Accelerated Evolution of Enhancer Hotspots in the Mammal Ancestor

Alisha K. Holloway1,2, Benoit G. Bruneau1,3, Tatyana Sukkonik1, John L. Rubenstein4, and Katherine S. Pollard1,2,5
1Gladstone Institute of Cardiovascular Disease, San Francisco, CA
2Department of Epidemiology and Biostatistics, University of California, San Francisco, CA
3Department of Pediatrics and Cardiovascular Research Institute, University of California, San Francisco, CA
4Nina Ireland Laboratory of Developmental Neurobiology and Department of Psychiatry, University of California, San Francisco
5Institute for Human Genetics, University of California, San Francisco, CA

*Corresponding author: E-mail: katie.pollard@gladstone.ucsf.edu.
Associate editor: Gregory Wray

Abstract

Mammals have evolved remarkably different sensory, reproductive, metabolic, and skeletal systems. To explore the genetic basis for these differences, we developed a comparative genomics approach to scan whole-genome multiple sequence alignments to identify regions that evolved rapidly in an ancestral lineage but are conserved within extant species. This pattern suggests that ancestral changes in function were maintained in descendants. After applying this test to therian mammals, we identified 4,797 accelerated regions, many of which are noncoding and located near developmental transcription factors. We then used mouse transgenic reporter assays to test if noncoding accelerated regions are enhancers and to determine how therian-specific substitutions affect their activity in vivo. We discovered enhancers with expression specific to the therian version in brain regions involved in the hormonal control of milk ejection, uterine contractions, blood pressure, temperature, and visual processing. This work underscores the idea that changes in developmental gene expression are important for mammalian evolution, and it pinpoints candidate genes for unique aspects of mammalian biology.

Key words: comparative genomics, development, enhancer evolution, mammal-specific traits

Introduction

Understanding the molecular underpinnings that produce physiological, cognitive, and morphological differences between organisms is a major challenge for biologists. Using forward genetics approaches, scientists have begun to establish direct connections between lineage-specific traits and molecular evolution. For example, to achieve complete ventricular septation, which allows for separation of oxygenated and deoxygenated blood, developing mammalian hearts express the transcription factor Tbx5 in a pattern distinct from their ancestors (Koshiba-Takeuchi et al. 2009). Furthermore, in stickleback fish, loss of a single tissue-specific enhancer of the Pitx1 gene causes loss of pelvic spines (Chan et al. 2010). In flies, nucleotide changes in multiple enhancers for shavenbaby lead to species-specific trichome patterns (McGregor et al. 2007; Frankel et al. 2011). However, an approach that surveys large portions of the genome is required to gain a more comprehensive view of how molecular evolution relates to lineage-specific traits. One such approach is comparative genomics, which allows for scanning of entire genomes for DNA sequences that have a key evolutionary signature: Differences in nucleotides suggesting that these sequences are responsible for making one species different from another. A scanning technique to identify conserved genome sequences with many changes in one extant lineage was first developed to discover human accelerated regions (HARs) (Pollard et al. 2006; Prabhakar et al. 2006; Koshia-Takeuchi et al. 2009) and has been used to identify genomic regions with elevated substitution rates in many other lineages (Chan et al. 2010; Lindblad-Toh et al. 2011). Others have identified lineage-specific conserved noncoding elements within a group that are either deposited by mobile elements (Bejerano et al. 2006; Lowe et al. 2007; Sasaki et al. 2008) or are otherwise absent in other groups (Mikkelsen et al. 2007). We extend this framework for testing conservation and acceleration on terminal branches to enable identification of changes in the rate of molecular evolution on an ancestral lineage. Specifically, we identify both coding and noncoding sequences over the entire genome that are present in vertebrates, evolved rapidly in the ancestral lineage of therian mammals but are then conserved in its descendants. Such a pattern suggests that the ancestors experienced a change in function, and the change in sequence was then constrained in descendants to maintain this new function.

We tested the branch leading to therian mammals to understand the genetic basis for differences in morphology and physiology from other vertebrates (fig. 1A). Therian mammals consist of marsupials (e.g., wallaby, opossum) and eutherians (e.g., human, dog) to the exclusion of the third group of extant mammals—the monotremes (e.g., platypus, echidna).
Therians diverged from monotremes ~190 Ma (Luo et al. 2011). Therians are diverse and widespread, and they share several unique traits that distinguish them from other vertebrates, such as having live young, a regular heartbeat, a higher metabolic rate, erect limb posture, and true nipples. Conversely, monotremes retained many traits present in the earliest mammals, some of which are shared with extant reptiles, including the cloaca (a single opening for urinary, defecatory, and reproductive functions), a sprawling gait, an uncoiled cochlea in the inner ear, lack of or reduced sweat glands, and egg laying (Rich 2005; Widelitz et al. 2007; Wakefield et al. 2008; Warren et al. 2008; Bickelmann et al. 2012; Ashwell 2013). Thus, therians are distinct from monotremes and other vertebrates in their morphology and physiology. We set out to discover the genetic basis for therian-specific traits.

Results and Discussion

Comparative Genomics Identifies Hundreds of Sequences That Distinguish Therian Mammals from Other Vertebrates

We used phyloP combined with the phastCons program in PHAST (Pollard et al. 2010; Hubisz et al. 2011) to scan whole-genome alignments in vertebrates for sequences that are present in therian and nontherian vertebrates but changed significantly in the therian mammal ancestor and remained highly conserved during therian diversification. We identified 177,346 vertebrate genomic regions that are conserved highly during therian diversification. We set out to discover the genetic basis for therian-specific traits.

The distribution of TSARs with respect to human gene features is similar to that of TCRs (fig. 1B). TSARs are largely within genic regions (supplementary table S2, Supplementary Material online), in contrast to HARs and the mammalian-conserved elements from which HARs were identified, both of which occur most frequently in intergenic DNA (Pollard et al. 2006; Lindblad-Toh et al. 2011). This distribution difference may be because intergenic sequences typically evolve more rapidly and are therefore more difficult to align in the TSAR analysis, which adds a requirement of syntenic alignment to nonmammalian vertebrates (see Materials and Methods).

To explore the hypothesis that TSARs are uncharacterized regulatory elements, we used functional genomics data from the ENCODE project (Dunham et al. 2012) and other sources (Griffith et al. 2008; Visel et al. 2009; Blow et al. 2010; Rada-Iglesias et al. 2011; Shen et al. 2012) to investigate how many TSARs overlap features indicative of enhancer or promoter activity in various human and mouse cell types (supplementary table S1, Supplementary Material online). We found widespread evidence that TSARs are active regulatory sequences. First, from human ENCODE data, 89% of TSARs overlap with DNaseI hypersensitive sites, H3K27 acetylation marks, or P300 peaks, all of which indicate regulatory activity. Furthermore, the chromHMM analysis of human ENCODE data (Ernst and Kellis 2012) predicts that 14% of TSARs are promoters and 35% are enhancers, and only 12.7% of TSARs are in intergenic regions without any predicted regulatory function. Furthermore, 98% of TSARs that overlap chromHMM-predicted enhancer sequences are supported by at least one additional experiment showing evidence of enhancer activity, and 89% are supported by multiple additional experiments. These include 366 TSARs that overlap elements from experiments designed to identify cis-regulatory regions in murine tissues (Shen et al. 2012), and 129 in putative regulatory regions in the Open Regulatory Annotation database (Griffith et al. 2008).
A small number of TSARs have been characterized experimentally in previous studies. TSARs overlap 17 early developmental enhancers in humans (Rada-Iglesias et al. 2011) and 105 putative heart enhancers in mouse (Blow et al. 2010). Transgenic enhancer assays from the VISTA Enhancer Browser (Visel et al. 2007) show that 51 TSARs have enhancer activity in e11.5 mouse. These enhancers are primarily active in the brain, eye, nose, and limbs (Visel et al. 2007). Overall, 4,285 of the 4,797 TSARs display evidence of regulatory function, yet they are largely uncharacterized. Future studies of these loci may offer clues about therian-specific biology.

Mutations in TSARs Alter Protein Sequences
More than half of the TSAR sequences overlap exonic regions of the human genome. Therefore, we explored mechanisms of evolution and the putative functional consequences of substitutions in the protein-coding regions of TSARs. Rapid evolution in protein-coding regions could be due to selection on protein function, rates of protein synthesis, or DNA- or RNA-binding elements. Using only the coding portions, we estimated rates of nonsynonymous (dN; amino acid altering) and synonymous (dS; non–amino acid altering) substitutions using Phylogenetic Analysis by Maximum Likelihood (Yang 2007) on the mammal ancestor and on the therian ancestor branches. Our analysis is conservative in that we required sequence to be present in both marsupials, the monotreme and at least one eutherian and one nonmammalian vertebrate. Median estimates of these substitution rates indicate that protein-coding TSARs evolved rapidly at synonymous sites and only slightly higher at nonsynonymous sites in the therian ancestor compared with the mammal ancestor (supplementary table S3, Supplementary Material online). However, some TSAR-containing genes have high rates of nonsynonymous substitution compared with synonymous substitution on the therian ancestor branch, which may indicate adaptive protein evolution (fig. 2A). These genes are involved in diverse yet vital functions that include spermatogenesis and neural development and many are either secreted or transmembrane proteins. Mutations in these genes are associated with cardiovascular disease, anhidrosis, skeletal abnormalities, hearing loss, and neurological diseases. Nonsynonymous substitutions in these genes had the potential to alter therian-specific traits.

TSARs Cluster in Loci Associated with Disease and Development
In the genome, TSARs are significantly closer to one another than are TCRs (supplementary fig. S2, Supplemental Material online). We defined a “cluster” of TSARs as any set of three or more TSARs where the neighboring TSARs are ≤50 kb apart. We then associated these clusters with nearby genes. We speculated that TSAR clusters might pinpoint genes whose functions were hotspots of modification in the therian ancestor, either via regulatory divergence or through structural changes to the encoded protein. Genes harboring clusters of TSARs were more often associated with urogenital defects and to a lesser extent brain development than genes near or containing a single TSAR. The largest cluster (n = 36) centres on the developmental transcription factor, Pax2, which is active in the development of the mid-hindbrain boundary, the retina, the renal system, and the inner ear.

Previous studies have suggested that distinct sets of genes undergo either protein-coding or regulatory evolution (Wray 2007). Consistent with this idea, we found that many TSAR clusters were either predominantly coding (>75% coding sequence) or noncoding (<25% coding sequence), with less than 30% of clusters having equal proportions of coding and noncoding sequence (fig. 2B and supplementary table S4, Supplementary Material online). Interestingly, we found further distinguishing characteristics within coding clusters—specifically that many protein-coding TSAR clusters (e.g., those in genes that encode UBE4A, DNAH2, HEATR1, Supt16H, and WDFY3) are evolving rapidly at synonymous sites, which might indicate changes in translational efficiency and expression (Warnecke and Hurst 2007; Mitarai et al. 2008) or regulatory elements overlapping exons (Birnbaum et al. 2012). Very few protein-coding TSAR clusters have high rates of nonsynonymous substitution (fig. 2A) (supplementary fig. S3A, Supplementary Material online). Taken together, the protein-coding TSAR clusters were not enriched for any particular biological processes. In comparison, with as clusters of TSARs in general, genes near or containing hotspots of noncoding TSARs are enriched for involvement in development of many brain regions and regulation of transcription. Because many TSARs that are clustered in these loci likely function as developmental enhancers or other modulators of expression, our results suggest that regulatory changes in the therian ancestor may have influenced neurodevelopment.

Multiple TSARs in the Lhx1 Locus Function as Developmental Enhancers
One large cluster of TSARs (n = 20) resides primarily in the intergenic region between Lhx1, a LIM-homeodomain transcription factor, and Mrm1, a mitochondrial rRNA methyl transferase (supplementary fig. S3B, Supplementary Material online). These TSARs are largely uncharacterized—only three overlap validated enhancers: TSAR.0067:OREG0042899 (Wederell et al. 2008), human ESC expression (Rada-Iglesias et al. 2011); TSAR.1565, expression in embryonic day 11.5 (e11.5) mouse limb (Visel et al. 2007); and TSAR.1586, expression in e14.5 brain (Shen et al. 2012). Using ENCODE data, chromHMM (Ernst and Kellis 2012) predicted that eight others are also enhancers. Although both Lhx1 and Mrm1 are widely expressed, the LIM-homeodomain transcription factors are important for mammal-specific forebrain development and many of them are expressed together in specific subregions (Abellan et al. 2010). To investigate if uncharacterized noncoding TSARs are novel enhancers of Lhx1, we used a transient transgenic reporter assay (Noonan 2009). In this assay, we determined if the mouse sequences of TSAR.0067 and TSAR.1586 show enhancer activity in e11.5 mouse embryos. We found that both TSARs drove reporter gene expression. TSAR.1586 produced a reproducible pattern of expression in the majority of embryos (fig. 3 and
supplementary fig. S4 and table S5, Supplementary Material online), but TSAR.0067 was expressed in many embryos, yet showed a variable expression pattern between embryos (supplementary fig. S4 and table S5, Supplementary Material online). This may be the result of testing at e11.5, because this enhancer is known to be active at earlier stages (e.g., human embryonic stem cells; Rada-Iglesias et al. 2011). TSAR.1586 was active in a small domain spanning the hypothalamus and telencephalon (fig. 3 and supplementary fig. S4, Supplementary Material online). Hypothalamic expression was probably in the anlage of the paraventricular nucleus and with expression continuing into the telencephalic stalk and extending into the anlage of the amygdala, resembling the pattern of Otp expression (Bardet et al. 2008; Garcia-Moreno et al. 2010). We repeated the transient transgenic experiment after replacing the mouse TSAR.1586 sequence with the orthologous chicken sequence (nontherian). This allowed us to test if changing this sequence in the therian ancestor altered the function of the Lhx1 enhancer. The construct with the chicken TSAR.1586 sequence did not show reporter gene expression in the hypothalamus and telencephalon. Interestingly, therian-specific hypothalamic expression occurred in the paraventricular neurons, which project into the pituitary gland where they release oxytocin or vasopressin (Okuno et al. 1965), all of which are critical features of therian-specific evolution. Further experimentation would be required to substantiate connection between this enhancer and regulation of hormones.

Evolution of Gata2 Enhancers May Be Important for Processing Visual Information

Gata2, a well-characterized developmental transcription factor, functions during hematopoietic, central nervous system (CNS), and urogenital development in the mouse and chicken, all of which are highlighted in the ontology overrepresentation analysis. A group of eight TSARs is located within the known regulatory boundaries of the developmental transcription factor Gata2 (seven shown in supplementary fig. S3, Supplementary Material online) (Zhou et al. 1998; Brandt et al. 2008). Of the eight, six are predicted to be enhancers (chromHMM) or are validated Gata2 enhancers (TSAR.0153:OREG0002950 and TSAR.3936:OREG0002949) (Wang et al. 2006). TSAR.3936 also has enhancer activity in e11.5 heart (Visel et al. 2009). Using our transient transgenic reporter assay (Noonan 2009), we tested whether three uncharacterized noncoding TSARs—TSAR.1137, TSAR.1622, and TSAR.2014—are novel enhancers of Gata2. At e11.5, all three sequences drove expression in one or more tissues (fig. 4 and supplementary fig. S5 and table S5, Supplementary Material online). As a positive control, we also tested TSAR.0153 and confirmed activity in embryonic vasculature (Wang et al. 2006).

The mouse version of TSAR.1622 showed strong enhancer activity in the CNS that recapitulated endogenous Gata2 expression (fig. 4 and supplementary fig. S5 and table S5, Supplementary Material online). Specifically, this enhancer was active at the midbrain/hindbrain patterning center (isthmus) and extended rostrally in the mantle zone along a longitudinal band in an intermediate dorsal/ventral position of the midbrain and caudal forebrain, terminating in the region of the pretectum (the latter regions also had extensive alar plate [dorsal] expression). In addition, activity in the hindbrain and spinal cord was robust along a ventral longitudinal domain that may include regions that produce motor neurons. Upon testing expression of the orthologous chicken TSAR.1622 sequence, we observed reporter gene expression in the spinal cord, similar to observations with the mouse version. Strikingly, the orthologous chicken sequence lacked reporter gene expression in the pretectum of the midbrain (fig. 4), which is involved in processing visual information and
mediating subconscious responses to light, the perception of pain, and REM (rapid eye movement) sleep. Thus, our discovery of a therian-specific enhancer of *Gata2* in the pretectum is particularly interesting because, early in their evolution, mammals exploited nocturnal niches, and this TSAR may be associated with vision (Heesy and Hall 2010).

*Gata2* also has an important role in regulating expression in the superior colliculus (SC), which mediates responses to visual inputs and is the primary integrating center for eye movements (Willett and Greene 2011). Two known enhancers of *Gata2* drive expression in the SC, OREG0002948, which is highly conserved in amniotes, and OREG0002951, which has no similar sequence in nonmammalian vertebrates. In the SC of mice, OREG0002948 alone can regulate *Gata2* (Nozawa et al. 2009). The orthologous OREG0002948 sequences in the human and chicken have only 10% sequence divergence, which may indicate that the chicken sequence is also an enhancer of *Gata2* in the SC. The pretectum and SC are both critical for processing visual information, so the evolution of enhancers that modulate *Gata2* expression may indicate that this gene is important in therian-specific visual evolution.

**TSARs Enhance Nearby Candidate Genes for Therian Traits**

Beyond TSARs within clusters, we also tested three other TSARs for enhancer activity based on their association with genes whose functions might be important in therian evolution. For all three, we confirmed enhancer function and...
observed significant differences in enhancer activity between the mouse and chicken sequences.

We chose to study TSAR.4350 (Griffith et al. 2008) because therians differ in skeletal morphology from monotremes and TSAR.4350 has been shown to have enhancer activity in the developing limb (Blow et al. 2010). The expression patterns of reporter constructs with mouse and chicken TSAR.4350 sequences are extremely reproducible within each species, yet show dramatically different patterns between species (fig. 5 and supplementary fig. S6 and table S5, Supplementary Material online). Both mouse and chicken TSAR.4350 sequences show expression in developing limbs, but the major differences are in the brain and CNS. Only the mouse enhancer produces reporter gene expression in the rhombic lip, rhombomeres 3–6, the tela choioidea, and the medial pallium. Expression is also present in the roof plate epithelium in regions that are BMP+ and WNT+, yet it is absent in regions that would be FGF repressed (Sur and Rubenstein 2005). Expression picks up in the roof plate epithelium of the midbrain and pretectum, stopping at the pineal gland. Conversely, the chicken version of the enhancer is repressed anteriorly with expression in the roof plate epithelium starting at the cervical/thoracic junction. TSAR.4350 is in an intergenic region on chromosome 5 between Npr3 and Tars. Although it is difficult to predict which gene TSAR.4350 is enhancing, expression patterns fit with those of NPR3 (or NPR-C), which is responsible for clearing natriuretic peptides, which regulate blood volume and pressure, in cerebrospinal fluid and other fluids (Potter et al. 2006). Also, the medial pallium, where TSAR.4350 is expressed, gives rise to the choroid plexus, which produces cerebrospinal fluid. Future experiments could test whether this sequence might have induced mammal-specific changes in how sodium levels are regulated in cerebrospinal fluid with respect to cerebral blood flow or pressure.

Finally, we investigated two TSARs located in a gene-dense region on human chromosome 10. In the previous work, TSAR.3328 had enhancer activity in the developing forebrain (Blow et al. 2010), which is consistent with the complex pattern of activity we saw in the forebrain (fig. 6 and supplementary fig. S7 and table S5, Supplementary Material online): In the telencephalon, a longitudinal domain mapped to the progenitor zone of the ventrolateral pallium. Mantle zone activity was also found in the caudomedial pallium; this domain showed activity in continuity with the thalamus, thalami, prethalamus, and possibly the paraventricular...
hypothalamus. The spinal cord had a longitudinal domain in an intermediate dorsoventral position. A separate domain was present as highly localized expression to a small region of the pretectum; this domain was specific to the mouse construct and completely absent in embryos with the chicken ortholog of the enhancer. For TSAR.0862, which has been identified as a putative liver enhancer in mouse (OREG0036240, Griffith et al. 2008), the mouse construct caused consistent expression of the reporter gene in the epidermis and inconsistent activity in the brain, while the chicken construct produced variable patterns of expression in many embryos (supplementary fig. S7 and table S5, Supplementary Material online). Both TSAR.3328 and TSAR.0862 enhancers lie within different introns of Adk (adenosine kinase) but are closer to the transcription start site of Kat6b, a histone acetyltransferase expressed during brain development (Campeau et al. 2012). These TSARs are also within 500 kb of the Ap3m1, Dupd1, Dusp13, and Samd8 transcription start sites. Further characterization of these TSARs would be necessary to determine which genes are targeted.

Conclusions

Comparative genomics provides and approach to pinpoint the changes in coding and noncoding elements that result in the morphological and physiological differences that make us mammals. Our analysis of TSARs identified clusters of uncharacterized enhancers and proteins that are likely critical for the
evolution of therian-specific traits. Of the TSARs, ~90% show some evidence of regulatory function. These findings support the hypothesis that TSARs are signposts for shifts in function in the therian ancestor. Although we anticipated identifying elements that may have helped maintain a higher metabolic rate, a more constant body temperature, and an erect limb posture, our set of TSARs may target some of these physiological and morphological features, but predominately indicate alterations in the development of the CNS and the sensory and urogenital systems. Thus, comparative genomics approaches (such as the internal branch test) are extremely useful for discovering both expected and unexpected loci. This work brings us a step closer to unraveling the molecular mechanisms that underlie unique aspects of our biology. Now, the key is to develop high-throughput methods for ascertaining the functional impact of elements like HARs and TSARs and to discover how they relate to our unique biology.

Materials and Methods

Comparative Genomics Methods

Genomic Sequence Data

The comparative genomic data we used were derived from the University of California–San Cruz (UCSC) 46-way multiple alignment files (Blanchette 2004) that use hg19 as the reference genome. We used five high-quality genomes as representatives of eutherian mammals. Those species include human (hg19), mouse (mm9), cow (bosTau4), dog (canFam2), and elephant (loxAfr3). We also included...
sequences from two marsupials—opossum (monDom5) and wallaby (macEug1)—and a monotreme—the duck-billed platypus (ornAna1)—which all have complete genome sequences. The data set includes three members of the reptilian lineage—chicken (galGal3), zebrafish (taeGut1), and anole (anaCar1)—a single amphibian (Xenopus tropicalis, xenTro2), and several fish. The fish include zebrafish (danRen6), Fugu rubripes (fr2), stickleback (gasAcu1), medaka (oryLat2), and Tetraodon nigritipes (tetNig2). The PHAST program called tree_doctor was used to prune other taxa from the 46 species trees (Hubisz et al. 2011).

**Evolutionary Rates and Relationships**

Rate matrices and branch lengths were obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/ (last accessed January 3, 2016) and http://genomewiki.ucsc.edu/index.php/Human/hg19/GRCh37_46-way_multiple_alignment (last accessed January 3, 2016), respectively. Given their different rates of substitution, autosomes were analyzed separately from the X chromosome.

**Data Filters**

We implemented several filters on the data to obtain orthogonal elements that were well represented in the relevant taxa. First, we filtered out genomic regions overlapping human segmental duplications (genomicSuperDups) and repeat masked regions (rmsk) (Fujita et al. 2010). Second, we required synteny between human and the chicken, fish (F. rubripes, Danio rerio), and/or frog (X. tropicalis). We used the netSynteny tracks from hg18 and then converted to hg19 coordinates using liftOver from the Kent library (Hinrichs 2006). Third, we required that each element have sequence covering at least 50% of bases from a marsupial, the monotreme, and at least one nonmammalian outgroup. This requirement focused our investigation on older elements that had a burst of change in the therian ancestor rather than on sequences that are found only in therians. We focused on elements that were important throughout vertebrate evolution but changed significantly in the therian ancestor. Furthermore, these older elements may have better functional annotation. We also required a minimum length of 45 nt and 30% coverage of nucleotides within the therian mammals.

**Identification of TCRs and TSARs**

We used the PHAST program phastCons to identify elements that are conserved only among therian mammals (Hubisz et al. 2011). For running phastCons, we set rho at 0.5 for the therian subtree, which enforces a maximum rate of evolution but changed significantly in the therian ancestor. Each phastCons element was annotated with genic information from data simulated under the null model to the likelihood under a model that allows for acceleration along the therian ancestor branch. The false positive rate was obtained from data simulated under the null model.

**Simulations to Test Power and Specificity**

Data simulated with acceleration on the therian ancestor branch can be used to test power and specificity of the internal branch test (supplementary fig. S1, Supplementary Material online). The median rate that TCRs underwent substitution in our analysis was one-third the rate of 4-fold degenerate sites in protein-coding regions. Therefore, to simulate data with similar power to reject the null hypothesis (as in our empirical data), we rescaled all branch lengths in our null model tree using tree_doctor (scale=0.33; supplementary fig. S1A, Supplementary Material online). We simulated data using phyloBoot (Pollard et al. 2010) with three different rates of substitution on the therian ancestor branch (supplementary fig. S1B, Supplementary Material online, branch in red). We used the null rescaled rate, 2× and 5× higher than the null rescaled rate. For each rate, we simulated 1,000 data sets at 3 sequence lengths: 150, 300, and 1500 bp. Median lengths of TCRs and TSARs are 177 and 359 bp, respectively. We used phyloOP to test the simulated data sets for acceleration along the therian ancestor branch (Pollard et al. 2010; Hubisz et al. 2011). The false positive rate was obtained from data simulated under the null model.

**Gene Annotation**

Each phastCons element was annotated with genic information from the hg19 knownGene track in the UCSC database (Hsu et al. 2006; Fujita et al. 2011). Genic regions included UTRs, coding and noncoding exons, and introns.

**TSAR Distribution and Clusters**

We tested whether TSARs occur closer or farther apart from each other compared with the TCRs from which they are drawn (supplementary fig. S2, Supplementary Material online). In 1,000 data sets, we randomly chose 4,797 elements without replacement from the full set of TCRs. The median distance between TSARs (34,208 bp) is considerably shorter than the distance between the shuffled data sets (87,600 bp). We then defined TSAR clusters as any set of three or more TSARs where two neighboring TSARs are 50 kb apart. We used a binomial test conditioned on at least one TSAR being
present in a cluster to test whether cluster regions are enriched for TSARs; the expected proportion of TSARs is based on the overall proportion of TSARs relative to phastCons elements. All clusters are enriched for TSARs at FDR < 0.2 and the vast majority are highly enriched (FDR < 0.05).

**Laboratory Methods**

In situ hybridizations were conducted using standard methods. In situ hybridization probes were obtained from Drs. Stuart Orkin (Harvard University) and Marjo Salminen (University of Helsinki). Transgenic reporter constructs were made using a modified hsp68-LacZ construct engineered to harbor Gateway recombination sites. Either polymerase chain reaction–amplified fragments (primers in supplementary table S6, Supplementary Material online) or synthesized DNA fragments (Integrated DNA Technologies, Inc., Coralville, IA) from the mouse or chicken genome were cloned into pENTR1A and then shuttled into the modified lacZ reporter plasmid (Invivogen, San Diego, CA). Linearized DNA samples were sent to Cyagen Biosciences, Inc. (Santa Clara, CA) for pronuclear injection into the mouse embryo. We examined multiple independent transgenic embryos, each with a unique integration site. Only constructs with reproducible activity were considered. Some embryos were visualized by optical projection tomography using a Bioptonics, Inc. (Edinburgh, UK) device and analyzed with Volocity three-dimensional image analysis software (Perkin Elmer, Waltham, MA).

**Supplementary Material**

Supplementary figures S1–S2 and tables S1–S6 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We thank MJ. Hubisz for implementing the --branch test in phyloP; J.D. Wythe and W.P. Devine for the transgenic reporter constructs and advice; C. Miller (Gladstone Histology Core) for performing optical projection tomography and virtual sections; F. Kreitzer for advice on incubating and dissecting the chicken egg; D. Kostka for discussions and scripts for testing distance between TSARs; and S. Orkin and M. Salminen for in situ hybridization probes. Transgenic mice were generated by Cyagen Biosciences, Inc. This work was supported by the Gladstone Institutes, NIH R01-GM82901 (A.K.H., K.S.P.), the Lawrence J. and Florence A. DeGeorge Charitable Trust/American Heart Association Established Investigator Award (B.G.B.), NIH/NCCR grant (C06 RR018928) to the Gladstone Institutes, William H. Younger, Jr (B.G.B.), and Nina Ireland and NIMH R37-MH049428 (J.L.R.).

**References**


Downloaded from https://academic.oup.com/mbe/article-abstract/33/4/1008/2579545/Accelerated-Evolution-of-Enhancer-Hotspots-in-the by guest on 15 September 2017


