Accelerated Evolution of Enhancer Hotspots in the Mammal Ancestor

Alisha K. Holloway[1,2], Benoit G. Bruneau[1,3], Tatyana Sukonnik[1], John L. Rubenstein[4], and Katherine S. Pollard*[,1,2,5]

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, CA

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

[3]Department of Pediatrics and Cardiovascular Research Institute, University of California, San Francisco, CA

[4]Nina Ireland Laboratory of Developmental Neurobiology and Department of Psychiatry, University of California, San Francisco

[5]Institute 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; Koshiba-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. 1*A*). 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).

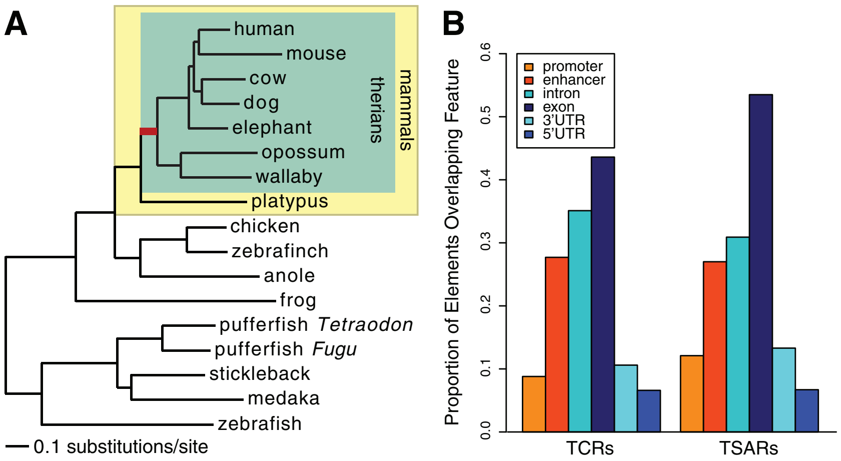

Fig. 1. (*A*) Phylogenetic tree of vertebrates used to identify TSARs; internal test for acceleration was conducted on the red branch, which unites therians. (*B*) TSARs and TCRs are distributed among genomic features in similar proportions with slightly more exonic and less intronic sequence in TSARs.

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 and possibly more constant body temperature, 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 among therians (therian conserved regions; TCRs), of which 4,797 have a strong signature for accelerated evolution in the therian ancestor (false discovery rate <1%; supplementary table S1, Supplementary Material online). We call these 4,797 sequences Therian-Specific Accelerated Regions (TSARs). Using simulations (see Materials and Methods), we established that the power of this method is sufficient to identify TSARs and that specificity is high for elements over 150 bp. For longer elements, power and specificity are remarkably high (supplementary fig. S1C, Supplementary Material online).

The distribution of TSARs with respect to human gene features is similar to that of TCRs (fig. 1*B*). 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. 2*A*). 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) centers 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. 2*B* 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. 2*A*) (supplementary fig. S3*A*, Supplementary Material online). Taken together, the protein-coding TSAR clusters were not enriched for any particular biological processes. In comparison, as with 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. S3*B*, 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

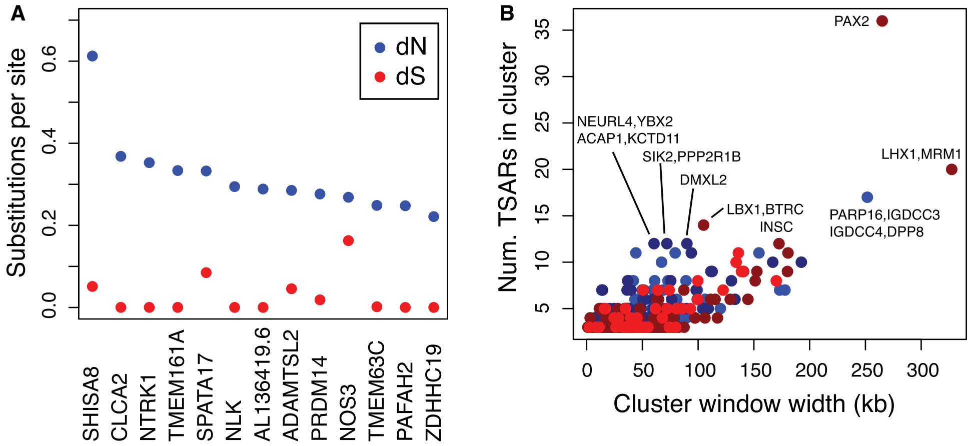

Fig. 2. (*A*) Thirteen human genes containing TSARs with nonsynonymous substitution rates (dN) at least 5× higher than the median and synonymous substitution rates (dS) less than half the median rate. This pattern may indicate adaptive protein evolution resulting in altered functions. (*B*) Clusters of three or more TSARs, where neighboring TSARs ≤50 kb apart are found in 383 genomic loci. For the densest clusters, the nearest genes are labeled. Clusters primarily composed of coding sequence are in blue (>75% dark blue; >50% blue); predominantly noncoding sequence clusters are in red (<50% coding sequence, red; <25% coding sequence, dark red).

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; García-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 into circulation (Vandesande and Dierickx 1975). Interestingly, oxytocin is involved in uterine contractions and milk ejection (Cross 1955), whereas vasopressin regulates blood pressure (Rocha E Silva and Rosenberg 1969) and temperature (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. S3C, 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

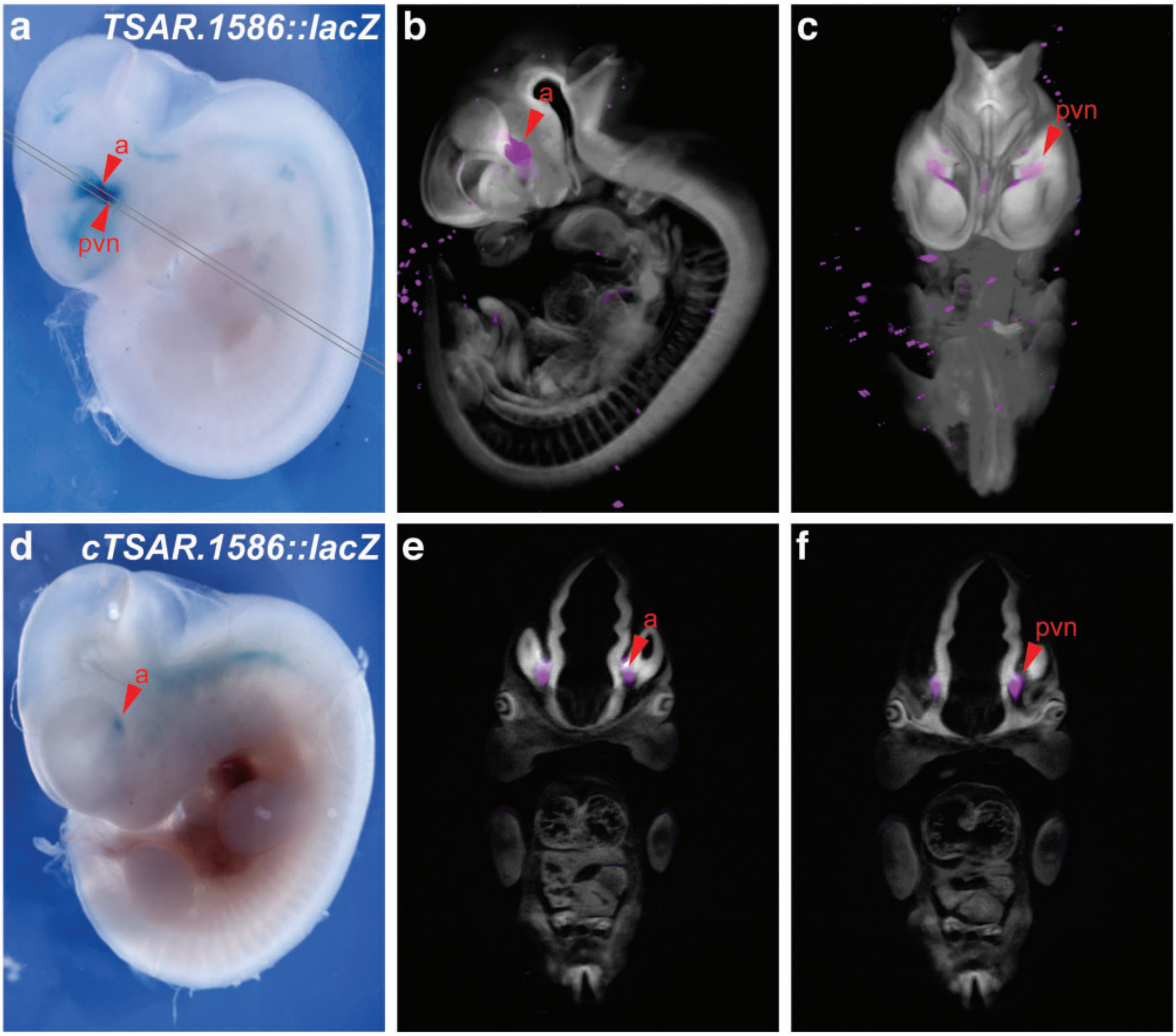

Fig. 3. Enhancer activity of TSAR.1586. (*A* and *D*) LacZ-stained and (*B*, *C*, *E*, *F*) reconstructed three-dimensional image from optical projection tomography (OPT) of e11.5 mouse embryos show the enhancer activity of TSAR.1622 for (*A*–*C*, *E*, *F*) the mouse sequence and (*D*) the orthologous chicken sequence. (*C*) Rostral view of reconstructed OPT shows expression in the hypothalamus and telencephalon for the mouse construct. (*E* and *F*) Coronal sections from the mouse construct show specific expression along the telencephalic stalk into (*E*) the anlage of the amygdala and (*F*) the anlage of the paraventricular nucleus. pvn: paraventricular nucleus; a: amygdala.

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 OREG002948 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

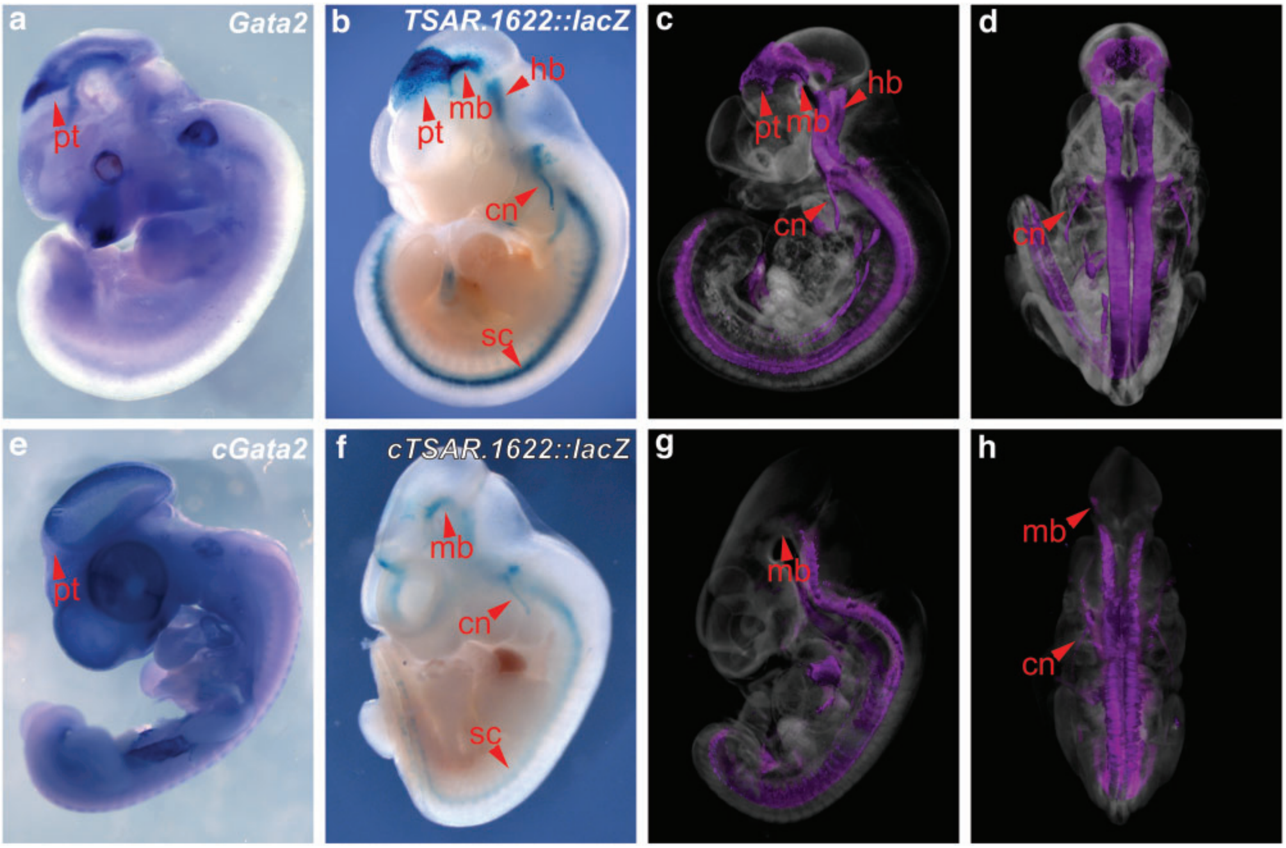

Fig. 4. Enhancer activity of TSAR.1622. (*A* and *E*) Endogenous *Gata2* expression by in situ hybridization in (*A*) e11.5 mouse and (*E*) stage HH25 chicken. (*B* and *F*) LacZ-stained e11.5 mouse embryos show enhancer activity of TSAR.1622 for (*B*) the mouse sequence construct and (*F*) the orthologous chicken sequence construct. Both overlap expression of *Gata2*. (*C–D, G–H*) Lateral and dorsal views of OPT in transgenic mice. (*B–D*) The mouse TSAR.1622 sequence shows strong expression in the spinal cord and mantle zone along a longitudinal band in an intermediate dorsal/ventral position of the midbrain, terminating in the pretectum, whereas (*F–H*) the chicken sequence does not show expression in the pretectum with (*G*) autofluorescence of the blood in the heart. pt: pretectum; cn: facial nerve; mb: midbrain; hb: hindbrain; sc: spinal cord.

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 chorioidea, 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 eminentia thalami, prethalamus, and possibly the paraventricular

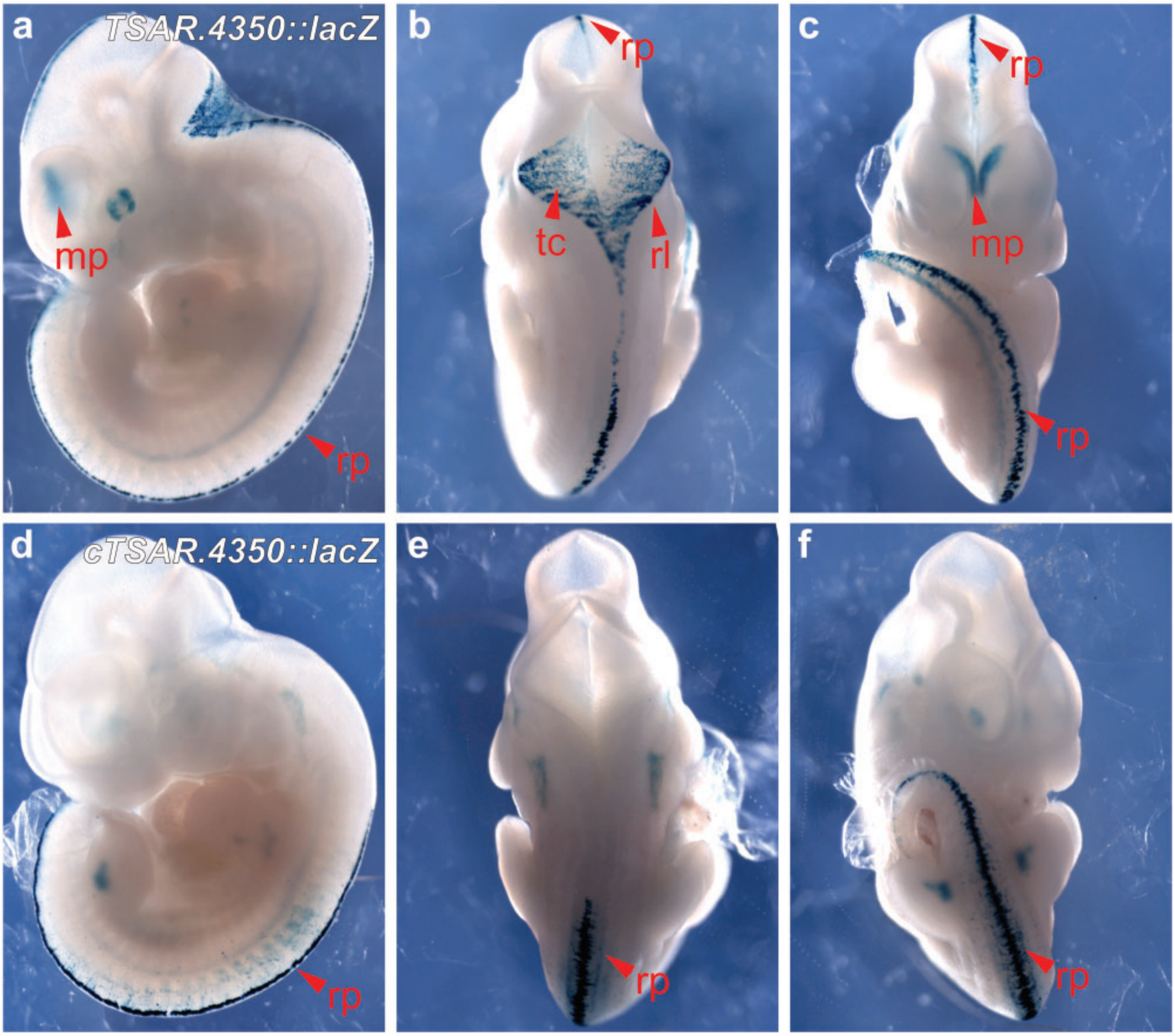

Fig. 5. Enhancer activity of TSAR.3328. (*A* and *D*) LacZ-stained e11.5 mouse embryo expression of enhancer TSAR.3328 for (*A*) the mouse sequence construct and (*D*) the orthologous chicken sequence construct. (*B* and *C*) Lateral and rostral view of reconstructed OPT shows expression patterns for the mouse construct. Both (*A*–*C*) mouse and (*D*) chicken versions are expressed in the ventrolateral pallium, prethalamus, hindbrain, and spinal cord. (*A*–*C*) Mouse-specific expression is in the pretectum and caudomedial pallium. (*E* and *F*) Coronal sections from the mouse construct show specific expression in (*E*) the caudomedial pallium and (*F*) pretectum. pt: pretectum; pTh: prethalamus; vlp: ventrolateral pallium; cmp: caudomedial pallium.

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

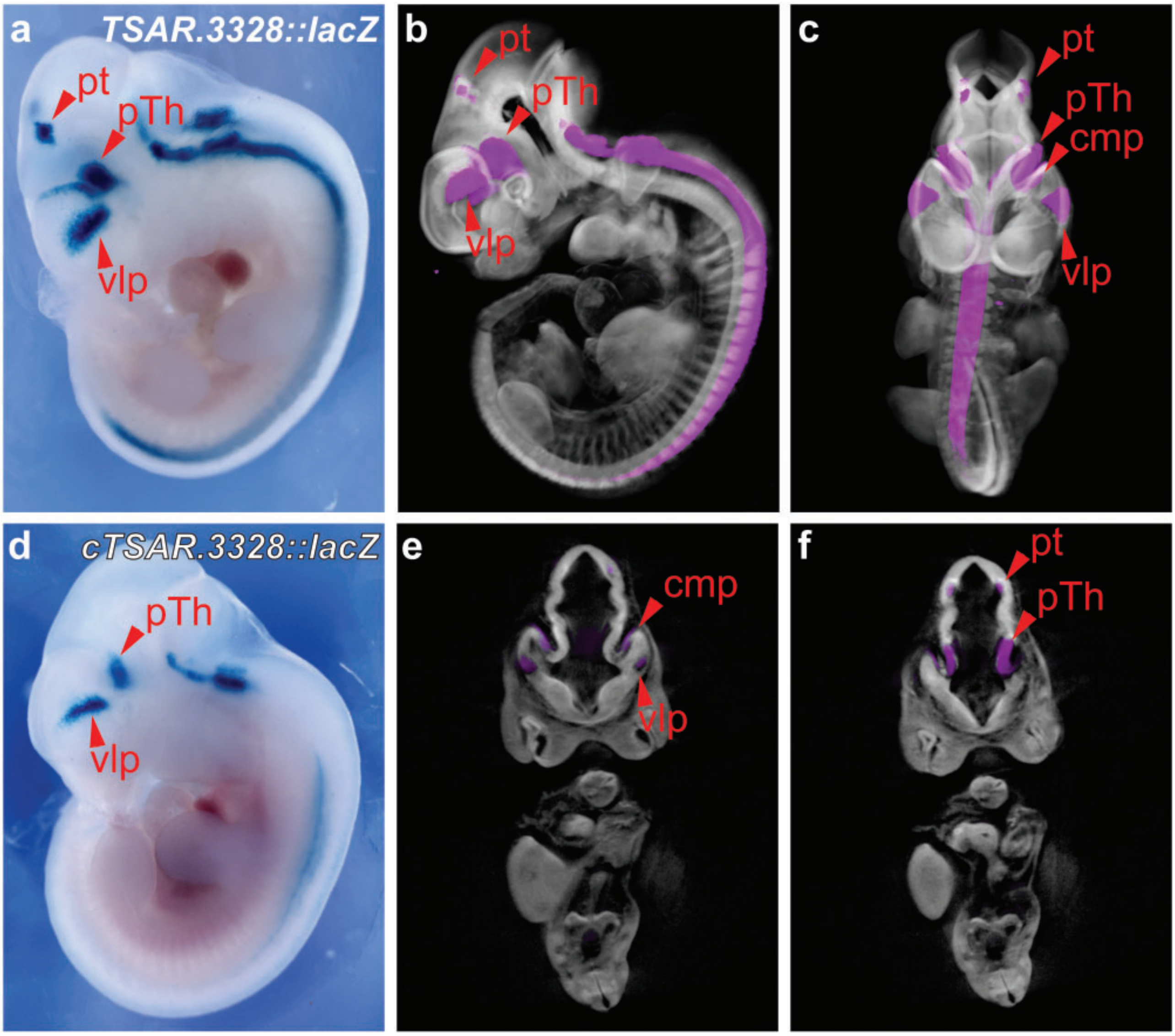

Fig. 6. Enhancer activity of TSAR.4350. (*A*–*F*) LacZ-stained e11.5 mouse embryos show enhancer activity for (*A*–*C*) the mouse construct and (*D*–*F*) the chicken construct. (*B* and *E*) Caudal view showing mouse-specific expression in (*B*) the tela chorioidea, rhombic lip, and dorsal roof plate for the mouse version and (*E*) the roof plate posterior to the cervical/thoracic junction in the orthologous chicken version. (*C* and *F*) Rostral view showing expression in the medial pallium and dorsal roof plate to the pineal gland are specific to the (*C*) mouse construct and, showing expression in the posterior roof plate for both (*F*) chicken and (*C*) mouse constructs. mp: medial pallium; rp: roof plate; tc: tela chorioidea; rl: rhombic lip.

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), zebrafinch (taeGut1), and anole (anoCar1)—a single amphibian (*Xenopus tropicalis*, xenTro2), and several fish. The fish include zebrafish (danRer6), *Fugu rubripes* (fr2), stickleback (gasAcu1), medaka (oryLat2), and *Tetraodon nigripes* (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 orthologous 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 of half the rate derived from 4-fold degenerate sites in coding regions. Stringent filtering followed by identifying conserved elements with phastCons resulted in 177,346 TCRs.

New code was developed in the PHAST program *phyloP* (Pollard et al. 2010; Hubisz et al. 2011) to implement the internal branch test (--branch option). Previously, it was possible to test a single terminal branch or an entire subtree. The null model is the same: It is a continuous time Markov model of nucleotide substitutions derived from neutral sequences. Here we used the REV model fit to 4-fold degenerate sites (separately for chromosome X and the autosomes). Each genomic region may be evolving faster or slower in all vertebrates; therefore all branch lengths of the phylogenetic tree from this model are then rescaled by a parameter to reflect the local expected substitution rate. This results in all branches of the tree being stretched or contracted by a single factor for each genomic region. The alternative model for the internal branch test is the null model with the therian ancestral branch scaled by a second parameter, which is constrained to be greater than 1 to reflect accelerated evolution in the ancestor of all therian mammals. Both models are fit to the multiple sequence alignment of each TCR by maximum likelihood. Thus, the internal branch test is a one-sided likelihood ratio test that compares the likelihood of data given the null model to the likelihood under a model that allows for acceleration along the therian ancestor branch. The *P* values from the internal branch test were adjusted for multiple comparisons using the false discovery rate (Benjamini and Hochberg 1995).

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. S1*A*, 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. S1*B*, 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 *phyloP* 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 $\leq$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–S7 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank M.J. 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/NCRR 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

Abellan A, Menuet A, Dehay C, Medina L, Retaux S. 2010. Differential expression of LIM-homeodomain factors in Cajal-Retzius cells of primates, rodents, and birds. *Cereb Cortex*. 20:1788–1798.

Ashwell K. 2013. Neurobiology of monotremes. Victoria (BC): CSIRO Publishing.

Bardet SM, Martinez-de-la-Torre M, Northcutt RG, Rubenstein JLR, Puelles L. 2008. Conserved pattern of OTP-positive cells in the paraventricular nucleus and other hypothalamic sites of tetrapods. *Brain Res Bull*. 75:231–235.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, James Kent W, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441:87–90.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 57:289–300.

Bickelmann C, Morrow JM, Müller J, Chang BSW. 2012. Functional characterization of the rod visual pigment of the echidna (*Tachyglossus aculeatus*), a basal mammal. *Vis Neurosci*. 29:211–217.

Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, Clarke SL, Wenger AM, Nguyen L, et al. 2012. Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res*. 22:1059–1068.

Blanchette M. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 14:708–715.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet*. 42:806–810.

Brandt W, Khandekar M, Suzuki N, Yamamoto M, Lim KC, Engel JD. 2008. Defining the functional boundaries of the *Gata2* locus by rescue with a linked bacterial artificial chromosome transgene. *J Biol Chem*. 283:8976–8983.

Campeau PM, Kim JC, Lu JT. 2012. Mutations in KAT6B, encoding a histone acetyltransferase, cause Genitopatellar syndrome. *Am J Hum Genet*. 90:282–289.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* 327:302–305.

Cross BA. 1955. The posterior pituitary gland in relation to reproduction and lactation. *Br Med Bull*. 11:151–155.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 9:215–216.

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* 474:598–603.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2010. The UCSC Genome Browser database: update 2011. *Nucelic Acids Res*. 39:D876–D882.

García-Moreno F, Pedraza M, Di Giovannantonio LG, Di Salvio M, López-Mascaraque L, Simeone A, De Carlos JA. 2010. A neuronal migratory pathway crossing from diencephalon to telencephalon populates amygdala nuclei. *Nat Neurosci*. 13:680–689.

Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, et al. 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucelic Acids Res*. 36:D107–D113.

Heesy CP, Hall MI. 2010. The nocturnal bottleneck and the evolution of mammalian vision. *Brain Behav Evol*. 75:195–203.

Hinrichs AS. 2006. The UCSC Genome Browser Database: update 2006. *Nucelic Acids Res*. 34:D590–D598.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* 22:1036–1046.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform*. 12:41–51.

Koshiba-Takeuchi K, Mori AD, Kaynak BL, Cebra-Thomas J, Sukonnik T, Georges RO, Latham S, Beck L, Henkelman RM, Black BL, et al. 2009. Reptilian heart development and the molecular basis of cardiac chamber evolution. *Nature* 461:95–98.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A*. 104:8005–8010.

Luo ZX, Yuan CX, Meng QJ, Ji Q. 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476:442–445.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature* 448:587–590.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* 447:167–177.

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J Mol Biol*. 382:236–245.

Noonan JP. 2009. Regulatory DNAs and the evolution of human development. *Curr Opin Genet Dev*. 19:557–564.

Nozawa D, Suzuki N, Kobayashi-Osaki M, Pan X, Engel JD, Yamamoto M. 2009. GATA2-dependent and region-specific regulation of *Gata2* transcription in the mouse midbrain. *Genes Cells* 14:569–582.

Okuno A, Yamamoto M, Itoh S. 1965. Lowering of the body temperature induced by vasopressin. *Jpn J Physiol*. 15:378–387.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 20:110–121.

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.

Potter LR, Abbey-Hosch S, Dickey DM. 2006. Natriuretic peptides, their receptors, and cyclic guanosine monophosphate-dependent signaling functions. *Endocr Rev*. 27:47–72.

Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470:279–283.

Rich TH. 2005. Independent origins of middle ear bones in monotremes and therians. *Science* 307:910–914.

Rocha E Silva M, Rosenberg M. 1969. The release of vasopressin in response to haemorrhage and its role in the mechanism of blood pressure regulation. *J Physiol (Lond)*. 202:535–557.

Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci U S A*. 105:4220–4225.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* 488:116–120.

Sur M, Rubenstein JLR. 2005. Patterning and plasticity of the cerebral cortex. *Science* 310:805–810.

Vandesande F, Dierickx K. 1975. Identification of the vasopressin producing and of the oxytocin producing neurons in the hypothalamic magnocellular neurosecretory system of the rat. *Cell Tissue Res*. 164:153–162.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457:854–858.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucelic Acids Res*. 35:D88–D92.

Wakefield MJ, Anderson M, Chang E, Wei KJ, Kaul R, Graves JAM, Grützner F, Deeb SS. 2008. Cone visual pigments of monotremes: filling the phylogenetic gap. *Vis Neurosci*. 25:257–264.

Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res*. 16:1480–1492.

Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol*. 24:2755–2762.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 453:175–183.

Wederell ED, Bilenky M, Cullum R, Thiessen N, Dagpinar M, Delaney A, Varhol R, Zhao Y, Zeng T, Bernier B, et al. 2008. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucelic Acids Res*. 36:4549–4564.

Widelitz RB, Veltmaat JM, Mayer JA, Foley J, Chuong CM. 2007. Mammary glands and feathers: comparing two skin appendages which help define novel classes during vertebrate evolution. *Semin Cell Dev Biol*. 18:255–266.

Willett RT, Greene LA. 2011. Gata2 is required for migration and differentiation of retinorecipient neurons in the superior colliculus. *J Neurosci*. 31:4444–4455.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*. 8:206–216.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*. 24:1586–1591.

Zhou Y, Lim KC, Onodera K, Takahashi S, Ohta J, Minegishi N, Tsai FY, Orkin SH, Yamamoto M, Engel JD. 1998. Rescue of the embryonic lethal hematopoietic defect reveals a critical role for GATA-2 in urogenital development. *EMBO J*. 17:6689–6700.