rich, gene-poor DNA [59]). Randomly-integrated RE1s are also likely to be found in regions of inactive chromatin, and may not be accessible to REST by virtue of occluding nucleosomes [60]. This is consistent with previous reports that REST recruitment is sensitive to chromatin environment [61], and that favourable nucleosome conformation requires the evolution of appropriate DNA sequence [62]. Thus the majority of such inserted RE1s will have little or no phenotypic effect and hence will be subject to gradual degeneration through neutral sequence mutation. However a small number of such inserted sites will fortuitously be capable of recruiting REST and repressing transcription of a nearby gene. Our findings suggest that selection subsequently favours any increase in affinity for REST and/or proximity to their target genes for these sites. In fact, the majority of deeply conserved RE1s are located within their target gene – often within the transcribed region. In contrast, almost all of the human-specific RE1s lie in gene desert regions. Taken together, these

findings point to a mechanism of random motif generation followed by refinement by which new gene-regulatory motifs are acquired by genes over time.

Clearly, to demonstrate the importance of cis-regulatory evolution it is not sufficient to simply identify non-conserved transcription factor binding sites. Such sites must be shown to affect the phenotype and fitness of species within that lineage. Indeed, it is possible that this is not the case for a large proportion of transcription factor binding sites [63], which we would thus expect to be neutrally evolving. Our analysis showed that a substantial proportion (30%) of RE1 sites in the human genome are primate-specific (ie they are not present in any other non-primate genomes analysed). Importantly, we found that these elements have characteristics of inter- and intra-specific variation consistent with their being under negative evolutionary selection. Finally these lineage-specific RE1s preferentially and significantly target neural genes. The majority of the latter do not have other nearby REST binding motifs, suggesting that these are not cases of turnover but rather represent the acquisition of novel target genes by REST during primate evolution. Our data are not consistent with lineage-specific RE1s representing neutrally-evolving evolutionary noise; rather, they suggest that neurodevelopmental gene regulatory networks in which REST participates have been remodeled to yield advantageous phenotypes during primate evolution.

Given the widespread use of mouse models for human disease, it is significant that only 34% of human RE1 sites are conserved between human and mouse, suggesting that significant differences in the gene regulatory organisation exist between these species. Such differences are not unique to REST: similar studies have shown that just 6% of Nanog-bound promoters are shared in human and mouse ES cells [17], while between 11-59% of promoter-binding by HNF factors in liver is conserved in both species [16]. A future goal will be to determine the biological impact of these differences, particularly with reference to mouse as a model for human disease and an assay platform for human therapeutics. Of particular scientific interest will be the changes in gene expression programmes mediating neuronal development: REST represses numerous neuronal genes in the developing neuroepithelium, and a REST homozygous knockout mouse died prior to birth with major developmental defects of the central nervous system[33]. Given the complexity and diversity of REST's biological roles, the innovations in recruitment profiles we have observed between mammals are likely to have diverse effects. The appearance of an RE1 within a new or existing target gene could have a number of outcomes: general reduction in expression; a change in its spatial or temporal expression pattern; altered response to upstream factors. Therefore, it may have to await investigation on a gene-by-gene basis to understand how RE1 evolution has contributed to organismal fitness. However, given the essential role played by REST in neurodevelopment, the appearance of new RE1 sites capable of regulating cell adhesion and other developmental genes may have contributed to the evolution of primate-specific traits. REST seems to play a profound role in setting the chromatin context of many important neural genes in development [64], and may be required for the appropriate activation of genes during terminal neuronal differentiation [65]. Furthermore, REST is likely to be interconnected with many other transcriptional programmes during this process. So it is likely that the appearance of a novel REST binding site in the vicinity of a developmental transcription factor or signaling molecular or adhesion molecule could significantly alter the behaviour of developing neurons and lead to major changes in the adult brain's organization and capabilities. It was with interest we noted that amongst the ontology terms connected consistently with new and ancient RE1s is that of "Cell Adhesion", since this term was that found previously to be associated with a class of

rapidly-evolving non-coding regions in human [28]. Therefore this study provides further support for the notion that the transcriptional regulation of cell adhesion genes, by REST and other factors, has been extensively reorganized during primate and human evolution.

It is possible that some of the primate-specific RE1 sites identified here are in fact conserved in other mammals, but that their sequence is incorrectly aligned; however, given the relatively large number of non-primate mammal species in this analysis (5), false positives of this type are not likely to be common. In contrast, the difficulty in aligning such divergent genomes as those of chicken, frog and fish, coupled with the relatively small number of these genomes compared to the mammal set, means that we expect the mammal-specific set to be artificially inflated by more ancient RE1s. To compound the potential bias toward classifying pre-mammalian RE1s as mammalian-specific, there is also an inherent inaccuracy of alignments of human to non-mammalian genomes [66]. Because of this we expect that the mammalian-specific set contains considerable numbers of more ancient RE1s which are not correctly aligned to non-mammal genomes. Consequently, we were careful to test hypotheses which should be largely unaffected by such considerations, and our findings are highly statistically significant regardless. In addition to the RE1 sequences identified here, it has recently been shown that REST can also bind to non-canonical RE1 sites in which the 5' and 3' half sites of the motif contain an extra 2-6 nucleotides between them [39]. Such non-canonical sites are not recognized by our PSSM and may contribute to an overestimation of primate and/or mammalian specific RE1s at the expense of more ancient classifications. However any such contribution by non-canonical RE1s will be small as they are found in <5% of bound regions in mouse ES cells [40]. Finally, the two experimental datasets used to validate our analysis were generated by similar but distinct methodologies. It is inevitable that there will be false-negative calls in both datasets, increasing the number of falsely-called lineage-specific binding sites in Figs 4A and 4B. Additionally, the fact that the data came from one pluripotent (mouse embryonic stem cell) and one differentiated (human Jurkat cell) cell type means that it is possible that cell-type specific binding sites could be confused with species-specific sites. Nevertheless, the fact that those PSSM-predicted RE1s which one would expect to be bound in both human and mouse are bound with similar rates in the two species (Fig 4A), strongly argues that the false negative rate of the experimental data, at least for those high affinity, conserved sites, is low.

In summary, the findings presented here suggest that non-coding regulatory DNA, including that regulating neural gene expression, has undergone sustained evolutionary innovation with the ongoing creation and refinement of new transcription factor binding sites. These novel sites have characteristics suggestive of function. The divergent patterns of transcription factor recruitment between species can be partially explained by DNA sequence evolution. These insights are likely to have important implications for our understanding of species diversity and for understanding human biology.

Acknowledgements

The authors wish to thank the following members of the Genome Institute of Singapore: Shyam Prabhakar and Guillaume Bourque for critical reading of the manuscript; Galih Kunarso for providing bioinformatics advice and tools; Ronald Law for help in cloning and purifying REST DBD construct; Anbupalam Thalamuthu, Tannistha Nandi, Vikrant Kumar, Prasanna Kolatkar for advice and discussions. Efithimios Motakis (Bioinformatics Institute, Singapore) advised on statistical analysis. This work was funded by the Singapore Agency for Science, Technology and Research (A*STAR).

References

1. Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869-872.
2. Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, et al. (2004) Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum Mol Genet 13: 489-494.
3. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, et al. (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309: 1717-1720.
4. Chou HH, Takematsu H, Diaz S, Iber J, Nickerson E, et al. (1998) A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. Proc Natl Acad Sci U S A 95: 11751-11756.
5. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.
6. Winter H, Langbein L, Krawczak M, Cooper DN, Jave-Suarez LF, et al. (2001) Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. Hum Genet 108: 37-42.
7. Pheasant M, Mattick JS (2007) Raising the estimate of functional human sequences. Genome Res 17: 1245-1253.
8. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The Evolution of Transcriptional Regulation in Eukaryotes
10.1093/molbev/msg140. Mol Biol Evol 20: 1377-1419.
9. Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165: 349-357.
10. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107-116.
11. Tautz D (2000) Evolution of transcriptional regulation. Current Opinion in Genetics & Development 10: 575-579.
12. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61: 995-1016.
13. Tsong AE, Miller MG, Raisner RM, Johnson AD (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell 115: 389-399.
14. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433: 481-487.
15. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, et al. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428: 717-723.
16. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature Genetics 39: 730-732.
17. Loh Y-H, Wu Q, Chew J-L, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genetics 38: 431-440.

18. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, et al. (2004) Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell 119: 1027-1040.
19. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296: 340-343.
20. Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430: 85-88.
21. Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 16: 962-972.
22. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 193: 723-750.
23. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, et al. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53
10.1073/pnas.0703637104. Proceedings of the National Academy of Sciences 104: 18613-18618.
24. Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, et al. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication
10.1093/nar/gkl525. Nucl Acids Res 34: 3862-3877.
25. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9: 397-405.
26. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell 133: 1106-1117.
27. Stone JR, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764-1770.
28. Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786.
29. Kim SY, Pritchard JK (2007) Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet 3: 1572-1586.
30. Ooi L, Wood IC (2007) Chromatin crosstalk in development and disease: lessons from REST. Nature Reviews Genetics 8: 544-554.
31. Schoenherr C, Anderson D (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. Science 5202: 1360-1363.
32. Chong J, Tapia-Ramirez J, Kim S, Toledo-Aral J, Zheng Y, et al. (1995) REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell 80: 949-957.
33. Chen Z-F, Paquette AJ, Anderson DJ (1998) NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. Nature Genetics 20: 136-142.
34. Sun Y-M, Greenway DJ, Johnson R, Street M, Belyaev ND, et al. (2005) Distinct Profiles of REST Interactions with Its Target Genes at Different Stages of Neuronal Development. Mol Biol Cell 16: 5630-5638.
35. Cheong A, Bingham A, Li J, Kumar B, Sukumar P, et al. (2005) Downregulated REST Transcription Factor Is a Switch Enabling Critical Potassium Channel Expression and Cell Proliferation. Molecular Cell 20: 45-52.
36. Kuwahara K, Saito Y, Takano M, Arai Y, Yasuno S, et al. (2003) NRSF regulates the fetal cardiac gene program and maintains normal cardiac structure and function. EMBO Journal 22: 6310–6321.

37. Ballas N, Grunseich C, Lu DD, Speh JC, Mandel G (2005) REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. Cell 121: 645-657.
38. Singh SK, Kagalwala MN, Parker-Thornburg J, Adams H, Majumder S (2008) REST maintains self-renewal and pluripotency of embryonic stem cells. Nature.
39. Johnson D, Mortazavi A, Myers R, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.
40. Johnson R, Teh CH-l, Kunarso G, Wong KY, Srinivasan G, et al. (2008) REST Regulates Distinct Transcriptional Networks in Embryonic and Neural Stem Cells. PLoS Biology 6: e256.
41. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708-715.
42. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15: 1451-1455.
43. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.
44. Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics: Oxford University Press.
45. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5: R98.
46. Nallamsetty S, Austin BP, Penrose KJ, Waugh DS (2005) Gateway vectors for the production of combinatorially-tagged His6-MBP fusion proteins in the cytoplasm and periplasm of Escherichia coli. Protein Sci 14: 2964-2971.
47. Jauch R, Ng CK, Saikatendu KS, Stevens RC, Kolatkar PR (2008) Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. J Mol Biol 376: 758-770.
48. Mortazavi A, Thompson ECL, Garcia ST, Myers RM, Wold B (2006) Comparative genomics modeling of the NRSF/REST repressor network: From single conserved sites to genome-wide repertoire
10.1101/gr.4997306. Genome Res 16: 1208-1221.
49. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, et al. (2008) The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res 36: D773-779.
50. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130.
51. Mullen RD, Colvin SC, Hunter CS, Savage JJ, Walvoord EC, et al. (2007) Roles of the LHX3 and LHX4 LIM-homeodomain factors in pituitary development. Mol Cell Endocrinol 265-266: 190-195.
52. Bruce AW, Donaldson IJ, Wood IC, Yerbury SA, Sadowski MI, et al. (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes
10.1073/pnas.0401827101. PNAS 101: 10458-10463.
53. Wood IC, Belyaev ND, Bruce AW, Jones C, Mistry M, et al. (2003) Interaction of the repressor element 1-silencing transcription factor (REST) with target genes. J Mol Biol 334: 863-874.
54. Belyaev ND, Wood IC, Bruce AW, Street M, Trinh J-B, et al. (2004) Distinct RE-1 Silencing Transcription Factor-containing Complexes Interact with Different Target Genes
10.1074/jbc.M310353200. J Biol Chem 279: 556-561.

55. Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38: 1452-1456.
56. Wu J, Xie X (2006) Comparative sequence analysis reveals an intricate network among REST, CREB and miRNA in mediating neuronal gene expression. Genome Biol 7: R85.
57. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207-219.
58. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, et al. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res 19: 24-32.
59. Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19: 68-72.
60. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, et al. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4: e1000216.
61. Ooi L, Belyaev ND, Miyake K, Wood IC, Buckley NJ (2006) BRG1 Chromatin Remodeling Activity Is Required for Efficient Chromatin Binding by Repressor Element 1-silencing Transcription Factor (REST) and Facilitates REST-mediated Repression. J Biol Chem 281: 38974-38980.
62. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772-778.
63. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6: e27.
64. Greenway DJ, Street M, Jeffries A, Buckley NJ (2007) RE1 Silencing Transcription Factor Maintains a Repressive Chromatin Environment in Embryonic Hippocampal Neural Stem Cells
10.1634/stemcells.2006-0207. Stem Cells 25: 354-363.
65. Kuwabara T, Hsieh J, Nakashima K, Warashina M, Taira K, et al. (2005) The NRSE smRNA specifies the fate of adult hippocampal neural stem cells. Nucleic Acids Symp Ser (Oxf): 87-88.
66. Kumar S, Filipski A (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res 17: 127-135.

Figure Legends

Figure 1: Sequence-based phylogenetic classification of RE1s.

(A) Strategy for identifying lineage-specific RE1s. The RE1 PSSM was used to score each 21bp sequence within the human genome. Subsequently, Multispecies aligned regions for each human RE1 were extracted, and themselves searched for RE1s in the same way. Shown is a hypothetical region of the human genome containing 3 RE1 sites (blue ovals). The RE1 site on the left is also present in aligned regions of chimp, mouse and fugu genomes, and hence is designated as a fish-specific (deeply conserved) RE1. The central RE1 site is a primate-specific RE1 because it is present in aligned sequence from chimp, but not mouse, while the fugu genome does not have an aligned sequence. The right-hand RE1 sequence represents a human lineage RE1 because it is not present in any other genome (in this case, because the region of the human genome does not align with any other).

(B) A heatmap representation of RE1 phylogenetic conservation. Each row of the heatmap represents a human RE1. Each column represents the RE1 conservation status of the orthologous region in a given species. Green indicates that an RE1 was identified in that sequence block. Black indicates that an aligned sequence block exists in that species, but does not contain an RE1. Red indicates that no orthologous sequence was found in that species. Data is separated into lineage classes which are shown on the left.

(C) A breakdown of the 1298 human RE1s by lineage category.

(D) Conservation of human RE1s in mouse. Human RE1s were specifically compared to orthologous loci in mouse and classified as: orthologous mouse loci that contains an RE1 (blue); orthologous mouse loci that do not contain an RE1 (grey) or human RE1s which have no aligned region in mouse (white).

Figure 2: Examples of human gene regulation by ancient and recent RE1s.

(A) The human *LHX3* gene (antisense strand), encoding a developmental homeodomain transcription factor, has a deeply-conserved RE1 in the proximal upstream region (red box). The RE1 PSSM scores for orthologous regions of 16 vertebrate species is shown to the right; species having a score below the cutoff of 0.91 are highlighted in grey, while species with no aligned genomic region are given a nominal score of zero. A local multiple alignment of multiple species orthologous DNA is shown below. The RE1 element is boxed and resides on the antisense strand.

(B) The *AGBL1* gene, encoding a protease of unknown function, contains two RE1s. The first, on the sense strand in the alignment above the figure, is primate specific with predicted orthologues in chimp and macaque. The second, expanded below the figure, is predicted to be specific to humans (green box). This RE1 resides within a primate-specific LINE2 insertion (highlighted in blue): note the absence of aligned sequence in non-primate species corresponding to the region of LINE1 (bottom track).

Figure 3: Experimental validation of lineage-specific RE1s by quantitative EMSA.

(A) Sequence of lineage-specific RE1: the EMSA probe (probe), the positive control *CHRM4* and negative control *CHRM4* Mut, and species-specific competitor DNA sequences for *LHX3*, *LARP6*, *SFRS8*, *TSNARE1* and *AGBL1*. Positions highlighted in black are well-conserved, in grey are moderately conserved and white weakly conserved. Nucleotides likely contributing to observed affinity differences are indicated in red. Hs – human, Pt – chimp, Rm – macaque, Rn – rat, Mm – mouse, Cf – dog, Et – hedgehog, Gg – chicken, Tn – tetraodon, Fr – fugu.

(B) Affinity of lineage-specific RE1s *in vitro*. EMSA of the RE1 probe with *LHX3*, *LARP6*, *SRFS8*, *TSNARE1* and *AGBL1* unlabelled competitors. One representative gel from four independent experiments is shown. For each gene set, the RE1 probe was electrophoresed alone (Probe), or with REST DBD protein alone (Protein), or with protein in the presence of unlabelled DNA competitor sequences: wild-type and mutant RE1s from the *CHRM4* gene (Control+ and Control- respectively) or the orthologous sequence from the indicated species. Specific DNA-protein complexes (REST DBD:DNA) (*b*, binding protein; *d*, binding of partially degraded protein) and free probe (DNA) (*f*) are indicated.

(C) Quantitation of relative binding affinities for lineage specific RE1s. The fraction of EMSA probe bound to REST in the presence of each competitor is shown; thus, low values represent a high affinity DNA sequence (mean ± standard deviation, n=4). Binding of each RE1 was compared to the relevant human sequence; statistical significance is represented by an asterisk ($P<0.05$, Student's *t*).

Figure 4: Validation of RE1 conservation using genome-wide ChIP data.

(A) Human RE1s were divided into categories based upon their conservation to mouse, as in Figure 1D. The percentages of these sites overlapping an experimental ChIP-Seq site in human [39] is shown by the black bars. We found the orthologous region in the mouse genome for all these sites, using the UCSC LiftOver tool. The percentages of mouse orthologous regions overlapping ChIP PET data from mouse [40] were similarly calculated and shown by grey bars. Statistical significance was calculated using the Chi-square test. The human RE1s fall into the following categories: those which have an orthologous mouse genomic region that also contains an RE1 ("RE1"); those which have an orthologous region that does not contain an RE1 ("No RE1"); those which have no orthologous region in mouse ("No alignment").

(B) RE1 overlap to experimental datasets was calculated as in (A), for RE1s with respect to lineage.

(C) The numbers of ChIPSeq sequencing tags for each RE1 is shown as a boxplot. The central bar denotes the median value, the box the interquartile range, and whiskers extend to 1.5 times the interquartile range beyond the box. Values lying outside the whiskers are defined as outliers. Median values are shown underneath the graph.

Figure 5: Ancient RE1 are more similar to the canonical RE1 sequence and more proximal to target genes.

The position-specific scoring matrix (PSSM) scores (A) and distance to the transcriptional start site (TSS) of the nearest Refseq gene (B) of lineage-specific RE1 sequences are plotted as box plots. Median values are highlighted in grey below.

Figure 6: Divergence of genomic REST binding profiles between human and mouse.

(A) The orthologous locations of experimentally-determined REST binding sites in human (ChIP-Seq in Jurkat cells [39]) and mouse (ChIP-PET in embryonic stem cells [40]) were compared.
(B) The PSSM score of each RE1 site bound in human and mouse ('Human AND Mouse'), as well as for 1242 sites which are specifically bound in human ('Human NOT Mouse') or 1770 sites specifically bound in mouse ('Mouse NOT Human'). For each human RE1, the orthologous region in mouse was determined by LiftOver and vice versa, and the highest scoring PSSM hit in each case was deemed to be the RE1. The 'Most Human' RE1 set were bound in human alone and have a PSSM score ≥0.1 compared that of the orthologous mouse region, and vice versa for 'Most Mouse'.
(C) PSSM score for the RE1s in each of the three categories present in human (H) and mouse (M) genomes plotted as box plots. The central bar denotes the median value, the box the interquartile range, the whiskers the range, and circles outliers.
(D) The distance from the centre of each experimentally-determined REST binding site to the nearest Refseq transcriptional start site for three categories of RE1 site in both human and mouse genomes. Data are presented in box pots as described in (C).
(E) Association of the REST binding site sets with LINE2 (grey) or other repeats (black) in the human genome. Background rates of association, were taken using a region 1kb downstream of each binding site.

Figure 7: Reduced inter-specific variation in RE1 nucleotides.

(A) Nucleotide substitution rates were estimated from the simple divergence rate of RE1s since divergence of human from macaque. Substitution rates were estimated from the rate of divergence of nucleotides in human-macaque sequence alignments. This rate was measured for the core half sites of the RE1 motif (black bars), and the background rate was estimated from a 200bp window around the RE1 motif. The statistical significance of the difference between these two values was assessed using the Chi-square test, comparing conserved and divergent nucleotides between RE1 and flanking DNA. This analysis was repeated for the set of Primate-specific RE1s, Mammal-specific RE1s and a set of control regions 1kb upstream of every Primate-specific RE1.
(B) As in (A) for human-mouse alignments.
(C) The program MONKEY was used to test for negative selection of RE1 motifs. With the RE1 PSSM, MONKEY was used to assign statistical significance to every human RE1 which could be aligned to macaque. For each case the reported P-value was recorded. The median P-value for the sets of Primate-specific and Mammal-specific RE1s are shown as grey lines. To estimate background, the RE1 PSSM was randomly shuffled 100 times and used to scan whole-genome human-macaque sequence alignments. The black

bars represent a histogram of resultant median reported P-values across all shuffled PSSMs.

(D) As in (C) for human-mouse alignments.

Figure 8: Reduced human variation in primate-specific RE1s.

(A) The total SNP count at each nucleotide position of the RE1 is summed for all human RE1s. Grey regions indicate nucleotides important for binding by REST. The number of SNPs within this region is significantly lower than for 200bp window around the RE1 (P=3.2E-4, χ^2 test).

(B) SNP density is significantly reduced for both primate-specific and mammal-specific RE1s, compared to 200bp of flanking DNA (P=5.0E-4 and P=2.4E-12, respectively by χ^2 test). Density was calculated by summing the number of single-nucleotide dbSNP129 polymorphisms at each position in the RE1 across all instances, then dividing by the number of instances. A set of control regions were constructed by taking equivalent RE1 and flanking sequences 1kb upstream of every primate-specific RE1.

Figure 9: A model for transcriptional network evolution.

In the scheme, the *x*-axis represents genomic distance from a hypothetical target gene, whose transcriptional start site is denoted by the black arrow). The *y*-axis represents the *in vivo* binding affinity of an RE1 (upper panel), or its repressive capability on target gene transcription (lower panel). New RE1s are constantly generated throughout the genome, often in gene-distal regions (grey arrows). Most sites have no phenotypic effect, or are detrimental to fitness and hence are lost (red crosses). A minority of sites may be weakly beneficial – for example the RE1 shown. Under selective pressure, the affinity of the RE1 sequence improves (via sequence mutation) and becomes more proximal to its target (dashed arrow), thereby improving its repressive capability. The most ancient RE1s take up a position proximal to the target gene TSS, and acquire improved sequence characteristics for *in vivo* recruitment, leading to maximal regulatory function.

Table 1: Primate-specific RE1s and their target genes.

RE1 ID	Chr.	Position	Distance (bp)	Target Refseq	Symbol	Description	ChIP-Seq
>RE1_10_71663060_-_0.9112	10	71663060	-126	NM_021129	PPA1	Inorganic pyrophosphatase	-
>RE1_16_2450245_-_0.9322	16	2450245	-139	NM_025108	C16orf59	Hypothetical protein LOC80178	-
>RE1_20_61798244_-_0.9168	20	61798244	211	NM_032945	TNFRSF6B	Tumor necrosis factor receptor superfamily	-
>RE1_22_30980262_-_0.9236	22	30980262	-1054	NM_014227	SLC5A4	Low affinity sodium-glucose cotransporter	-
>RE1_16_47871900_-_0.9587	16	47871900	-1283	NM_004352	CBLN1	Cerebellin 1 precursor	-
>RE1_2_218990717_-_0.9152	2	218990717	1362	NM_007127	VIL1	Vilin 1	-
>RE1_10_124661567_+_0.9383	10	124661567	-1370	NM_001029888	FAM24A	Family with sequence similarity 24, member A	BOUND
>RE1_22_43038512_-_0.9197	22	43038512	-1542	NM_001099294	KIAA1644		-
>RE1_16_164435_-_0.9504	16	164435	-1570	NM_000517	HBA2	Alpha2 globin	-
>RE1_19_50005757_-_0.9216	19	50005757	-1589	NM_001013257	BCAM	Basal cell adhesion molecule isoform 2	-
>RE1_22_22364989_+_0.9489	22	22364989	-1951	NM_153615	Rgr	Ral-GDS related protein	BOUND
>RE1_16_2818196_+_0.9501	16	2818196	1968	NM_145252	LOC124220	Hypothetical protein	BOUND
>RE1_7_131986096_+_0.9156	7	131986096	-2013	NM_173682	LOC286023	IQ motif and Sec7 domain-containing protein 2.	-
>RE1_X_53329672_-_0.9302	X	53329672	2161	NM_015075	IQSEC2	Hypothetical protein	-
>RE1_16_5085286_-_0.9217	16	5085286	-2446	NM_201400	FAM86A	Hypothetical protein LOC196483	-
>RE1_15_68931042_-_0.9216	15	68931042	-2500	NM_197958	LARP6	Acheron isoform 2	-
>RE1_22_48695402_+_0.9445	22	48695402	-2698	NM_024105	ALG12	Asparagine-linked glycosylation 12	BOUND
>RE1_19_44781793_-_0.9713	19	44781793	3201	NM_013268	LGALS13	Galactoside-binding soluble lectin 13	BOUND
>RE1_16_2851366_-_0.9405	16	2851366	3204	NM_022119	PRSS22	Brain-specific serine protease 4 precursor	BOUND
>RE1_14_69104215_+_0.9162	14	69104215	-3277	NM_020181	C14orf162	Chromosome 14 open reading frame 162	-

Shown are the 20 most proximal primate-specific RE1 / Refseq gene pairs. Chr.: Chromosome. Position refers to the first nucleotide of the RE1 motif. The location of the RE1 relative to the transcriptional start site (TSS) of the nearest Refseq gene is displayed; a negative value indicates a location downstream of the TSS. Those RE1s which were shown to recruit REST by genome-wide ChIP-Seq [39] are labeled as "Bound".

Figure 1

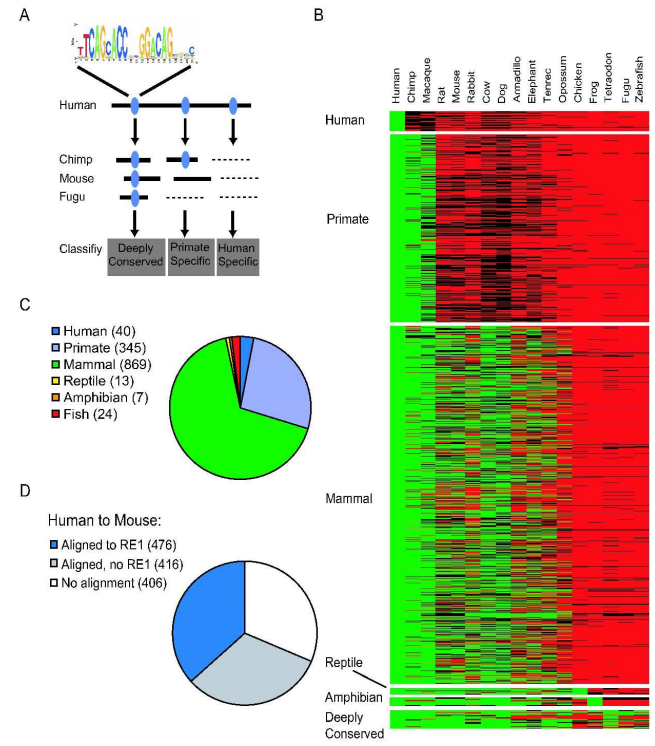
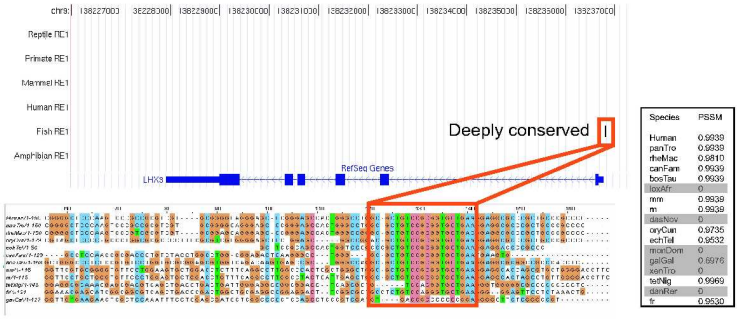


Figure 1
210x297mm (600 x 600 DPI)

Figure 2

A *LHX3*



B *AGBL1*

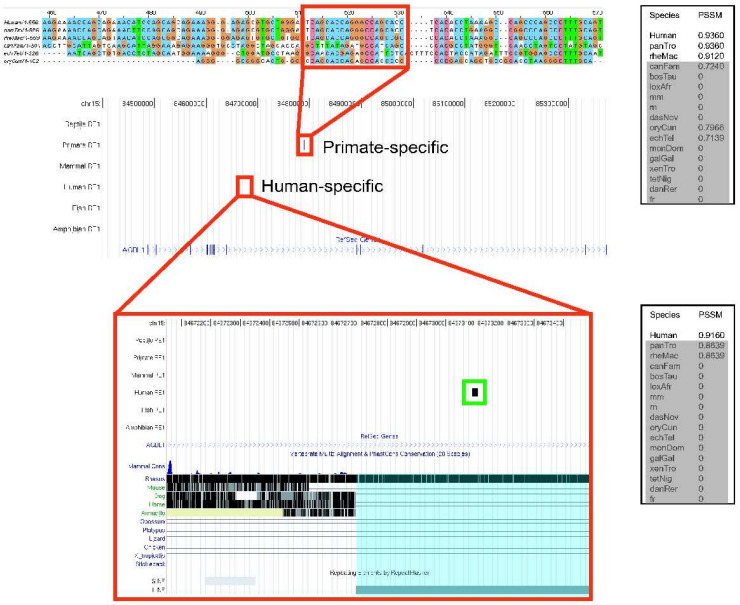


Figure 2
180x239mm (600 x 600 DPI)

Figure 3

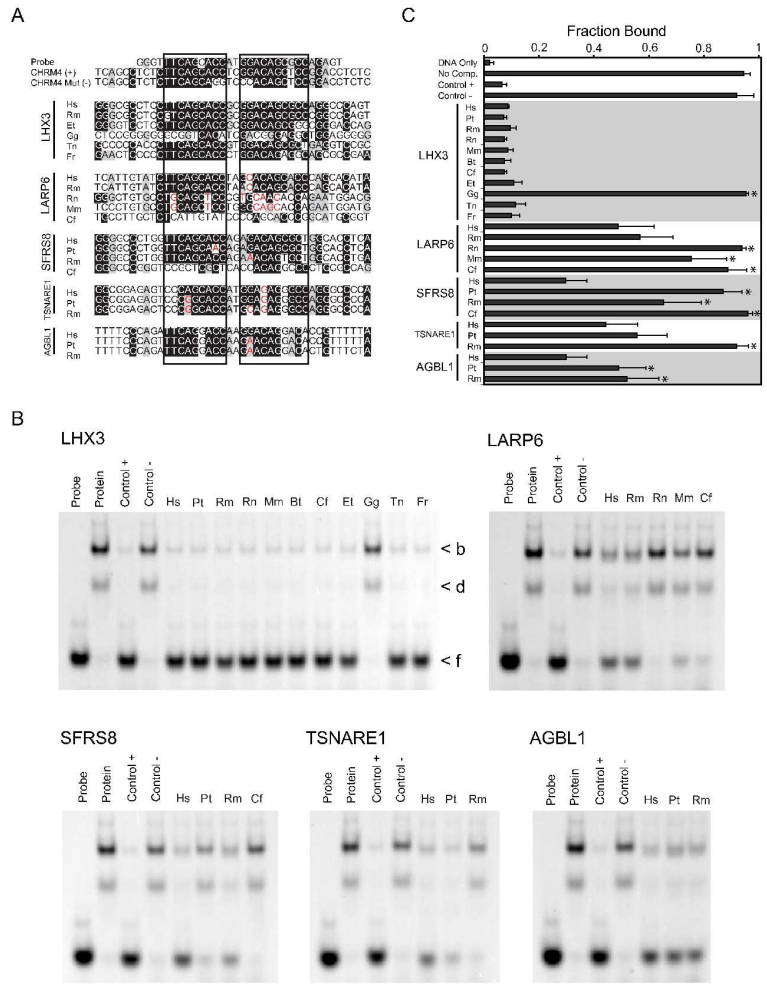


Figure 3
182x258mm (600 x 600 DPI)

Figure 4

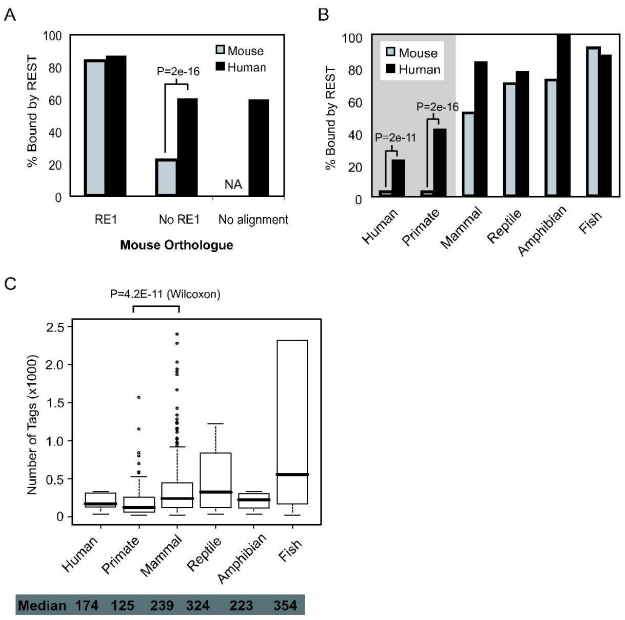


Figure 4
210x297mm (600 x 600 DPI)

Figure 5

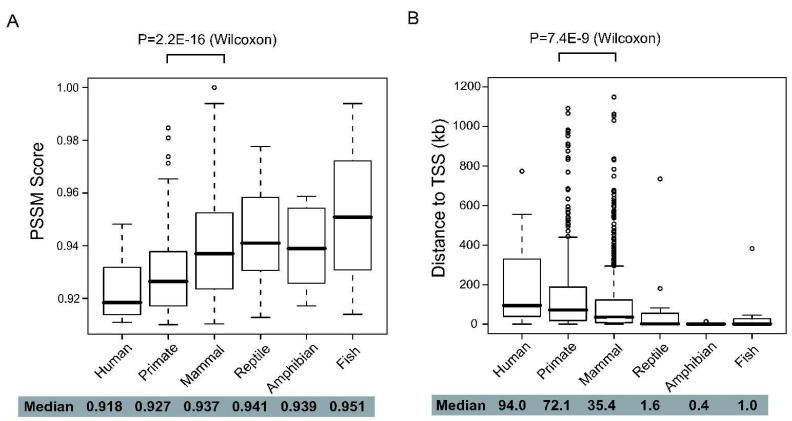


Figure 5
210x297mm (600 x 600 DPI)

Figure 6

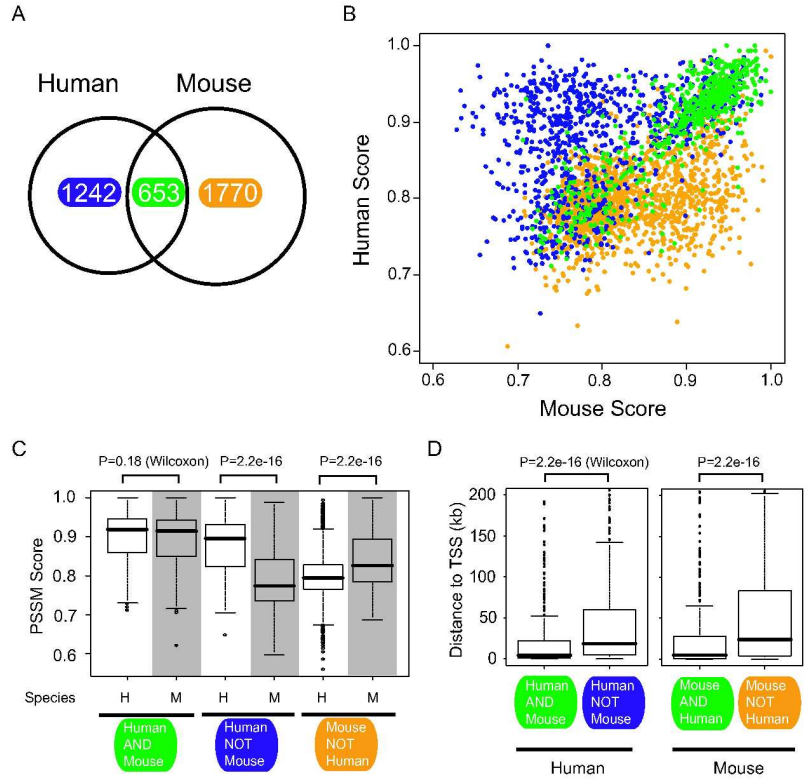


Figure 6
180x239mm (600 x 600 DPI)

Figure 7

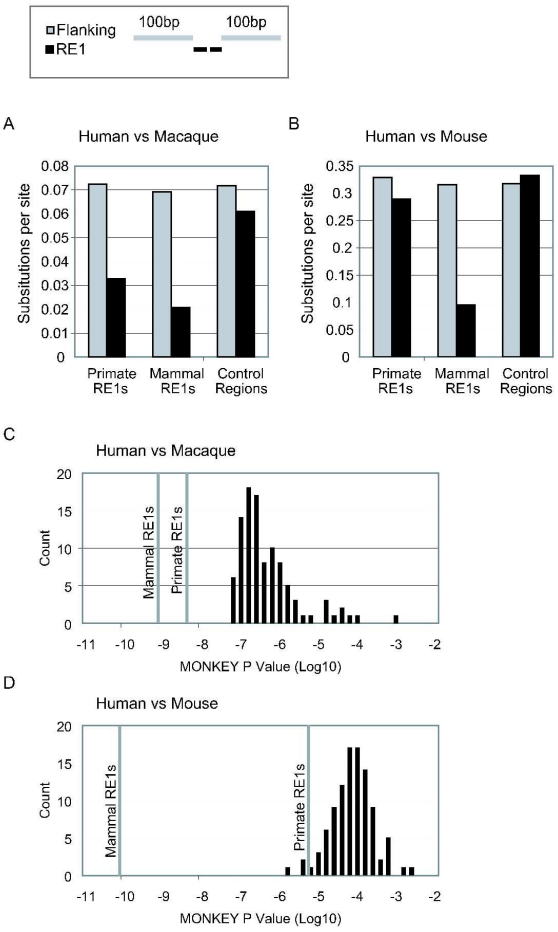


Figure 7
210x297mm (600 x 600 DPI)

Figure 8

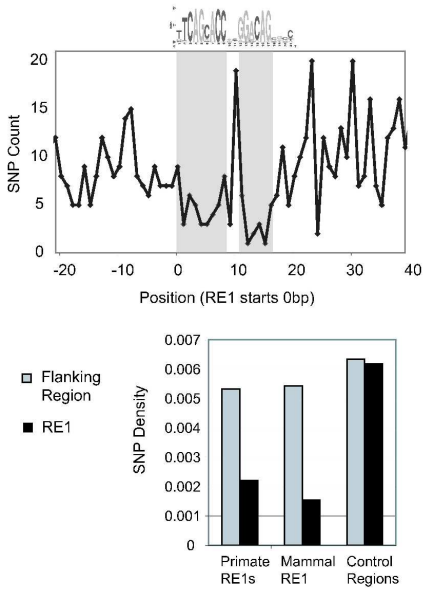


Figure 8
210x297mm (600 x 600 DPI)

Figure 9

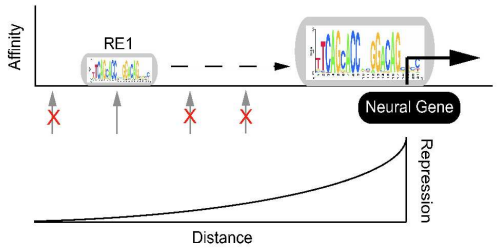


Figure 9
210x297mm (600 x 600 DPI)