TRANSPOSON EXAPTATION IN MAMMALIAN EVOLUTION

by

DONALD HUCKS

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN BIOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2008

ACKNOWLEDGEMENTS

I would like to express my appreciation for the contributions of several people to the research presented herein. Claudio Casola made significant contributions to the *pogo* project, including the mining of *pogo* transposons. Clément Gilbert collaborated in the analysis of *tigger1* transposons and the *tigd1* gene. In PCR and sequencing, I was assisted by Stephanie Ethofer, Ahmad Gaber, John McCormick, Marguerite Schaeffer, and Richard Cordaux, who designed the first set of primers. Mark Batzer graciously provided anthropoid DNA. Experiments performed by Swalpa Udit contributed to the definition of *mbs*, applied in the statistical model. John Pace provided technical support, including automation of the BLAST interrogations of anthropoid genomes. I would like to thank the UTA Genome Biology Group and the UTA Mobile DNA Group for helpful discussions. I would like to thank my thesis committee: Dr. Esther Betrán, Dr. Ellen Pritham, and Dr. James Robinson, for their assistance, encouragement, and support. I would like to thank my mentor and thesis advisor, Dr. Cédric Feschotte, for his guidance, for his encyclopedic knowledge of all things transposable, and for his boundless enthusiasm for science and the scientific life. Finally, I would like to thank my wife, Stacey, for virtues too numerous to list.

This research was supported by an NIH grant to CF.

April 15, 2008

ABSTRACT

TRANSPOSON EXAPTATION IN MAMMALIAN EVOLUTION

Donald Hucks, M.S.

The University of Texas at Arlington, 2008

Supervising Professor: Cédric Feschotte

A growing body of work suggests that the exaptation of transposon-derived sequences to perform beneficial cellular functions has played a significant role in eukaryotic evolution. In chapter 1, we present an analysis of 10 exapted *pogo* transposons in the human genome. We present evidence that all 10 are restricted to tetrapods, and that 8 of the 10 arose early in mammalian evolution, in several independent exaptation events involving diverse *pogo* lineages. We show that all 10 have been subject to stringent selection throughout mammalian evolution, with pseudogenization having occurred only infrequently. In 4 of these genes, we observed no cases of gene loss, consistent with a very high selective value for these genes. We also present evidence that all 10 genes encode sequence-specific DNA-binding proteins, each likely to bind a highly constrained, but as yet unknown, sequence somewhere within the mammalian genome.

In chapter 2, we present evidence that a motif occurring within the terminal inverted repeats (TIRs) of *Hsmar1* and *made1* transposons is subject to purifying selection at a number of loci in anthropoid genomes as binding sites for SETMAR, the protein product of an anthropoid-specific gene formed some 50 mya by fusion of an *Hsmar1* transposase gene with an extant histone methyltransferase gene.

Together, our analyses support the notion that this complementary nature of transposon exaptation, the host recruitment of transposase as DNA-binding protein and non-coding transposon sequence as binding site, has been a recurrent theme in mammalian evolution.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

RECURRENT DOMESTICATION OF *POGO* TRANSPOSONS IN MAMMALIAN EVOLUTION

1.1 Introduction

Transposable elements (TEs) are selfish, or parasitic, mobile genetic elements, capable of catalyzing their own transposition by mechanisms which, directly or indirectly, result in an increase in genomic copy number over time (Orgel and Crick 1980, Doolittle and Sapienza 1980, Hickey 1982, Charlesworth et al. 1994). Class 1 elements, or retrotransposons, transpose via an RNA intermediate, by a replicative "copy and paste" mechanism. Class 2 elements, or DNA transposons, transpose via a "cut and paste" mechanism involving the excision of the transposon by an element-encoded enzyme, a transposase, which binds a sequence occurring within the terminal inverted repeats (TIRs) of the transposon and catalyzes breakage and joining reactions, inserting the element at another locus. Although not directly replicative, host DSB (double strand break)-repair mechanisms may restore the transposon sequence at the original locus, resulting in an increase in copy number (Craig et al 2002).

Transposable elements are widespread across eukarya, and often comprise substantial portions of eukaryotic genomes (Craig et al 2002). In fact, transposable elements and their remnants comprise nearly half the human genome, mostly in the form of molecular fossils, long dead and accumulating mutational decay over time. All TEs known to be currently active in the human genome are Class 1 elements, most notably the Long Interspersed Nuclear Element (LINE), L1, and the Short Interspersed Nuclear Element (SINE), Alu, which together make up more than a third of the human genome, although only a small proportion of these elements remain active. Much less abundant, in humans, are Class 2 elements, remnants of which comprise ~3% of the human genome. No Class 2 elements have been active in primates in the last 40 myr (Lander et al. 2001, Pace and Feschotte 2007).

The high copy numbers of TEs in many eukaryotic genomes, especially those of plants and various metazoans, suggests that most individual TEs and transposition events are selectively neutral to mildly deleterious. Undoubtedly, TEs sometimes produce highly deleterious effects, as well. Such events may include chromosome breakage, disruption of exons, displacement of regulatory sequences, and chromosomal rearrangement by ectopic recombination of homologous TE sequences (Orgel and Crick 1980, Doolittle and Sapienza, Hickey 1982, Charlesworth et al. 1994, Kidwell and Lisch 2001). On occasion, however, a TE may undergo exaptation, that is, the acquisition of a novel cellular function, conveying a benefit to the host organism (Gould and Vrba 1982, Brosius and Gould 1992). This genomic co-opting of a molecular parasite is often referred to as molecular domestication (Miller et al, 1999). The exapted sequence may be coding or non-coding DNA. Conferring a beneficial function, it becomes subject to natural selection, and may be carried to fixation by positive selection and long maintained in a population by purifying selection. Although Class 2 TEs are much less abundant in the human genome than are Class 1 TEs, it appears that exaptation of Class 2 elements is much more common than exaptation of Class 1 elements. Of 47 human genes identified as exapted transposons, all but 4 were derived from Class 2 elements, or DNA transposons (Lander et al, 2001).

A common theme in transposon exaptation appears to be recruitment of transposase as a host DNA-binding protein (Feschotte and Pritham, 2007). An example is CENPB, a domesticated pogo transposase of 599 amino acid, which binds a highly conserved 17-bp motif, the CENP-B box, within the alpha satellite repeats of mammalian centromeres (Masumoto et al.1989; Yoda et al. 1992; Tanaka et al. 2001). Sequences displaying high identity to human CENPB have also been isolated in hamster, sheep and several primates (Haaf et al. 1995; Burkin et al. 1996; Goldberg et al. 1996; Yoda et al. 1996; Bejarano and Valdivia 1996). The mouse CENPB homolog, which is 92% identical to the human protein, has been observed to bind a DNA motif highly similar to the CENP-B box in humans. Recently, a similar motif was

identified within the centromeric satellite repeats of the marsupial Macropus rufogriseus (Bulazel et al. 2006). A fragment containing the motif was bound in vitro by recombinant human CENPB protein (Bulazel et al. 2006), suggesting that it binds a yet to be discovered CENPB homolog.

Although some authors have reported the presence of sequences similar to the CENPB box in the satellite DNA repeats of Xenopus, insects and plants (Coelho et al. 1996; Lopez and Edstrom 1998; Weide et al. 1998; Heslop-Harrison et al. 1999; Nonomura and Kurata 1999; Lorite et al. 2004; Mravinac, Plohl, and Ugarkovic 2004; Edwards and Murray 2005), the similarity is weak and no homolog of CENPB has been detected in any non-mammalian species. The precise function of CENPB at the centromere is unknown and it has been noted that Cenpb knockout mice are viable and fertile, with only mild phenotypic effects and no obvious deficiency in chromosome segregation (Hudson et al. 1998; Kapoor et al. 1998; Perez-Castro et al. 1998; Fowler et al. 2000).

Pogo transposons, with which *cenpb* shares homology (Smit and Riggs 1996; Kipling and Warburton 1997), are members of the *Tc1/mariner* superfamily of Class 2 transposons (Capy et al 1998, Plasterk et al 1999) and are widely distributed across eukaryotic genomes (Robertson 1996, Smit and Riggs 1996, Kapitonov and Jurka 1999, Feschotte and Mouches 2000). *Pogo* transposons are currently active in a wide variety of eukaryotic lineages, including several metazoans, but no active *pogo* element has been described in vertebrates. The results of the mouse experiments, together with evidence for partial redundancy among a trio of centromere-binding proteins in fission yeast (Baum and Clarke 2000; Irelan et al 2001; Nakagawa et al 2002), which exhibit homology to each other and, incidentally, to a lineage of fungal *pogo* (Murakami et al 1996; Lee et al 1997; Irelan et al 2001; Casola et al 2008), has given rise to speculation that the mammalian genome may encode protein(s) complementary to CENPB with redundant centromeric function (Hudson et al 1998; Kapoor et al 1998; Irelan et al 2001; Nakagawa et al 2002). *Cenpb* is one of 10 human refseq genes (all with status validated or reviewed), known to have been derived from *pogo* transposons, based on sequence

similarity; Smit and Riggs 1996; Kipling and Warburton 1997; Zeng et al 1997; Dou et al 2004). The others are *jerky* (Toth et al. 1995, Zeng et al. 1997)*, jerky-like* (Dou et al 2004), and 7 "*Tigger*-derived" genes, *tigd1-7* (Smit and Riggs 1996; Kipling and Warburton 1997; Dou et al. 2004). All are single copy genes, in humans, comprised of a single uninterrupted open reading frame (ORF). They are predicted to encode proteins ranging in size from 471 to 593 amino acid. *Jerky* is known to encode a DNA- and RNA –binding protein with activity in neurons. Knockout mice are prone to epileptic seizures and tremors (Toth et al. 1995). Nothing is known, however, about the functions of the remaining 8 POGO-derived proteins. Therefore, it seems reasonable to wonder whether 1 or more of these genes sharing homology with *cenpb*, might function with some redundancy at the centromere.

As a first step in elucidating the functions of this group of genes, we resolved to gain insight into their evolutionary histories. We wished to determine when they were exapted, and whether they arose by a single exaptation followed by multiple gene duplication events or, rather, from several independent exaptations. We also wished to gain insight into their biological significance, as signified by stringency of constraint and frequency of pseudogenization. Finally, we sought intragenic patterns of evolution, which would allow us to test our hypothesis that each of these genes, like *cenpb* and *jerky*, encodes a sequence-specific DNA-binding protein. Such an observation would be consistent with both the theoretical prediction that exaptation involves the acquisition of a novel function by an existing structure (in this case a molecular structure), and numerous empirical observations, suggesting transposon exaptation is an important mechanism by which novel DNA-binding proteins arise in evolution (Feschotte and Pritham 2007). In pursuit of these objectives, we utilized publicly available genomic databases and applied the statistical principles of molecular evolution.

1.2 Methods

We obtained amino acid sequences for CENPB, JRK, JRKL, and TIGD1-7 from the University of California at Santa Cruz Genome Browser. (Accession numbers: CENPB

NM_001810, JRK NM_001077527, JRKL NM_003772, TIGD1 NM_145702, TIGD2 NM_145715, TIGD3 NM_145719 , TIGD4 NM_145720, TIGD5 NM_032862, TIGD6 NM_030953, TIGD7 NM_033208). We used each sequence as a query against the translated nucleotide sequences (TBLASTN) of all organisms in the NCBI genomic databases (Altschul et al. 1990). Synteny was confirmed, where possible, by using each putative ortholog as a BLAT query against its respective genomic sequence in the UCSC Genome Browser (Kent 2002; Kent et al. 2002; Karolchik et al. 2003; Karolchik et al. 2004). We then compared flanking genes, as annotated therein. Transposon sequences, used in the Neighbor-Joining tree, were mined and analyzed for the presence of TIRs and target site duplications (TSD), and phylogenetically characterized using Mr. Bayes. Pairwise alignments were constructed using ClustalX (Chenna et al. 2003) and manually refined using Bioedit v7.0.5.3 (Hall 1999) and GeneDoc. v2.6.002 (Nicholas and Nicholas 1997). K-estimator (Comeron 1999) was used to tabulate the number of non-synonymous sites and substitutions as well as frequencies for each of three classes of synonymous sites and substitutions (2-S, 2-V, and 4-fold). Using these data, we calculated dN/dS by the Pamilo-Bianchi-Li method (Pamilo and Bianchi 1993, Li et al 1993). For the maximum likelihood method, we utilized the free ratio branch model of the codeml program in the PAML suite (Yang 1997). The maximum likelihood analog of dN/dS, returned by codeml, is termed omega (ω) and is calculated based on inferred ancestral sequences at each node of an input phylogeny. To formally test whether the observed ω on each individual branch was significantly < 1, we ran a series of models in which all branches were free, except 1, which was constrained to ω=1. Likelihood ratio tests were then compared to a chi-squared distribution, with 1 degree of freedom. For the partitioned estimation of omega by domain, we used the 1-ratio site model in codeml in mgene mode. We tested whether omega for each partition was significantly < 1 by the method described above, with degrees of freedom calculated according to Yang (1997). We tested whether the difference in omega between each pair of partitions was significant, using a null model in which the two partitions were combined into a single unit with

homogenous omega. In both cases, we applied a Bonferoni correction for multiple tests. We confined this analysis to full length sequences comprising unambiguous ORFs, using the following input trees, aided by Treeview:((((CENP-B_hs, CENPB_Papio_anubis),(CENP-B_Ateles_geoffroyi,CENP-B_Aotus_nancymaae, CENPB_Callithrix_jacchus)), CENP-B_Lemur_catta), ((CENP-B_mus,CENP-B_rat), CENPB_hamster),CENP-B_opo); (((jrk_horse,(jrk_dog,jrk_cat)),jrk_cow),((((jrk_hs,jrk_chim),jrk_mac),jrk_gal),(jrk_rat,jrk_mus)),jrk_opo); (((jrkl_hs,jrkl_mac),(jrkl_rat,jrkl_mus)),((jrkl_cat,jrkl_dog),jrkl_cow),jrkl_opo); (((tigd1_hs,tigd1_chimp),tigd1_mac),(tigd1_dog,tigd1_cow),tigd1_tenrec); (((tigd2_hs,tigd2_chim), tigd2_mus), ((tigd2_horse, tigd2_dog), tigd2_cow), tigd2_opo); ((((tigd3_hs,tigd3_chim),tigd3_tree_shrew),(tigd3_mus,tigd3_rat)), (tigd3_horse,tigd3_dog),tigd3_cow); (((((((tigd4_hs,tigd4_chim), tigd4_mac), tigd4_gal), (tigd4_mus,tigd4_rat)), (((tigd4_cow, (tigd4_horse,tigd4_dog)), tigd4_bat), tigd4_hdg)), tigd4_arm),tigd4_ele,tigd4_plat); ((tigd5_hs, (tigd5_mus, tigd5_rat)), (tigd5_dog, tigd5_cow), tigd5_opo); (((tigd6_chim, tigd6_hs), tigd6_sqrl),(tigd6_dog,tigd6_horse), tigd6_opo); ((((tigd7_hs, tigd7_chim), tigd7_mac), tigd7_squi), (tigd7_horse, tigd7_dog), tigd7_eshr); The Neighbor-Joining tree in Figure 2 was obtained using MEGA 3.1 (Kumar, Tamura, and Nei 2004). The consensus logo (Schneider and Stephens 1990) was constructed using the program Weblogo (Crooks et al. 2004).

In the phylogenetic analysis of *tigd1*, we conducted a BLAT search (UCSC Genome Browser) of the human, dog, and cow genomes, using the nucleotide sequences of *tigd1* from human, dog, and cow as respective queries. From each of these genomes, we then retrieved the full length *Tigger1* transposon within which the gene resides. We then used these full-length transposon sequences as BLAT queries against their respective genomes, and selected *Tigger1* sequences, based on size and similarity to the query. We produced multiple alignments, using ClustalX, and used GeneDoc to edit the alignments. We produced the

Neighbor-Joining trees using Mega 3.1, with Kimura's 2-parameter model and pairwise deletion of gaps.

1.3 Results

1.3.1 Identification of Orthologs

With 2 exceptions, described below, all interrogations of the NCBI genomic databases, using predicted protein sequences for human *pogo*-derived genes as queries, yielded results that fell into 1 of 2 categories. The first category was comprised of sequences > 70% identical to the query. The second category was comprised of sequences < 35% identical to the query. We considered those sequences in the former category to be putative orthologous genes worthy of further analysis. The latter category appears to be populated by various members of the remaining *pogo*-derived genes (which, when translated, share identity ranging from ~20% to ~30%) as well as molecular fossils of various ancient *pogo* transposons. The exceptions are *tigd2* and *jerky-like*, which share ~65% amino acid identity over their entire lengths.

For eight of these queries, all of the putative genes (> 70% identical to the query) were retrieved from the genomes of mammals. In addition to this distribution, an apparent ortholog of *tigd4* was also found in the genome of the lizard, *Anolis carolinensis*. Apparent orthologs of *tigd5* were retrieved from the genomes of the chicken, *Gallus gallus*, and the frog, *Xenopus tropicalis*. The large number of completely sequenced animal genomes publicly available (3 non-mammalian vertebrates: chicken, the squamate *Anolis carolinensis*, and the amphibian *Xenopus tropicalis*; 1 echinoderm, 3 ascidians, 12 flies, 3 mosquitoes, 1 beetle, 1 lepidopteron, 3 nematodes, 2 flatworms and 1cnidarian), several with high coverage (> 6X), in addition to on-going sequencing projects, suggests that the restriction of these sequences to the aforementioned lineages is no mere technical artifact, but is, rather, an accurate reflection of their taxonomic distribution. This is also supported by previously published findings which identified *pogo* transposons in a variety of eukaryotic lineages, while finding homologous genes only within the distribution described above (Casola et al 2008).

Next, we analyzed each of the sequences for evidence that some or all of them represent orthologous and extant, functional genes. First, we evaluated the sequences for the presence of open reading frames (ORFs), as indicated by the presence of a start codon and the absence of frameshift and nonsense mutations. Thus, we observed several apparent pseudogenes, with multiple frameshift and nonsense mutations. In several cases, a sequence with high identity to the query included a single, or very few, such mutations. In these cases, we interrogated the trace files available in the NCBI databases for evidence that these could have been sequencing errors, not uncommon in low coverage genome sequencing. Such cases in which we could not obtain evidence for a sequencing error, due to limited data, were noted as uncertain regarding the presence of an ORF. We will return to this distinction below.

Next, we evaluated synteny, that is, the occurrence of the same gene sequence in the same genomic environment among diverse genomes. An observation of synteny is evidence of orthology. To this end, we utilized the BLAT function in the UCSC Genome Browser. Using each sequence as a query against its respective genome, we identified the unique genomic locus of each sequence. Then we noted the annotated genes which flanked the query on either side. In this analysis, we were, of course, limited to those sequences from complete genomes, annotated in the Genome Browser (human, chimp, macaque, mouse, rat, cat, dog, horse, cow, opossum, and platypus). With the exception of *tigd1*, to be discussed below, we were able to confirm synteny of all sequences retrieved from applicable genomes. In Figure 1.1, we present a schematic representation of observed synteny in *tigd5*.

Next, we sought evidence of purifying selection, consistent with conserved cellular function. Evolutionary theory predicts that, in a host protein-coding sequence, mutations will occur stochastically with roughly equal probability at all positions in the sequence (neglecting variability in base composition). Natural selection will, however, decrease the probability that deleterious mutations become fixed in a population relative to the fixation probability of neutral mutations. The phenomenon of *wobble* provides a basis for the statistical evaluation of natural

selection at the molecular level. *Wobble* refers to redundancy in the genetic code, whereby most third position mutations do not alter the encoded amino acid, whereas most first position, and all second position, mutations result in an amino acid change, and are termed replacement mutations. Nucleotide changes which do not alter amino acid sequence are referred to as silent mutations. On the assumption that replacement changes are more often deleterious than beneficial, theory predicts that silent mutations, being selectively neutral, have a greater probability of reaching fixation in a population, by genetic drift, than do replacement mutations,

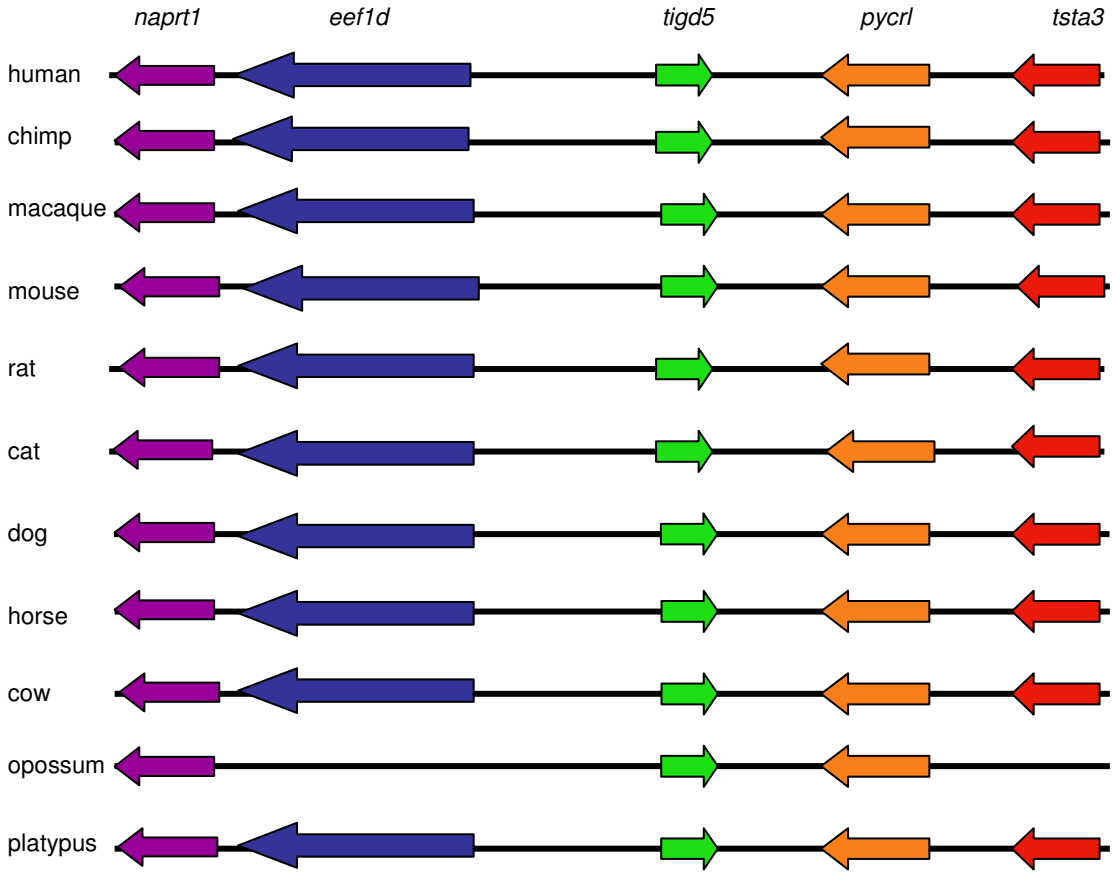


Figure 1.1 Schematic representation of synteny. Tigd5 is presented as an example. Arrows represent genes, with directionality of ORFs. Gene names appear above. The figure only depicts linear gene order and does not address variation in intergenic distances.

the fixation of which is often impeded by purifying selection. Therefore, on an evolutionary timescale, one expects mutations at silent sites to become fixed at higher rates than those at replacement sites. In comparing orthologous protein-coding sequences between species, the ratio of silent substitutions per silent site to replacement substitutions per replacement site is termed dN/dS (Ka/Ks, ω (omega)). For a neutrally evolving sequence, as a pseudogene, this ratio is expected to be approximately 1. A dN/dS ratio significantly < 1 is consistent with the action of purifying selection and is taken as evidence of a beneficial function. A dN/dS ratio significantly > 1 suggests that beneficial mutations have been driven to fixation by positive selection, the molecular mechanism of adaptive evolution (Li 1997).

We evaluated the dN/dS ratios for all sequences using the Pamilo-Bianchi-Li method, in pairwise comparisons against the human sequence. In sequences in which the ORF was disrupted by frameshift and/or nonsense mutations, we edited these positions to restore the proper reading frame. All sequences with uninterrupted reading frames exhibited dN/dS significantly < 1 and usually < 0.30. The only sequences for which the observed dN/dS was not significantly < 1 were those mentioned above which also harbored multiple frameshift and/or nonsense mutations, and we designated these sequences as pseudogenes (indicated by ψ in Figure 1.2). Those sequences which displayed dN/dS significantly < 1, but harbored apparent frameshift and/or nonsense mutations, we provisionally denoted as probable genes, on the premise that the ostensibly disabling mutations are likely to be sequencing artifacts. We concede, however, the possibility that these may represent bona fide cases of very recent pseudogenization.

1.3.2 Taxonomic Distribution and Origins

Next, we examined the phylogenetic distributions of these genes, to determine when they were exapted. The results suggest that 9 of the 10 are amniote-specific, and 8 are restricted to mammals (Figure 1.2). Of these, 4 are present as apparently functional genes in the platypus, a monotreme, placing their likely exaptation near the root of the mammalian tree.

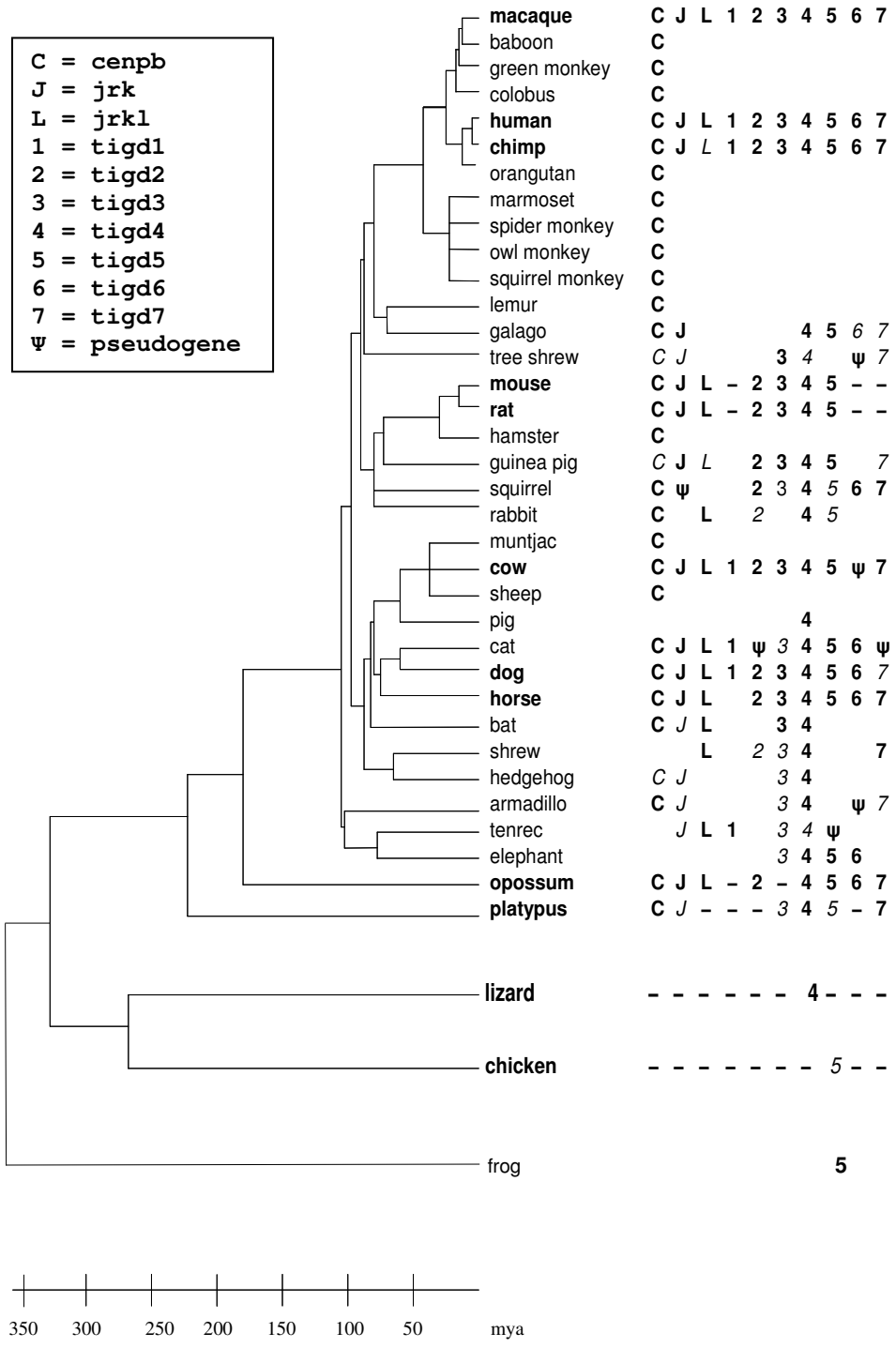


Figure 1.2 Phylogenetic distributions of 10 *pogo*-derived genes suggest recurrent domestication in mammalian evolution. 9 pogo-derived genes are amniote-specific. 8 are restricted to mammals. Italics indicate “probable” gene or pseudogene. Completely sequenced and assembled genomes are in bold face.

Three more genes (*jerky-like*, *tigd2*, and *tigd6*) appear in the opossum, suggesting domestication after the divergence of placentals from monotremes, about 230 mya, but prior to divergence of placentals and marsupials, some 180 mya (Figure 1.3). The remaining gene, *tigd1*, appears to have arisen near the time of the eutherian diversification, about 100 mya, and will be discussed in more detail, below.

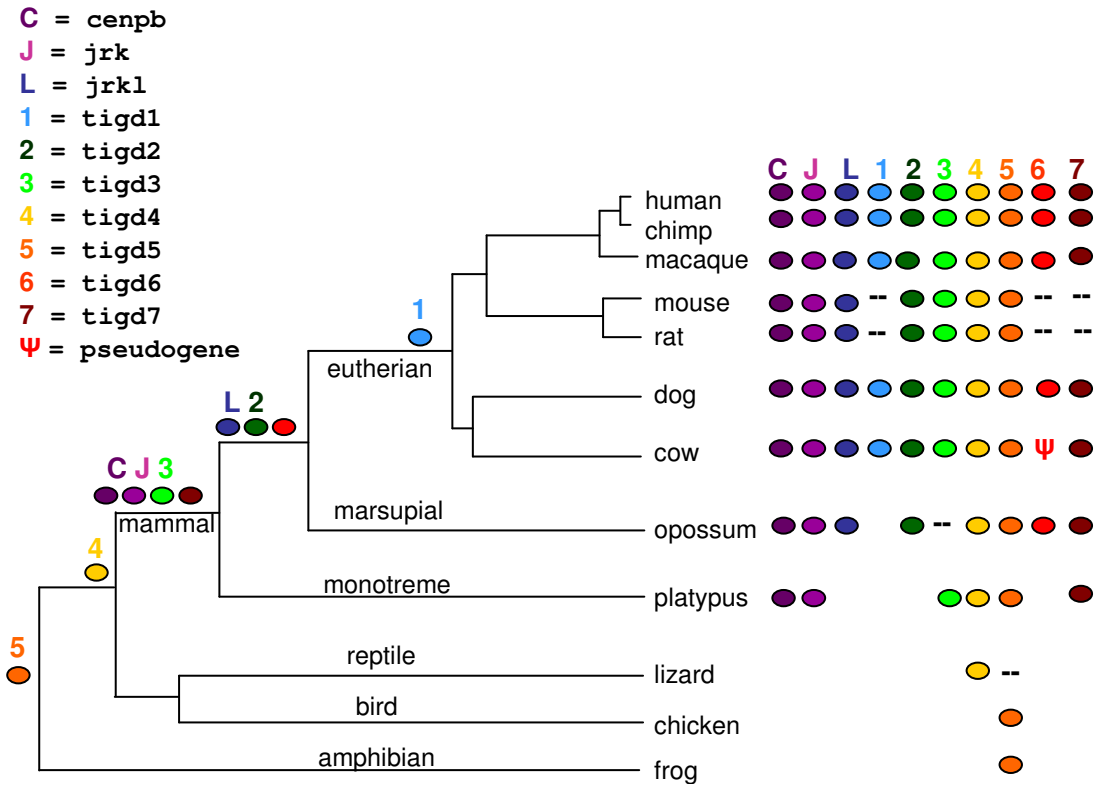


Figure 1.3 *Pogo*-derived genes arose at multiple points between the advent of tetrapods and the eutherian diversification. Ovals on the right indicate the presence of each gene within the indicated genomes. Ovals on the left indicate the apparent timeframe for the origin of each gene.

Next, we sought to determine whether these 10 genes arose by a single exaptation followed by multiple gene duplication events, or by multiple, independent exaptations of pogo transposons. First, we constructed a multiple alignment of all full-length sequences for all 10 genes. Then, we built a Neighbor –Joining tree and used it as a guide for a Maximum Likelihood

analysis, using codeml in PAML. We used the Ks branch lengths, returned by PAML, as estimators of neutral evolution, and applied a widely cited estimate of the average rate of neutral evolution in mammals (2.2 e-9 mutations / site / year) (Kumar and Subramanian 2002) to arrive at approximate coalescence times. In parallel, we considered the ratios of external to internal branch lengths, to estimate coalescence times, assuming a eutherian coalescence ~100 mya, and mammalian coalescence ~230 mya. This is a somewhat crude approach, but we merely wished to determine whether the inferred coalescence times were roughly compatible with gene duplication. Under a gene duplication hypothesis, we would expect inferred coalescence times consistent with taxonomic distributions. With the exception of a single gene pair (*jrkl* and *tigd2*), the Ks estimates produced coalescence times measured in billions of years, a full order of magnitude too great to be consistent with mammal-specific and, in some cases, therian-specific gene duplications. These results suggest, instead, that these several genes arose by independent domestication from diverse pogo lineages, on several occasions, most of them early in mammalian evolution. By far the most closely related gene pair in our dataset is the *jrkl* / *tigd2* pair, with identity ~65% at the amino acid level. The Ks estimates for the branches separating these genes suggest a coalescence time ~ 300 - 400 mya. As this is only a rough approximation, this result would seem to be compatible with either a single domestication, and subsequent gene duplication, in the ancestral therian, within the past 230 myr, or a pair of independent therian-specific domestication events of distinct *pogo* elements. It is, of course, another possibility that a single exaptation and gene duplication could have occurred prior to divergence of monotreme and therian lineages, followed by loss of both genes in the monotreme. However, we note that *jrkl* and *tigd2* are among the most highly constrained in our dataset, with treewide omega values of 0.0819 and 0.0849, respectively and with only 1 observed gene loss out of 29 observations. Loss of both genes in monotreme seems unlikely, and a therian origin for each of these genes seems more parsimonious.

Phylogenetic analysis also supports multiple origins of these various genes from distinct *pogo* families. Consistent with a previously published Bayesian analysis on a dataset which included more transposons and a small subset of genes, a Neighbor-Joining tree supports this model (Figure 1.4). There are 2 hypotheses regarding the topology of a tree composed of various animal *pogo* sequences and *pogo*-derived genes. Under a hypothesis of a single exaptation followed by multiple rounds of gene duplication, we expect the genes, being more closely related to each than to any of the more distantly homologous transposon sequences, to nest within a single clade, flanked by transposons. Under a hypothesis of multiple exaptations from distinct *pogo* lineages, we expect to observe several distinct clades, comprised of both genes and transposons. The observed topology is similar to the latter and suggests at least a half dozen independent domestication events. As noted above, the relative sequence divergence estimates among these genes suggests independent exaptations for all genes, except, possibly, either *jrkl* or *tigd2*.

1.3.3 Evaluation of Constraint

Our results indicate relatively few cases of pseudogenization (7 out of 173 mammalian sequences). We note the absence of a detectable ortholog for 3 of these genes (*tigd1*, *tigd6*, and *tigd7*) within the completely sequenced genomes of mouse and rat (Figure 1.1). Mean dN/dS values ranged from 0.1042 for *cenpb* to 0.2597 for *tigd1*. Only *tigd6* showed substantial gene loss, with 3 observed pseudogenes. For 3 of these genes (*cenpb*, *jrkl*, *tigd4*), we identified no apparent pseudogenes. We note that 2 of these, *cenpb* and *jrkl*, are among the three most highly constrained genes in our dataset, as measured by pairwise dN/dS and omega values, described below. For all full-length sequences, we performed a parallel analysis by a Maximum Likelihood method, using the codeml program in the PAML suite and obtained very similar results. In this analysis, *tigd5* exhibited the lowest omega at 0.0777, closely followed by *cenpb* at 0.0781. *Tigd1* had the highest omega, at 0.2742, and was the only gene with a value greater

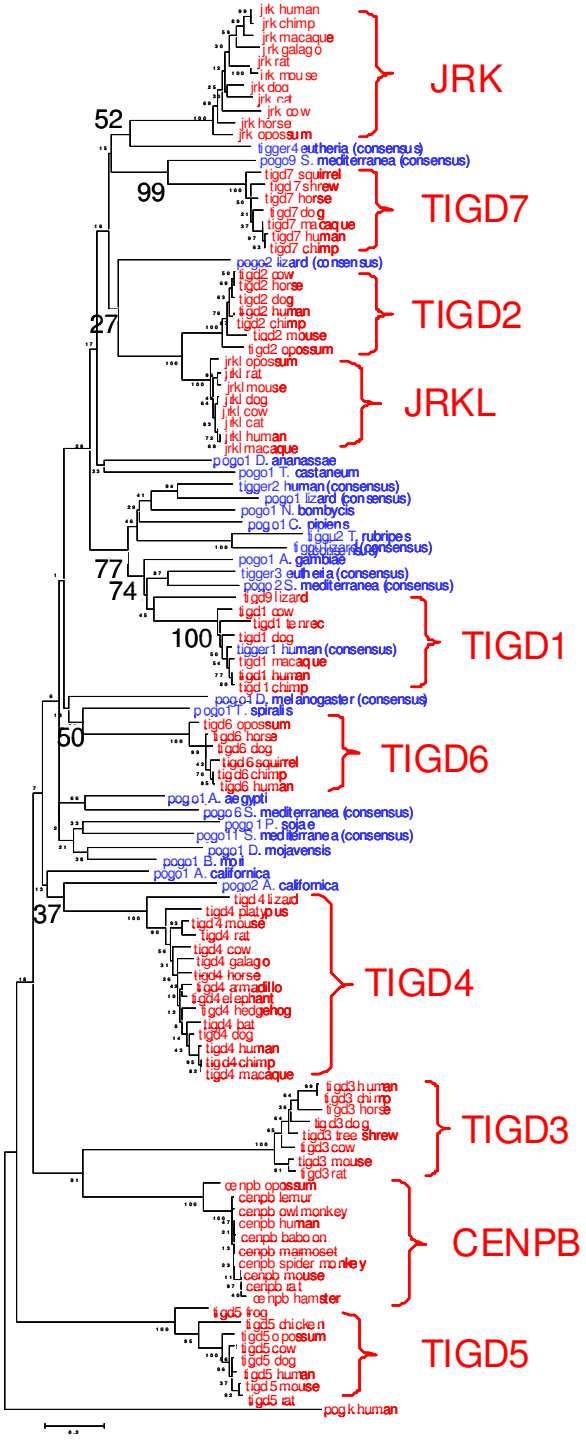


Figure 1.4 Phylogenetic analysis supports multiple independent exaptation events from diverse lineages of pogo transposons. The Neighbor-Joining tree was constructed using Mega 3.1. Genes are in red. Transposons are in blue.

than 0.2. We note that the omega values have a more narrow range than that observed for the pairwise dN/dS estimates. This is due to the inclusion of partial sequences in the latter dataset, some of which resulted in higher estimates, due to regional heterogeneity in the efficacy of selection, to be described in detail below. We, therefore, consider the omega values to be more accurate estimates of treewide parameters (Table 1).

Table 1. Treewide ω Values for Genes and Gene Regions

Gene	Entire Gene	DNA-binding	Catalytic	Acidic	Dimerization
cenpb	0.0781	0.0146	0.0425	0.1312	0.0250
jrk	0.1471	0.0127	0.1064	0.3055	0.1774
jrkl	0.0819	0.0133	0.0796	0.1474	0.0345
tigd1	0.2742	0.1075	0.2414	0.3023	0.3016
tigd2	0.0849	0.0158	0.1091	0.2316	0.0001
tigd3	0.1839	0.0038	0.1908	0.4545	0.1256
tigd4	0.1541	0.0090	0.1106	0.2126	0.2031
tigd5	0.0777	0.0002	0.0441	0.0748	0.0775
tigd6	0.1406	0.0267	0.0916	0.2813	0.2053
tigd7	0.1503	0.0504	0.1376	0.2935	0.1191

1.3.4 Regional Heterogeneity in Magnitude of Constraint

The 10 proteins in our study have been previously described as POGO-derived, by virtue of significant similarity to POGO transposases. Specifically, the highest similarity is observed in the N-terminal DNA-binding domain and in the adjacent catalytic domain (a DDE endonuclease). Over this expanse of approximately 340 amino acids, *Homo sapiens* CENPB is nearly 30% identical and nearly 50% similar to the active POGO transposase in *Drosophila melanogaster*. Additionally, *Homo sapiens* CENPB is approximately 24% identical and 45% similar to the translated consensus sequence of the Homo sapiens *pogo*-like family, *Tigger1*, over the homologous region. In active *pogo* lineages, the DNA-binding domain exhibits affinity for a sequence which occurs within the terminal inverted repeats (TIRs) of the transposon. Previous studies have dissected the tertiary structure of CENPB and elucidated the binding specificity of the N-terminal domain, to the level of resolving the individual amino acid residue-nucleotide interactions which participate in binding of the protein to its target sequence, the so-

called CENPB box, within the alpha satellite repeats of the mammalian centromere (Tanaka et al. 2001). At the C-terminus of CENPB lies a dimerization domain, which has also been thoroughly characterized, and which may function in concatemerization of CENPB at the centromere (Tawaramoto et al. 2003). In active *pogo* lineages, homo-dimerization, mediated by C-terminal domains, may be necessary for transposition. The amino acid sequences of *H. sapiens* CENPB, *H. sapiens* TIGGER1, and *D. melanogaster* POGO do not, however, exhibit significant similarity over this domain. This observation is consistent with the theoretical prediction that as transposon lineages diversify, the facility of a transposase to homo-dimerize while eschewing dimerization with closely related transposases will increase its efficiency of transposition, thereby increasing its abundance, resulting in an elevated rate of evolution on this part of the sequence. Intervening between the catalytic and dimerization domains of CENPB lies an expanse of about 72 amino acids showing marked enrichment for acidic residues (D and E, aspartic acid and glutamic acid, respectively.) More than 73% of residues in this region are acidic. The *D. melanogaster* POGO and *H. sapiens* TIGGER1 also show enrichment for acidity in this region, albeit less dramatically (*D. melanogaster* POGO: 35% acidic over 69 residues, *H. sapiens* TIGGER1: 36% acidic over 41 residues). Excluding this acidic run, only about 10% of residues are acidic in each of these sequences. We note that the remaining POGO-derived proteins display a comparable level of enrichment for acidity in this region, with the exception of JERKY, for which this region is only ~ 14% acidic.

In order to investigate the domain architecture of the remaining proteins in our dataset, we utilized the Conserved Domains (cds) database on the NCBI website. We used each of the 10 queries against the NCBI conserved domain architecture database. For each of the queries, the software predicted an N-terminal CENPB-like DNA-binding domain adjacent to a DDE-type endonuclease domain.

To facilitate investigation of possible heterogeneity in the magnitude of selective constraint among domains, we produced a multiple alignment of the 10 amino acid sequences

from human and used *H. sapiens* CENPB as a guide in partitioning the sequences corresponding to the domain architecture of CENPB (Figure 1.5). Thus, we were able to divide the corresponding nucleotide sequences for all genes and all species for which we possessed full-length sequences into four partitions, corresponding to the putative DNA-binding, catalytic, acidic, and dimerization domains of the encoded proteins. In theoretical terms, exaptation refers

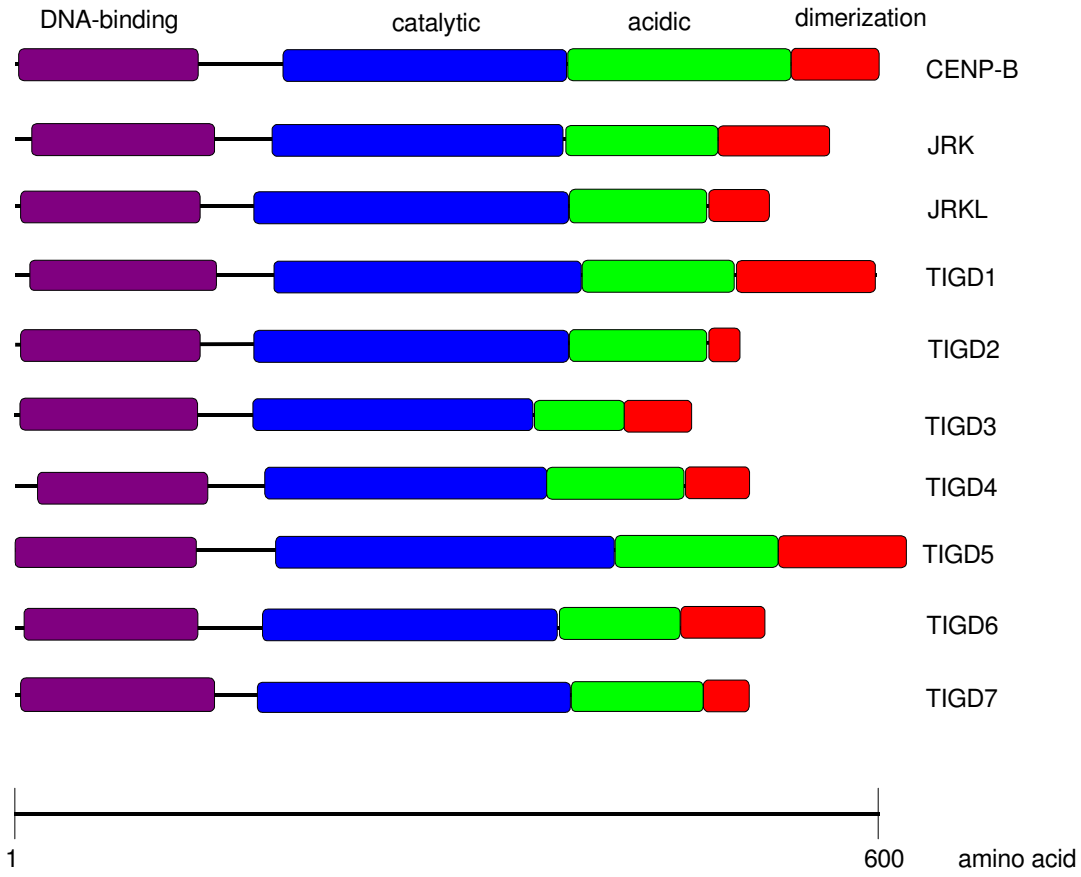


Figure 1.5 Ten pogo-derived proteins exhibit similar domain architecture.
Amino acid positions based on human sequences as annotated in the UCSC Genome Browser.
Domain annotation of Hs CENPB was used as a guide.

to the adaptation of an existing structure to a new function. Theory predicts that transposase exaptation should involve the cellular recruitment of an extant transposase mechanism to a novel function, beneficial to the host organism, bringing a formerly neutrally evolving sequence under the influence of natural selection. The major functions of transposases are, in general,

site-specific DNA-binding, DNA cleavage, and, in some cases, protein-protein interactions (especially homo-dimerization) (Craig et al. 2002). Empirical evidence from previous studies suggests that exaptation for DNA-binding activity is a common theme in molecular domestication. Indeed, the 2 proteins in our dataset which have been functionally characterized (CENPB and JRK) are known to bind DNA. We, therefore, hypothesized that these additional POGO transposases were exapted for this same proclivity and that evidence of such should be detectable in heterogeneity in the magnitude of selective constraint acting on these various domains. Specifically, we should expect to see particularly stringent constraint on the DNA-binding domain.

Using the codeml program in PAML, we estimated tree-wide omega values for each of the 4 partitions within each gene. We then compared the omega values among partitions and tested the observed differences for significance, as described in Methods. We note that all regions of all 10 genes exhibit omega values significantly < 1 (LRT, all values of $p < 0.0001$), consistent with purifying selection acting on all regions in each of these genes. In addition, we observed highly consistent patterns of regional heterogeneity in omega (Figure 1.6) (Table 1). Omega for the region encoding the DNA-binding domain ranged from 0.0002 (*tigd5*) to 0.1075 (*tigd1*) and, in every case, was significantly lower than that for the region encoding the DDE domain. With a single exception (*tigd1*) the DNA-binding domain showed significantly lower omega than the acidic run. With 2 exceptions (*tigd2* and *tigd7*), omega for the region encoding the DNA-binding domain was significantly lower than for the region encoding the dimerization domain. The observation of strict constraint on the region encoding the DNA-binding domain in each of these genes is consistent with conservation of sequence-specific DNA-binding activity in the encoded proteins. We note that for 3 of these genes (*cenpb*, *jrkl*, and *tigd2*), omega for the region encoding the dimerization domain is significantly lower than for the regions encoding both the DDE and acidic domains. Each of these genes has an omega value < 0.05 over this region, and in the case of *tigd2*, omega over this region (0.0001) is significantly lower even than

omega for the region encoding the DNA-binding domain (0.0158). We note that a high level of constraint on the dimerization domain of CENPB is consistent with its observed homo-dimerization in vivo. The observation of a similarly stringent level of constraint acting on the 3' region of *jrkl* and *tigd2* suggests that the corresponding protein domain may mediate

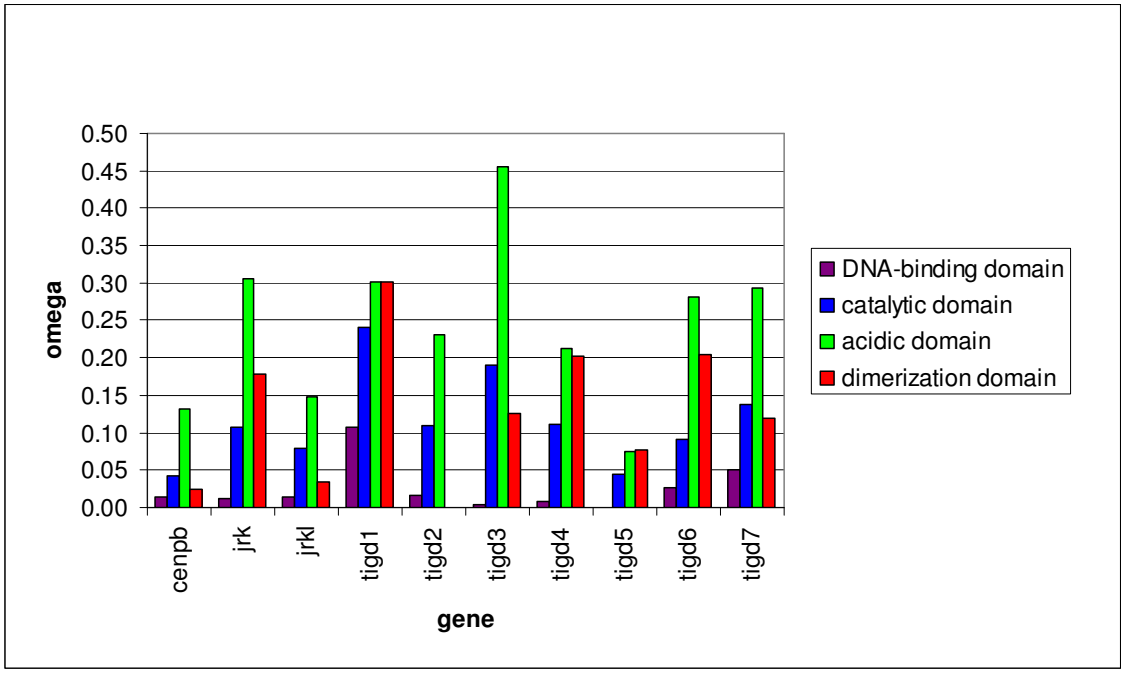


Figure 1.6 Nucleotide sequence encoding the DNA-binding domain is highly constrained. All values were estimated using site models in the codeml program in the PAML suite.

homo-dimerization activity, as well. The observation of higher omega values, but still significantly < 1, on the region encoding the acidic domain, for most of these genes, suggests that selection may act to preserve an overall level of acidity in this region, with relaxed constraint on non-acidic residues, and perhaps even relaxed constraint on some individual acidic residues. Some authors have suggested that acidic amino acids may enhance the activity of DNA-binding proteins by interacting with basic residue on histones, thereby relaxing local

chromatin structure, facilitating binding of the protein to its target DNA. Alternatively, these acidic regions may provide an interface for interaction with other host proteins.

1.3.5 Motif Conservation and Evidence of Functional Diversity

As further evidence that each of these proteins functions in DNA-binding, we note the presence of a highly conserved motif located near the C-terminal end of the N-terminal domain (Figure 1.7). The motif S_GWλ_RΩ_R, in which _ is variable, λ indicates L or F, and Ω is W or F is present in every sequence we observed, for all 10 proteins. In CENP-B, this sequence comprises 1 of 8 α-helices involved in DNA-binding. It contains 3 of the 7 residues known to participate directly in specific base recognition, 2 of which are perfectly conserved in all sequences observed (in bold above). This same motif is perfectly conserved in 2 centromere-binding proteins in the fission yeast Schizosaccharomyces pombe, and conserved with only 1 replacement in another. All 3 share homology to transposases encoded by fungal pogo transposons, and exhibit apparent partial functional redundancy by binding distinct centromeric sequences (Halverson et al. 1997; Lee et al. 1997; Ngan and Clarke 1997; Irelan et al. 2001; Nakagawa et al. 2002). In addition, this motif is present in the H. sapiens TIGGER1 consensus and partially conserved (with 2 replacements) in the transposase of an active pogo in D. melanogaster (Figure 1.8).

We also note, however, considerable diversity among these 10 proteins with regard to the DNA-binding domain, with pairwise amino acid identity over this domain ranging from ~24 - 40%, among the human amino acid sequences. This is a bit higher than the 18-28% observed identity over the entire sequences and somewhat lower than the ~45-50% identity exhibited over the DNA-binding domains among the centromeric trio in yeast. In addition, we note that each of these proteins exhibits at least 2 perfectly conserved amino acid changes, versus CENP-B, among the 7 positions involved in base recognition (in CENP-B). In each of these proteins, except JRK, we observe at least 1 non-conservative amino acid replacement, present in all species, among these 7 residues. As an example, we note a residue within the α-helical

motif given above, which is involved in specific base recognition and which is highly variable among these proteins but exhibits perfect conservation within each protein, over all species

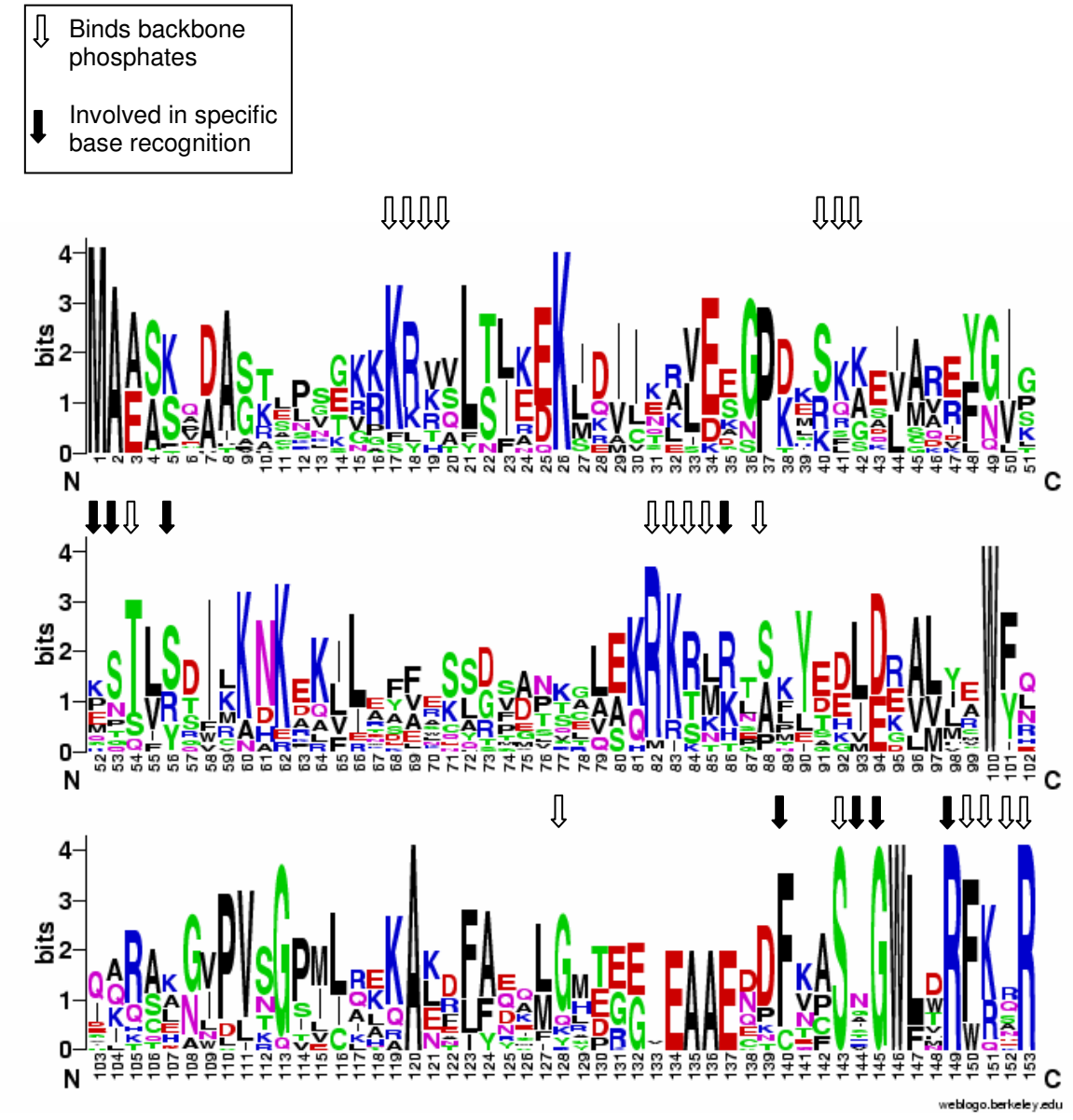


Figure 1.7 Comparison of DNA-binding domains in 10 POGO-derived proteins reveals evidence of conserved DNA-binding activity and possible diversity in binding site specificity. Logo was produced from multiple alignment of all full-length sequences for all 10 pogo-derived genes. Height of each character is proportional to the representation of that base at that position in then alignment. Open arrows indicate residues which bind DNA at backbone phosphates. Solid arrows indicate residues which recognize specific bases.

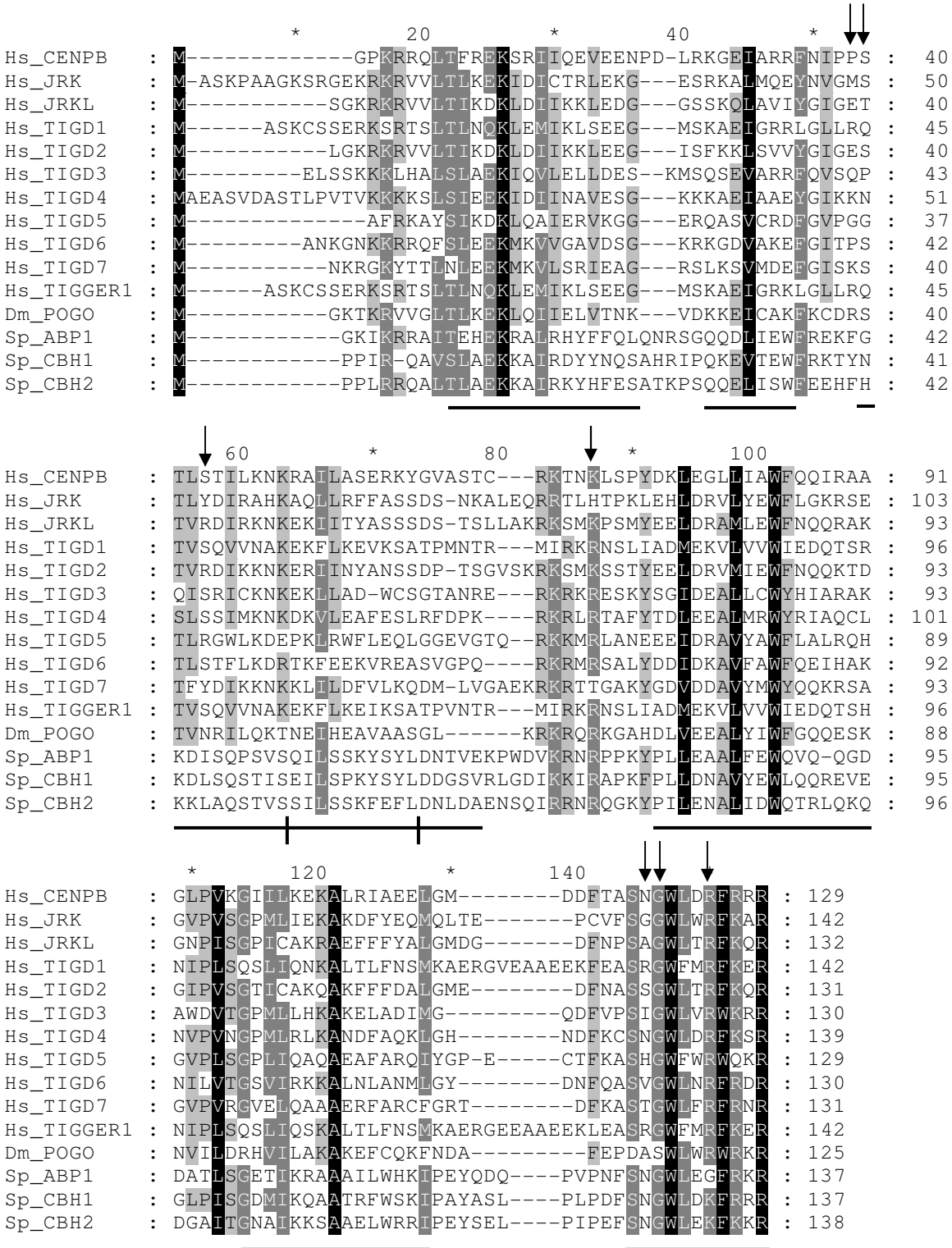


Figure 1.8 N-terminal DNA-binding domain from pogo transposases and pogo-derived cellular proteins. Alpha helices involved in DNA recognition are underlined. Arrows indicate residues involved in specific base recognition.

observed. We are hesitant to make predictions about binding affinities for these various proteins, in the absence of crystal structures, and based solely on alignment of primary sequences against CENP-B. However, we note that while our data are consistent with conservation of site-specific DNA-binding activity for each of these proteins, we do not observe a level of identity which we would consider compelling evidence that any of these proteins shares a binding site with CENPB.

1.3.6 Orthology in Tigd1

Our interrogation of the genomic databases with the translated *H. sapiens tigd1* returned apparent full-length orthologs in chimp, macaque, dog, cow, and tenrec, plus a partial sequence in cat. All sequences exhibited pairwise dN/dS (vs. human) ranging from 0.2424 to 0.3153, and omega values between 0.19 and 0.29 (except human, 0.35). All values were significantly < 1. The tree-wide omega value was 0.2742 (0.1075 over the region encoding the DNA-binding domain). Amino acid identities to the human query ranged from 85% to 99%.

However, interrogation of the UCSC Genome Browser, using BLAT, revealed that the locus at which this sequence resides in the carnivores (chr36:29,140,383-29,142,158 in dog and scaffold_16027:709-807in cat) is not syntenic to the position at which *tigd1* resides in primates (chr2:233,121,023-233,123,470 in human, chr12:96,440,859-96,441,072 in macaque, and chr2b:238,748,241-238,750,687 in chimp, all of which exhibit synteny). Furthermore, the locus at which the sequence resides in cow is not syntenic with either of these loci (Figure 1.9). Now, *tigd1* is also unique among the genes in our dataset in that it appears to be restricted to eutherians, making it at least 80 myr younger than any of the other known *pogo*-derived genes. This relative youth allows for identification, not only of the conserved coding sequence, but of the flanking which harbors the remnants of TIRs. For all the other *pogo* –derived genes, these flanking sequences have been neutrally evolving too long, and are consequently too degenerate, to be identified. Because *tigd1* was domesticated from a younger, less divergent, transposon family, researchers have been able to unambiguous identify the *pogo* family from

which *tigd1* was domesticated. In fact, the *tigd1* in human resides within an annotated *Tigger1* transposon, as observed in the Genome Browser. Likewise, the putative *tigd1* orthologs in dog, cat, and cow lie within annotated *Tigger1* transposons. This fact, as it happens, adds another problem to the question of orthology in *tigd1* in that no *Tigger1* in the human genome has yet been observed to occupy an orthologous locus in any non-primate species. This amounts to a lack of evidence for vertical transmission of *Tigger1* in eutherians (above the ordinal level).

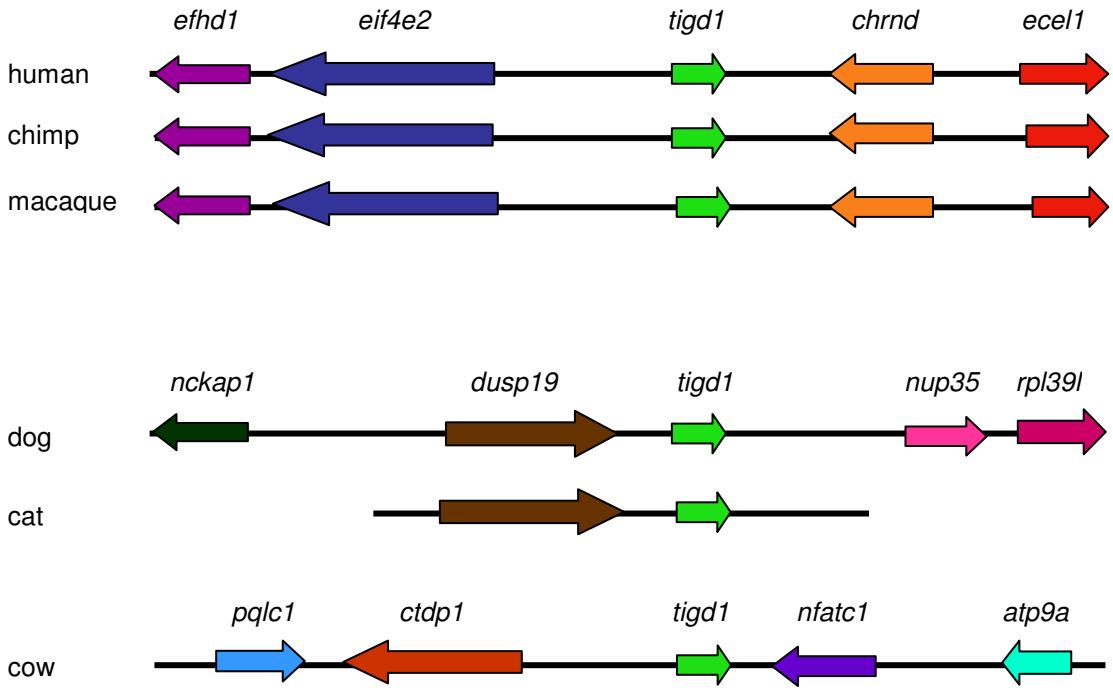


Figure 1.9 Lack of synteny in *tigd1*. Arrows represent genes, with directionality of ORFs. Gene names appear above. The figure only depicts linear gene order and does not address variation in intergenic distances.

Thus, we are confronted with the possibility that the *tigd1* observed in primates is not the same gene observed in carnivores, nor in cow. Thus, the data present a pair of mutually exclusive hypotheses. Under the first hypothesis, no *Tigger1* elements were present in the common ancestor of eutherians, but colonized these genomes independently some time after ordinal diversification, perhaps by multiple horizontal transfers. In this case, *tigd1* must have

arisen independently in at least 4 different lineages, and from distinct populations of *Tigger1* transposons. Under the second hypothesis, *Tigger1* was present in the ancestral eutherian, probably at low copy number, with most of its activity occurring after the ordinal diversification of eutherians, resulting in a dearth of observed orthologous insertions. Under this hypothesis, it is plausible that *tigd1* was domesticated once, in the common ancestor, as a still active transposon, which subsequently transposed to a novel locus in at least 3 lineages. Alternatively, under this hypothesis, the *tigd1* genes in these various lineages could have arisen independently from distinct, but closely related *Tigger1* transposons after the eutherian diversification.

We have applied a phylogenetic approach to the comparison of these hypotheses. For this analysis, we selected those copies of *Tigger1* in human, which exhibit the greatest similarity to human *tigd1*, those copies of *Tigger1* in dog showing the greatest similarity to *tigd1* in dog, and likewise for cow. We also included the copies of *Tigger1* within which the *tigd1* genes reside, in each of these species, plus chimp and macaque. We then aligned the sequences and constructed unrooted Neighbor-Joining phylogenies. Under a horizontal transfer / independent domestication hypothesis, according to which the *tigd1* genes arose from distinct *Tigger1* populations, we should expect the primate *tigd1*'s to form a clade with the human *Tigger1* transposons, the dog gene to group with the dog transposons, and the cow genes to group with the cow transposons. Under a single domestication / vertical transmission hypothesis, we should expect the genes to form a single clade.

Wishing to minimize the tendency of sequences under purifying selection to cluster together amid neutrally evolving sequences, we constructed 2 trees. For the first, we used only 3rd codon positions from the gene / transposase region of the alignment (Figure 1.10). In the second, we used only non-coding sequence flanking the gene / transposase region of the alignment (Figure 1.11). (The edited alignments were comprised of 590 and 610 nucleotides, respectively.) The resulting trees exhibited similarly ambiguous results. Neither the dog nor cow

tigd1 genes form a clade with the primate *tigd1*'s, thereby presenting no direct evidence of vertical inheritance of *tigd1* above the ordinal level. On the other hand, the very short, unsupported internal branches and general lack of lineage-specific clustering of transposons

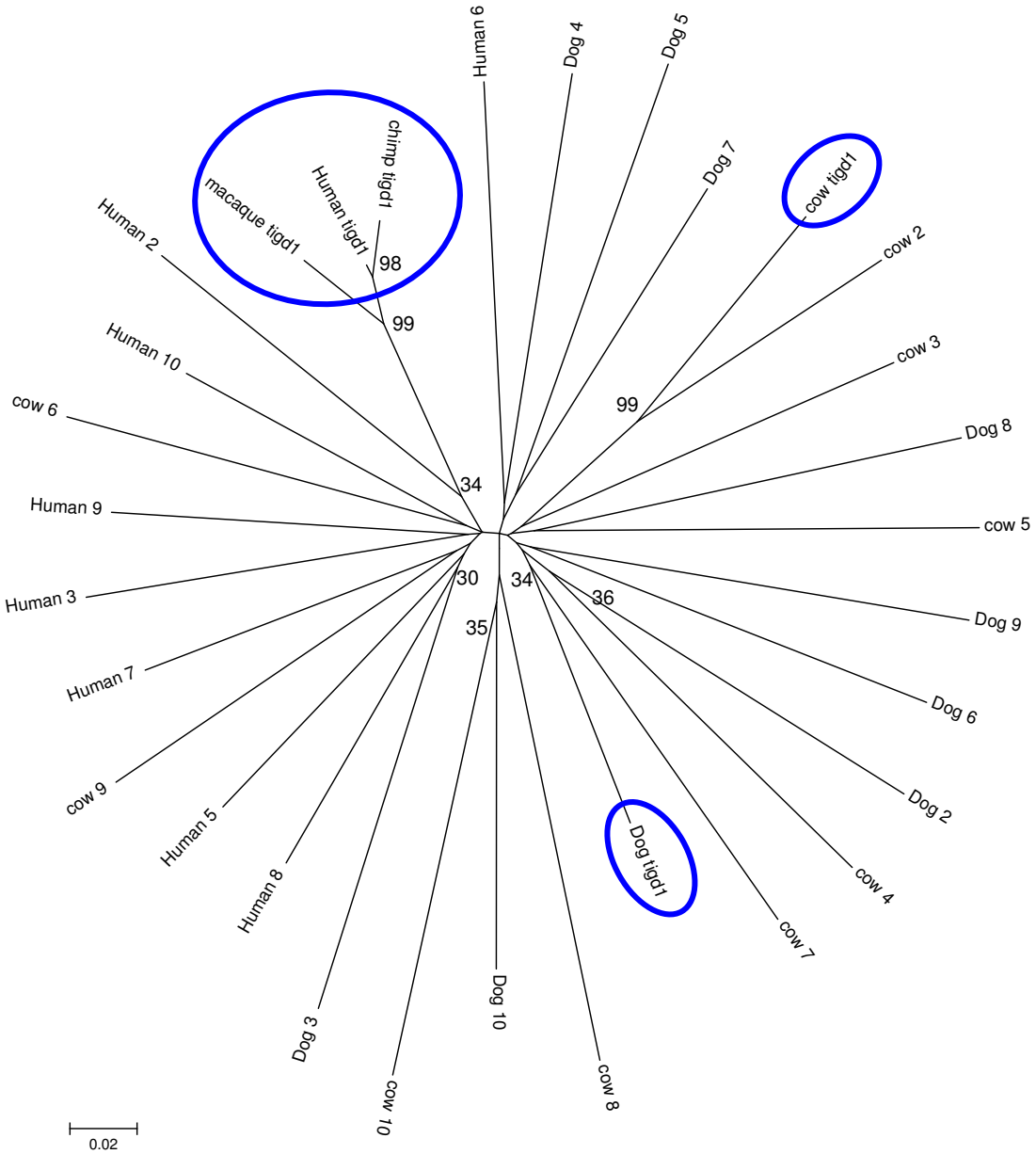


Figure 1.10 Neighbor-Joining tree of *Tigger1*, using 3rd codon positions. Sequences containing *tigd1* genes are circled. Bootstrap values < 30 are omitted.

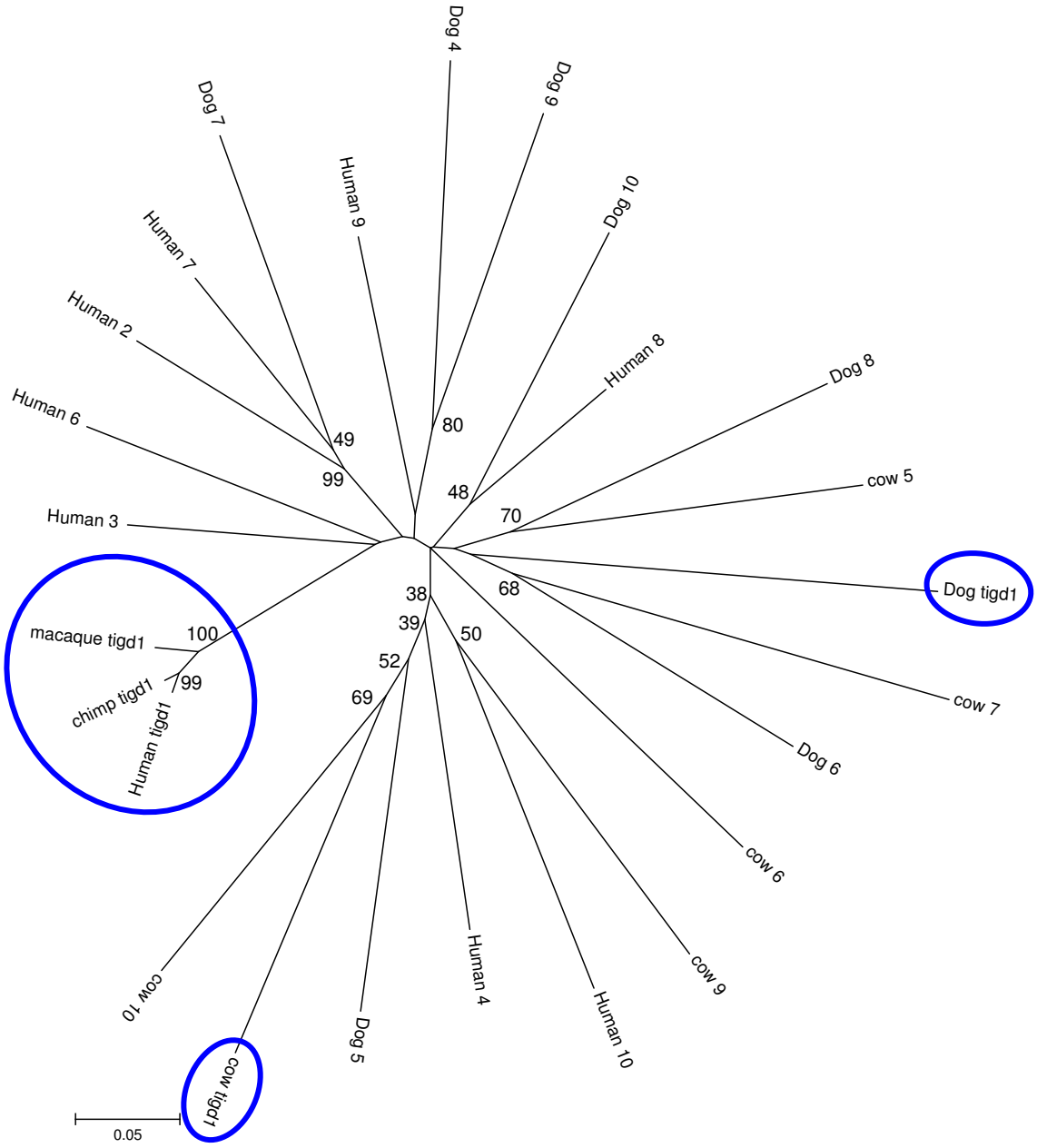


Figure 1.11 Neighbor-Joining tree of *Tigger1* non-coding sequence. Sequences containing *tigd1* genes are circled. Bootstrap values < 30 are omitted.

provide no direct evidence that these various eutherian orders were invaded by diverse populations of *Tigger1*. The trees would seem, rather to support the proposition that the *tigd1* genes observed in the various eutherian lineages were derived from highly similar, if not

identical *Tigger1* transposons. Pending the results of a more sophisticated phylogenetic analysis, the origin of *tigd1* remains unresolved.

We note, however, that despite the lack of observed inter-ordinal orthology in *Tigger1*, we have recently discovered evidence that the very copy of *Tigger1* which gave rise to the *tigd1* these lineages. It has been documented that when a transposon is excised by its cognate transposase, it may leave a "footprint." This may include a target site duplication (TSD), produced at the time the transposon was inserted into the locus. *Tc1/mariner* transposons are known to produce a TA target site duplication, and are thus typically flanked by a pair of TA's. Additionally, transposon excision sometimes leaves behind a small fragment of a TIR, comprised of a few terminal nucleotides (Craig et al. 2002). Aided by the Phastcons vertebrate conservation track on the UCSC Genome Browser, we have identified a putative footprint in the genomic sequences of cat and horse (Figure 1.12). The sequence is TACTGTA, in both species. The CTG match the last three nucleotides of the TIR from *Tigger1*. This triad is flanked by the TA…TA, characteristic of the *Tc1/mariner* TSD. The probability of a 5-nucleotide motif (the 7 above, less the TA residing at the locus prior to insertion) occurring by chance, while not prohibitively small (~ 0.001), makes it appear reasonably likely that this is a bona fide excision scar from an ancient *Tigger1* excision. This observation places *Tigger1* in the genome of the ancestral eutherian prior to ordinal diversification, rendering a horizontal transfer/ multiple infestation hypothesis unparsimonious and making the hypothesis of a single *tigd1* domestication, and subsequent transposition, a possibility. Alternatively, distinct *Tigger1* copies, already present in the ancestral eutherian, or spawned by transposons that were, may have been domesticated independently in each of these lineages, following the eutherian diversification.

1.4 Discussion

Our analysis of 10 exapted *pogo* transposons in the human genome suggests that all 10 are restricted to tetrapods, and that 8 arose early in mammalian evolution. Our data further

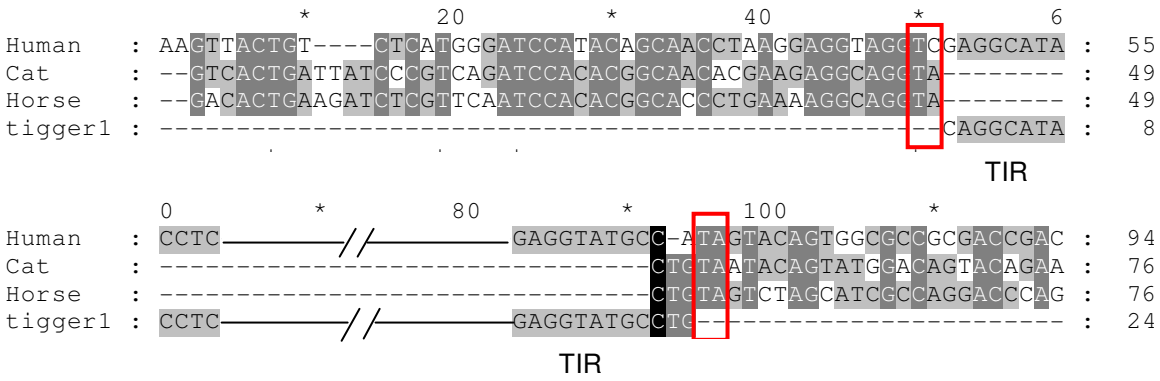


Figure 1.12 Putative *Tigger1* excision footprint at site orthologous to human *tigd1*. Partial TIR's of *Tigger1* are shown. Boxes indicate apparent TSD. CTG fragment between TSD in cat and horse is consistent with imprecise repair following transposon excision.

suggest that, with the possible exception of the *jerky-like* / *tigd2* pair, gene duplication does not appear to have occurred, and that, rather, these genes arose by multiple, independent domestication events. Our results show that several of these genes have been very highly constrained in mammalian evolution. Furthermore, we have presented evidence that all 10 genes are likely to encode site-specific DNA-binding proteins. It seems likely that the observed stringency of constraint on the DNA-binding domain of these POGO-derived proteins reflects constraint on binding affinity to a similarly conserved, but as yet unknown (except for CENPB), set of binding sites within the mammalian genome. In the case of CENPB, the binding site is known to be a highly constrained sequence occurring within the alpha satellite repeats of mammalian centromeres. The occurrence of a trio of partially redundant POGO-derived centromeric proteins in yeast, together with the observation of mild phenotype in *cenpb* knockout mice has lead to speculation about the presence of mammalian proteins with centromeric activity functionally redundant to that of CENP-B. We are wary of drawing broad comparisons between CENP-B in mammals and the centromeric proteins in *S. pombe*. However, the observation that, in yeast, a lower overall level of variation and a comparable level of diversity at residues likely to participate directly in specific base recognition allows these proteins to differentiate among distinct targets makes it seem unlikely that any of these POGO-

derived mammalian proteins binds the same site as CENP-B (the CENP-B box). It remains possible that any of these proteins might bind a distinct centromeric sequence, as is the case in yeast. However, if such a sequence is being bound by a protein which has been highly conserved throughout mammalian (or at least eutherian) evolution, we should expect this sequence to exhibit broad conservation, as the CENPB box does. To date, no such sequence has been detected, and apart from the CENPB box, the mammalian centromere is rapidly evolving, with little sequence similarity among taxa (Henikoff et al. 2001). These observations cast doubt on the prospect of a redundant centromeric role for one or more of these POGO-derived proteins. Future work will focus on confirming the DNA-binding activity of these proteins in vivo and elucidating the genomic loci at which they bind.

We have presented evidence that the *tigd1* gene in primates has homologs in cat and dog, cow, and tenrec. These sequences enigmatically reside at distinct loci among orders making it difficult to determine whether their homology is to an ancestral cellular gene or, rather, to a transposon. At present, we cannot distinguish between 2 hypotheses regarding the origin(s) of *tigd1*. The data seem to be compatible with either a single exaptation followed by several lineage-specific transposition events, making *tigd1,* for some time, a true "jumping gene," to resurrect McClintock's early description, or with a hypothesis that these *tigd1* genes were domesticated from distinct, but closely related *Tigger1* transposons independently in at least 4 eutherian lineages, probably following a single genomic infiltration prior to the eutherian diversification.

We also speculate about the possibility of a hybrid model, as follows. Let us suppose that, just prior to the eutherian diversification, the ancestral genome was inhabited by a small population of *Tigger1* transposons, none of them fixed. A single *Tigger1* element transposed into a locus at which binding by transposase conferred some beneficial cellular function. This exapted binding site then became subject to natural selection. All transcriptionally competent *Tigger1*'s would have been likewise exapted, their encoded proteins potentially executing both

selfish and beneficial functions. These individual protein sources, however, would have continued to evolve neutrally, so long as multiple copies were present. Over time, some protein sources would have been lost by mutation as new sources arose by ongoing transposition. Suppose then that the amplification of *Tigger1* straddled the eutherian diversification, with most copies being generated in a lineage-specific manner and, most, if not all, of the relatively few copies that existed prior to the diversification being subsequently lost (the fixation probability of a neutral allele being equal to its frequency divided by twice the effective population of its host), leaving little or no evidence of orthology. Over time, as individual transposons accumulated mutations in a stochastic manner, the population of exapted genes, encoding proteins to bind the exapted site would dwindle, until at some point, only 1 remained. This last suitable copy (the last one in an appropriate genomic context, with the right expression profile befitting its cellular function) only then would come under selection. Under such a model, even if we observed an abundance of *Tigger1* copies at orthologous loci among these eutherian orders, we should not be surprised if this last functional copy, this *tigd1*, should, by chance, reside at distinct loci among these genomes.

We note that such a model could explain not only the lack of synteny among *tigd1* genes, but their failure to cluster phylogenetically, while having an obvious parsimony advantage in requiring only a single exaptation event. Ongoing work will focus on testing these hypotheses, by more sophisticated phylogenetic methods, and will seek clarification regarding the origin(s) and evolutionary history of *tigd1*.

Together, our observations of recurrent domestication of *pogo* transposons early in mammalian evolution raise the question as to whether *pogo* transposons are uniquely suited to potential exaptation. It is tempting to speculate that, in addition to the features common to Class 2 transposons, the region enriched for acidity in *pogo* transposons, and retained, in varying degrees, in 9 of these 10 exapted proteins, may enhance their ability to interact with DNA and/or cellular proteins, or to do so in a broader range of chromatin contexts, thereby

augmenting their potential for exploitation by the host. However, mounting evidence suggests that transposon exaptation, exotic though it may seem, is far from anecdotal and has likely been a recurrent and significant source of genomic innovation throughout eukaryotic evolution. The 47 currently identified cases of exapted transposons residing in the human genome may be just the tip of the iceberg. As our knowledge of the extent of transposon exaptation in various taxa grows, we will have a clearer idea as to the degree to which *pogo* transposons may or may not have been highly favored by natural selection in the molecular struggle to exist.

CHAPTER 2

PURIFYING SELECTION ON MARINER BINDING SITES IN ANTHROPOID EVOLUTION

2.1 Introduction

In contrast to the very high copy numbers of some Class I transposons in the human genome, notably the Long Interspersed Nuclear Element *L1* and the Short Interspersed Nuclear Element *Alu*, which together comprise a third of our DNA, Class II transposons exhibit much lower copy numbers in humans (Craig et al 2002). One of the most abundant Class II family in humans, with a copy number of ~8000, is *made1*, an 80-bp deletion derivative of *Hsmar1*, a member of the mariner superfamily, which is widespread among eukaryotes. There are ~200 full-length copies of *Hsmar1* in the human genome (Robertson and Zumpano 1997). *Made1* is an example of a miniature inverted-repeat transposable element (MITE). MITEs do not encode transposase, but can be mobilized in trans by the transposase encoded by the full length elements from which they are derived (Hartl et al 1992). *Hsmar1* and *made1* are found in anthropoid primates, that is, the platyrrhines (new world monkeys) and catarrhines, which are comprised of cercopithecoids (old world monkeys) and hominoids (apes). They are not observed in prosimians, which include tarsier, galago, and lemurs. This distribution places their colonization of the primate genome between ~58 mya, when prosimians and anthropoids diverged, and ~40 mya, when platyrrhines and catarrhines diverged (Robertson and Zumpano 1997). Molecular dating of these elements also places their origin within this time frame (Robertson and Zumpano 1997; Pace and Feschotte, 2007).

Like all Class 2 elements found in the human genome, *made1* and *Hsmar1* have been inactive in the human lineage for at least the last 40 myr (Lander et al 2001, Pace and Feschotte 2007). At some point during the interval between 58 and 40 mya, a single copy of *Hsmar1* transposed into a locus just downstream of a pre-existing gene encoding a SET histone

methyltransferase (Robertson and Zumpano 1997; Cordaux et al 2006). Histone methylation is involved in maintenance of chromatin structure and modulation of gene expression. *Set* is a single copy gene, which is highly conserved throughout vertebrates (Cordaux et al 2006). In anthropoids, *set* is fused, in frame, to the *Hsmar1* transposase gene. The primary transcript encodes a hybrid protein, called SETMAR. The precise function of SETMAR is unknown, but analysis of the molecular evolution of *setmar* has demonstrated that the regions encoding both the SET and MAR domains have evolved under purifying selection (Cordaux et al 2006). Furthermore, in vitro analyses have shown that SETMAR retains both histone methyltransferase activity (Lee et al 2005) and DNA-binding activity, and that the latter is accomplished by the N-terminal region of the MAR domain, derived from the *Hsmar1* transposase gene (Cordaux et al. 2006). The exaptation of SETMAR is an example of a recurrent theme in the molecular domestication of Class II transposons, namely the cellular co-opting of transposases as novel host DNA-binding proteins (reviewed by Feschotte and Pritham 2007). The binding sites for such a protein are likely to be descended from the sites bound by the ancestral transposase, and located within the terminal inverted repeats (TIRs) of the transposon. Indeed, SETMAR has been shown to bind, in vitro, a 19-bp motif occurring within the consensus sequences of *made1* and *Hsmar1* TIRs, referred to as *mariner binding site*, or *mbs* (Cordaux et al 2006).

In vitro analysis also suggests that there are 2 positions at either of which a single nucleotide substitution inhibits binding. At the remaining positions, a single substitution does not appear to inhibit binding. It appears that paired substitutions within the sequence inhibit binding, although not all combinations have been tested. Additionally, there are 2 adjacent positions within the sequence, which do not appear to participate in interaction with the protein, as both positions may be substituted simultaneously with no observable effect on binding (Figure 2.1).

We have hypothesized that SETMAR binds the same 19-bp motif in vivo, and that such binding, at least at some loci, confers a beneficial effect on phenotype, bringing these binding sites under selective constraint. Our objectives in the study described herein were twofold. First,

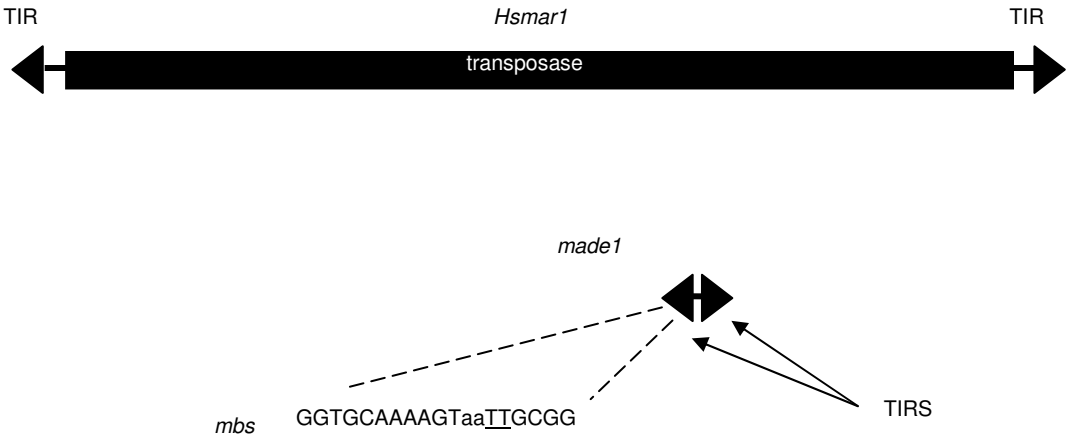


Figure 2.1 Anatomy of *Hsmar1* and *made1* transposons, and *mariner binding sites*. The nucleotides in lower case do not affect binding in vitro, and are not considered in assessment of *mbs* conservation. Binding is inhibited by a single substitution at either underlined position.

we sought evidence of purifying selection acting on some of these sequences. This task was complicated by the fact that the human genome harbors approximately 1200 *mbs*, with fewer than 2 substitutions, mostly within the TIRs of *made1* transposons, representing a large pool of potential binding sites for SETMAR in vivo (Cordaux et al 2006). Additionally, *made1* and *Hsmar1* transposons in the human genome exhibit a mean divergence of only ~8% from their consensus sequences (excluding hypermutable CG dinucleotides, with an average divergence of ~40%) (Robertson and Zumpano 1997). Therefore, we expect most copies of this motif to have just 2 or 3 substitutions, under neutral evolution, and we would not be surprised to find more than a thousand copies with fewer than 2 substitutions, simply by chance. It seems quite possible that many, perhaps even most, of these sequences are, in fact, neutrally evolving. Our challenge was to detect a (possibly faint) signal of selection under the white noise of genetic drift. Second, we wished to identify individual loci which show some evidence of purifying selection, while eliminating other loci, in order to reduce to a manageable number the list of candidate loci for in vivo analysis of binding by SETMAR, and to facilitate a search for any patterns that might begin to emerge from the genomic distribution of such loci.

In pursuit of these dual objectives, we applied the technique of phylogenetic shadowing (Boffelli et al. 2003). This technique is based on the assumption that the observation of a higher level of identity than would be expected by chance in a sequence residing at orthologous loci in multiple genomes is evidence that the sequence has been conserved in these lineages by the action of purifying selection to maintain some beneficial function. In our analysis of *mbs*, we applied phylogenetic shadowing in a sequential screening process, combining both in silico and experimental interrogations of several anthropoid genomes. We began with *mbs* identified in the human genome, then evaluated the conservation of sequence at orthologous loci, first in an old world monkey, then in a new world monkey, then in several additional anthropoids. At each step, we retained only those loci at which we observed shared conservation of *mbs*, and used these as queries in the next step.

To estimate the proportion of *mbs* expected to be conserved, by chance, we constructed a simple probabilistic model based on estimated neutral rates of evolution. We also performed a parallel procedure on an arbitrarily defined control sequence. Our results show that such a model accurately predicts imperfect conservation of a specified motif, within a precisely defined evolutionary space. We also present evidence that a subset of *mbs* have been subject to purifying selection during anthropoid evolution, as reflected in a significantly greater proportion of cross-species conservation than is expected by chance. Additionally, we note the identification of several individual loci exhibiting broad conservation, making them promising targets for in vivo analyses.

2.2 Methods

2.2.1 Identification of Mbs in Human, Macaque, and Marmoset Genomes

Based on the results of a previous study of in vitro binding by SETMAR, we defined a *mariner binding site* (*mbs*) as a copy of the sequence indicated in Figure 1, and occurring within the TIRs of *made1* and *Hsmar1* transposons (Smit and Riggs 1996, Robertson and Zumpano 1997), with no more than 1 nucleotide substitution (versus the consensus, shown) among the 17 positions indicated in upper case. The positions indicated in bold face are intolerant positions

at either of which a single substitution inhibits binding by SETMAR, in vitro. Therefore, in our model, an *mbs* may not harbor a substitution at either of these positions. Henceforth, we will use the terms *mbs* and *mariner binding site* to denote only those sequences which satisfy these criteria. Homologous sequences with more than 1 substitution, or with a substitution at either of the intolerant positions, will not be referred to as *mbs* but, rather, as degenerate TIRs.

We used the *mbs* consensus as a BLASTN query and parsed the output to retain all sequences with no more than 1 mismatch. We then intersected the genomic coordinates for these sequences with a table of coordinates for all *made1* sequences as annotated in the Table Browser on the UCSC Genome Browser (Karolchik et al. 2004, Kent et al. 2002), and extracted the coordinates for those *made1* which harbored putative *mbs*. We then used the sequences of these *made1* elements, augmented by 100 nucleotides of flanking sequence, as queries in a BLASTN interrogation of the macaque genome in the NCBI wgs database (Altschul et al. 1990). We retained all loci which returned hits in the macaque genome. All sequences retrieved had e values < 1e -30, and exhibited a high level of identity throughout flanking as well as transposon sequence. It was at this point that we examined the dataset, confirming the presence of a bona fide *mbs* in the query, and discarding any loci at which the query contained only degenerate TIRs. (Our initial search of the human genome had returned some sequences containing a single mutation residing at either intolerant position.) Then we evaluated each macaque sequence for conservation of *mbs* and tallied the results for statistical analysis. From this first screen, we retained those loci at which the *mbs* observed in human was conserved in macaque and used the human sequences from these loci, again with 100 nt of flanking, as BLASTN queries, this time against the marmoset genome. We analyzed the output from the marmoset interrogation as described above for macaque.

2.2.2 PCR and Sequencing

We designed oligonucleotide primers for 44 loci (Table 2.1, 2.2), aided by multiple alignments produced using ClustalX. The loci were selected on the basis of apparent amenability to PCR, especially for the absence of repetitive elements in flanking sequence. We

performed PCR in 25-μL reactions, using 280 nM of each oligonucleotide primer, 320 μM dNTPs, in 2.0 mM $MgCl_2$ at pH 8.5, with 1.5 units of Taq polymerase, and 20 ng of genomic DNA from 2 apes (chimp, gorilla), 1 OWM (green monkey), 5 NWM (red-bellied tamarin, squirrel monkey, owl monkey, woolly monkey, dusky titi). After an initial denaturation step of 95℃ for 5 min, we performed 30 cycles of 30 s at 95 ℃, 30 s optimal annealing temperature (60℃ for set 1, 55℃ for set 2), and 1 min extension at 72 ℃. DNA from tamarin, squirrel monkey, and titi was obtained from the Coriell Institute for Medical Research. DNA for chimp, gorilla, green monkey, owl monkey, and woolly monkey, plus human Hela cell DNA (as a positive control), was obtained from Mark Batzer (Louisiana State University). We separated the PCR products on 1.5% agarose gels, stained with ethidium bromide, and visualized by UV fluorescence. We performed direct sequencing of PCR products, using an ABI 3130xl automated sequencer, following exo-sap purification protocol and ethanol precipitation, according to manufacturer's specifications. The trace files were visualized and edited using Bioedit v7.0.5.3 (Hall 1999).

Table 2.1 Oligonucleotide Primers for 44 *Mbs* (Set 1)

Hs coordinates	primer sequence 5'--> 3'
chr1:39031259-39031338-f	AATGGCAGAATCGTCTGGAG
chr1:39031259-39031338-r	TCAGGTATTTTGACTGAACATAGGA
chr1:61658139-61658185-f	GCTACCCTGAAATTCACATGC
chr1:61658139-61658185-r	TGTCACTTGTCCTGGCTCTG
chr1:176411046-176411126-f	ACAGGAAGGTGGAGCAGAGA
chr1:176411046-176411126-r	CCTGCCCTAGACTTTTCCAG
chr1:181527181-181527227-f	TGGCTCACTCTTCCCATACC
chr1:181527181-181527227-r	ACTGGTGGGGATGTTTTTGA
chr2:138492190-138492269-f	TGATTGAGGAGGAAAGCAAAA
chr2:138492190-138492269-r	AACATTGGCTTGGGACTGAG
chr2:144850789-144850834-f	GGATGTCAGCAGGTACACCA
chr2:144850789-144850834-r	GCTCGCCAAAAGAGTCATTA
chr3:197175692-197175772-f	TGGGCTGCTGCTTTATTTTT
chr3:197175692-197175772-r	AAAACAAAAGGGCTTTCTTATCAA
chr4:31433807-31433881-f	TGCACAGAGGTGGCAGATAA
chr4:31433807-31433881-r	CAAGGACTGAGGTATGGCAGA

Table 2.1- *Continued*

Hs coordinates	primer sequence 5'--> 3'
chr4:101408285-101408365-f	GGTGTCTGTGAAGCACAAATG
chr4:101408285-101408365-r	GAGCCTGAACGTGGTAAACTG
chr4:166952345-166952419-f	GCTACTCAGTTTTACGTCTGTGC
chr4:166952345-166952419-r	TCATCCCTGTGGTAAGAGGA
chr5:71351746-71351826-f	TAGCTTGGTGGCTGGAAGTT
chr5:71351746-71351826-r	CGGCTGAATTGTCCTCATTT
chr5:106794912-106794963-f	TGGCCAGAAAAGAATGGTTC
chr5:106794912-106794963-r	ACATTTCAAAGGGTGGACAA
chr5:163315201-163315274-f	GCTTAGTTAGATCATGGCCACA
chr5:163315201-163315274-r	TCCAGAGGCTCAGAGAACAGA
chr6:52056481-52056558-f	TTCCAGTCCTGGAGAGAAAAA
chr6:52056481-52056558-r	TAAGCAGGCAAGGCAAGAAT
chr6:105062768-105062849-f	TGATGCCTGCTCTGCATTTA
chr6:105062768-105062849-r	ACTGTTCTGGGGCAAATGAG
chr6:111956048-111956102-f	AAATGACTCTGACTTATTACTTGTGC
chr6:111956048-111956102-r	GCCCATCACAAAAACAAGAA
chr6:130991857-130991912-f	GGGTGGGAGGAATTGATTTT
chr6:130991857-130991912-r	GGAATACCACCTCACTGGAAA
chr6:160311434-160311514-f	GGAGGTAGGGTGCATAACGA
chr6:160311434-160311514-r	AAGGAACCAAGAAGAATGAAACC
chr6:169688860-169688936-f	GCAGCTATGGCTCTAGCTCTG
chr6:169688860-169688936-r	GCTCTTGCAGGTGGTTGATAG
chr7:31767544-31767590-f	TCTGTAACAGCTTGGTTTATCACA
chr7:31767544-31767590-r	TCAACATTCAAACTTACATCTGTGT
chr7:98822072-98822147-f	GTGTGCACCCCAACATCTC
chr7:98822072-98822147-r	ATCTCATGCCTCTGGCAAAC
chr8:7748433-7748513-f	CTGGAAAGGTGGCATTCTGT
chr8:7748433-7748513-r	GCAGCTTTCTTTAATCCCCATA
chr8:66519456-66519533-f	TGGGTTGGGTTTGAGGTTTA
chr8:66519456-66519533-r	GCTGACCAAATTCTTCCCTCT
chr8:113557005-113557085-f	CCTACAGCGTGGCCTAATTT
chr8:113557005-113557085-r	TTGGCCTGACTTATTATTCCA
chr9:70832017-70832067-f	CACAGGCTCTGCGTAAGATG
chr9:70832017-70832067-r	AACCTCTTTGAACTTGTTTTCACA
chr9:88308708-88308793-f	ATTGCTTCCGACGTAAAACG
chr9:88308708-88308793-r	CCCAAGAGCTTTTCTGTAAATCA

Table 2.1- *Continued*

Hs coordinates	primer sequence 5'--> 3'
chr10:115287615-115287660-f	GTCCAGGGTCTAAGGGGAAG
chr10:115287615-115287660-r	TGAGCATCACATGGGGTAGA
chr12:4950874-4950953-f	GCTACTCAAGCCTTGCATCC
chr12:4950874-4950953-r	CACCCTCCACTTCACCTCAC
chr13:24718671-24718747-f	ACATGGAAAAGCCAACAAAA
chr13:24718671-24718747-r	TTTTTGCAGAAGGTCAGCTTT
chr13:82331799-82331878-f	GCTATCGTTCAACCATGCAA
chr13:82331799-82331878-r	CGGAGAAGTGAATAAGTGAAGG
chr13:107086387-107086465-f	TAATTCGGCCGTGTGATGTA
chr13:107086387-107086465-r	GGGAAAATAAATAGGGGAAGC
chr13:108801761-108801840-f	TTATAGTGGCATGTCCGTAACTT
chr13:108801761-108801840-r	TGTTTGGCTGAGTTCAGAGATT
chr15:45037029-45037103-f	TCCTTTTCTCTCCTCTCATTTCTC
chr15:45037029-45037103-r	TGCAGAGTTCTTCATTTTTACCA
chr15:61426464-61426537-f	ACCCCAGGTGGGAAAAATTA
chr15:61426464-61426537-r	TCAATGTCCTCAGGGGTTTC
chr15:92425336-92425415-f	TGCTTCATTGGTTCTTGTGG
chr15:92425336-92425415-r	TCTTTGCCTTGAGTTTTAAGTCG
chr16:23932066-23932183-f	CTGGGTTCTTTGTCACCATGT
chr16:23932066-23932183-r	TCAGCACACCACATACCTCTTC
chr16:77684359-77684433-f	TAAGGCCCATTTTCATGGAT
chr16:77684359-77684433-r	ACGTAACCAACAGCTCAGCA
chr18:21134586-21134666-f	GAGTGTCATCAAAGCAAACTTAAA
chr18:21134586-21134666-r	AGGAGGGAGGAGCAGAAAAG
chr18:51981646-51981698-f	TCTTCACTACCCTTAAGAGGAACTTT
chr18:51981646-51981698-r	CACTTGGCAAGACCAAACCT
chr18:67645193-67645272-f	TCTGGATTCTAGAAGTGTTGGTG
chr18:67645193-67645272-r	TGCCTAAGTGAACACACATGAA
chr21:38199634-38199671-f	TGGGGAAATGTGAAGGAGTC
chr21:38199634-38199671-r	CCTGCCTGATCAATCCTCAT
chrX:83286905-83286983-f	GGCTGGAGCTTCTGTGAAAT
chrX:83286905-83286983-r	TCGGGTCTGAACAAAGGTTA
chrX:123130513-123130585-f	CTCTGGCCTGCATCAAAGTA
chrX:123130513-123130585-r	TTCCACAGTTTAAAGAGCAGACA
chrX:134948504-134948534-f	TGAAGGCAGCTTATCTCTCATGT
chrX:134948504-134948534-r	TGAGGGAGATAAAGGATGCTATG

Table 2.2 Oligonucleotide Primers for 44 *Mbs* (Set 2)

Hs coordinates	primer sequence 5'--> 3'
chr1:39031259-39031338-f	CTCTATGTCTCAGTTTCCCTACCTGC
chr1:39031259-39031338-r	GGATTCATCTTGCTTGGGGAATCTG
chr1:61658139-61658185-f	GGGCATACAACACAACTAGGATGG
chr1:61658139-61658185-r	GCAACTCCCATCCCCTGG
chr1:176411046-176411126-f	GCTGAAAATCATTTCCCCTGGAAAGTC
chr1:176411046-176411126-r	GAGAAAAGGGCATTGCAAACCAT
chr1:181527181-181527227-f	CATTGGTCTCGGTTCCTCTTGG
chr1:181527181-181527227-r	CATTCTCTGGGGGCAGTCG
chr2:138492190-138492269-f	TGCCAAGCAGTGCACTTGG
chr2:138492190-138492269-r	GCTCTGTGAGACAGTGTGCTC
chr2:144850789-144850834-f	GTGATCCACCCACCTCAGC
chr2:144850789-144850834-r	GAGAAAATTCCATTCTGGGGAACC
chr3:197175692-197175772-f	GGGACCGCCAGTTTAAATAGC
chr3:197175692-197175772-r	CGAGATGGTGAAACCCCGTC
chr4:31433807-31433881-f	GGAAAGTCAGGGAAAATGCACAGAGG
chr4:31433807-31433881-r	CAAGGACTGAGGTGTGGCAG
chr4:101408285-101408365-f	GGTGTCTGTGAAGCACAAATGAG
chr4:101408285-101408365-r	GGTCTGGCTGTTTCCATGGC
chr4:166952345-166952419-f	GCTACCAGAGCCATGTTTTGCC
chr4:166952345-166952419-r	GACTTTTCTCAGCTGCACCTGG
chr5:71351746-71351826-f	GGGTGAGCACTAGACCAGC
chr5:71351746-71351826-r	GGCTTCCACTCTCCACAGG
chr5:106794912-106794963-f	CTCTGGAGGAATACAGAGCTGACC
chr5:106794912-106794963-r	GCCTCGATCTCCTGACCTCG
chr5:163315201-163315274-f	CTGTTCCCTGGCAATGTTTCACC
chr5:163315201-163315274-r	CGCTCAGACAAATAAGCAGAACACTG
chr6:52056481-52056558-f	TCCCACTCTTTAATCAGGTGGC
chr6:52056481-52056558-r	CACTCCCTCTCTAGGCAGTCTAG
chr6:105062768-105062849-f	GGGCATTGGTCACATGTATGTTGG
chr6:105062768-105062849-r	CCTCCATATCTTCTAACACTACCACCC
chr6:111956048-111956102-f	AGGCATGTGTGTACTAGGTCTTC
chr6:111956048-111956102-r	ATGTTGCATATATCCTGCCCCTCAC
chr6:130991857-130991912-f	CTGCCCCTCCAGCAACC
chr6:130991857-130991912-r	CATTCAGATTGACTAAGAACTCCCACC
chr6:160311434-160311514-f	GGAAAGGCTCCAAACTGCAGG
chr6:160311434-160311514-r	GGTGGAAGCCTGCAGGG

Table 2.2- *Continued*

Hs coordinates	primer sequence 5'--> 3'
chr6:169688860-169688936-f	GCTCAGGATAAAGACTGCGATTTGC
chr6:169688860-169688936-r	CTGTGCTAGCACCCTCAGACC
chr7:31767544-31767590-f	GTCAAGGATCTGTCCCTCTGGC
chr7:31767544-31767590-r	CCAGGTCTGTTGGGCAGTG
chr7:98822072-98822147-f	CTCTCCCTGTGTGCACCC
chr7:98822072-98822147-r	CCTGGTGGGCTAGGCG
chr8:7748433-7748513-f	GTCTGGAAAGGCGGTTCTCTG
chr8:7748433-7748513-r	CTGCACATCACAAAACTTTGAGCC
chr8:66519456-66519533-f	GGTGTGGGTTGGGTTTGAGG
chr8:66519456-66519533-r	GCAGAAAATGAAACTGCTCATGGG
chr8:113557005-113557085-f	CTGATGAGTGCCTACAGTGTGGC
chr8:113557005-113557085-r	GGTCTGTCTTGCTTTTATCTTCAACC
chr9:70832017-70832067-f	CCTGCTGAATTTTGAACACAGGC
chr9:70832017-70832067-r	CCTTCAAAGGCATTCCAAGAGTCAAC
chr9:88308708-88308793-f	GGAACCACTGCTCTGTGACATCAC
chr9:88308708-88308793-r	TAGCTCACTTGAAGGTTCCTGC
chr10:115287615-115287660-f	CCCGTTCTTCCTCAGCACC
chr10:115287615-115287660-r	GCAACCTAGACTGCATGGCC
chr12:4950874-4950953-f	CTGTAGCTGGGTCCCACAG
chr12:4950874-4950953-r	GTACAGGTCCTGCACCCTC
chr13:24718671-24718747-f	CCAGAGTCCCAACCCCTAACTAC
chr13:24718671-24718747-r	GGTTGGGTAGCAGAGCAGG
chr13:82331799-82331878-f	CCTAAAAATTGCCTGTGGTAGGGC
chr13:82331799-82331878-r	GCAATCCACAATGTTGGGCAG
chr13:107086387-107086465-f	GAGAGATGTGATTGACCTGCTTCC
chr13:107086387-107086465-r	CAGAAGCCTATCCCAAATTTGCAGG
chr13:108801761-108801840-f	CAGGGCCCTCATTTGTTGC
chr13:108801761-108801840-r	GTTTGGTGGAGTTCAAGAGATTCCG
chr15:45037029-45037103-f	GTGCTATCTCTTTCCTGTCAGGACC
chr15:45037029-45037103-r	GAGCAGCAGTCTACATATGCAGAG
chr15:61426464-61426537-f	CCTGCTCATCCAAGCTGGC
chr15:61426464-61426537-r	CAGGGGTTTCTCATCAGGAAGGTC
chr15:92425336-92425415-f	GTAGGGAGACTGTCTCTCTGTGC
chr15:92425336-92425415-r	GGCATTATCTAGAGCTTTGTCCGAG
chr16:23932066-23932183-f	GCCAAGCACTGTTCTAAGTGC
chr16:23932066-23932183-r	CTGACCTCCCTGACCAGG

Table 2.2- *Continued*

Hs coordinates	primer sequence 5'--> 3'
chr16:77684359-77684433-f	GAGGCTGTCTCACTCAGGC
chr16:77684359-77684433-r	CACGTCCTCCCAGCCAC
chr18:21134586-21134666-f	CTCATGTTGGCATCACTTTCCATCTC
chr18:21134586-21134666-r	CCACCTTTGTGCTTGCAATGC
chr18:51981646-51981698-f	GCATATTATGAAGCAGGAGCATTTGG
chr18:51981646-51981698-r	CTGCTGGGATAAAGCAGCAC
chr18:67645193-67645272-f	GGTGTTAGGGAGGCCAGG
chr18:67645193-67645272-r	GAACATTGCCTGAATGAACACACATG
chr21:38199634-38199671-f	GCAGAAAACATCCAGGGATGTGG
chr21:38199634-38199671-r	CTCTGATTTCCATGGTGTGGTAGACC
chrX:83286905-83286983-f	GAGGCTGGAGTTTGTGTAAAATCCC
chrX:83286905-83286983-r	CCTGATGTGAAAAGCCCCAGG
chrX:123130513-123130585-f	CAATAGTGCCAGGGTTGATGAAC
chrX:123130513-123130585-r	GGCACAAAGCCATGGATAAATTTAGG
chrX:134948504-134948534-f	GCTCAAGTGATCCTCCCACCTC
chrX:134948504-134948534-r	CAACCGAGGCCAAGTGTAGAC

2.2.3 Model

We constructed a simple probabilistic model for predicting the proportion of mbs observed at time 1 which should be conserved, by chance, at time 2, assuming neutral evolution. For the macaque data, time 1 resides at the divergence of the hominoid and cercopithecoid lineages, ~25 mya. For the marmoset data, time 1 corresponds to the divergence of the catarrhine and platyrrhine lines, ~40 mya (Figure 2.2). In selecting appropriate rates of neutral evolution, we were aware of the problem posed by heterogeneity of molecular clocks not only among lineages, but within genomes (Laird et al. 1969; Li et al. 1996; Nei and Kumar 2000; Kumar and Subramanian 2002; Yi et al. 2002; Kumar 2005; Li 1997; Kim et al. 2006). Therefore, rather than relying upon estimated rates of nucleotide substitution, available in the literature, we calculated these rates directly from the data. In this estimation, we restricted our analysis to loci at which we observed an mbs in human, macaque, and marmoset, assuming that any local variation in neutral rate which might bias conservation of *mbs* would

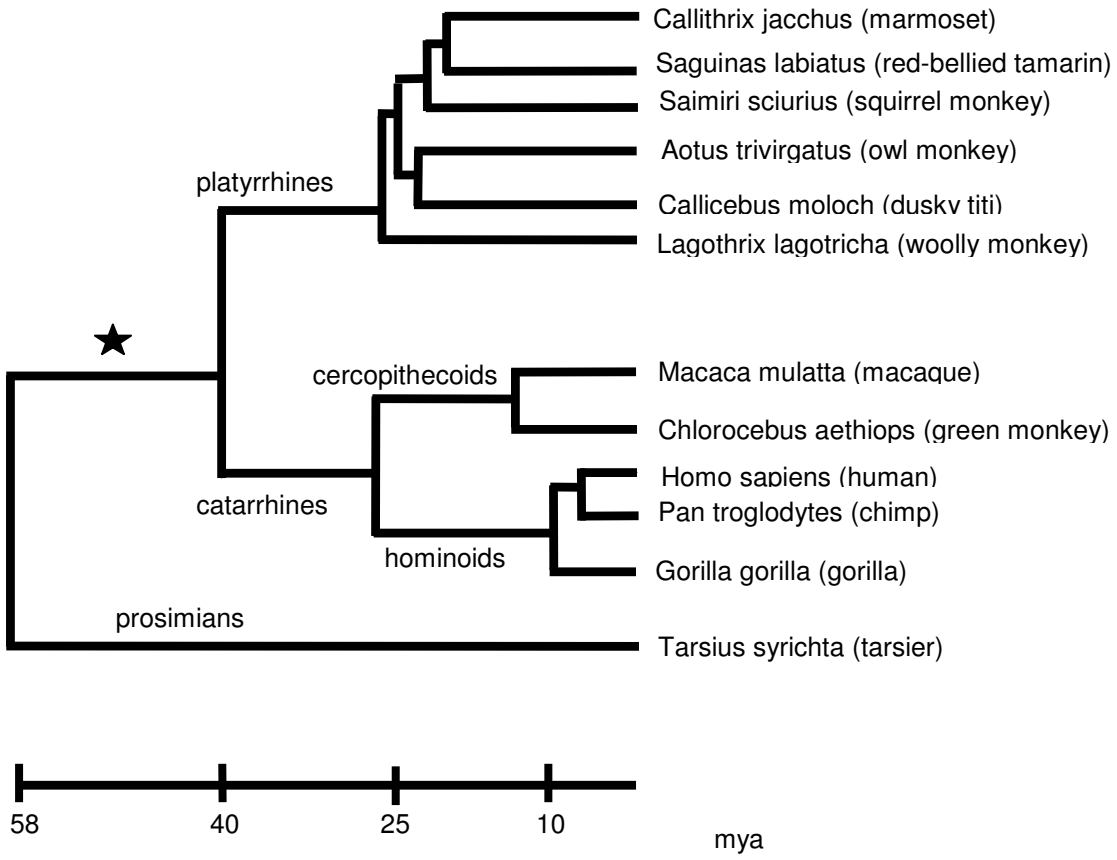


Figure 2.2 Genome colonization by *Hsmar1* and *made1* transposons, and formation of SETMAR, occurred 40 – 58 mya in the anthropoid lineage. The star indicates the branch on which these events occurred.

likewise be reflected in the estimation of neutral rates thus obtained. As it has been observed that CG dinucleotides are hypermutable in mammals (Bird 1980, Labuda and Striker 1989, Batzer et al. 1990, Kim et al. 2006), we calculated separate rates for CG and non-CG positions. We performed the estimation of non-CG rates as follows. For the macaque, we evaluated flanking sequence in 3-species alignments, with marmoset as outgroup. This allowed us to assign substitutions, occurring after divergence of catarrhines and platyrrhines to 1 of 3 branches: hominoid, cercopithecoid, or prior to divergence of these lineages. We excluded

ambiguous positions, at which we observed 3 character states. To infer the neutral rate in marmoset, we utilized the same alignments, but instead of evaluating flanking, we examined the *made1* sequences, excluding *mbs*. This allowed us to use the *made1* consensus as a de facto outgroup. Based on these analyses, we calculated rates of neutral substitution for marmoset and macaque lineages, following their respective divergences from catarrhine and hominoid lineages. The hypermutability of CG dinucleotides makes estimation of substitution rates at these sites problematic, with a likelihood of undercounting such sites. We therefore estimated divergence of these positions versus the *made1* consensus, excluding *mbs*. We did not apply a correction because we were interested in observed conservation of mbs, which should be predicted by observed, as distinguished from inferred, rates of substitution. Applying a correction produces an inferred rate of substitution greater than the observed rate, in order to compensate for reversion. Such a correction, applied to our model, would decrease the calculated probability that an *mbs* observed in human is also observed in another species. However, it is likely that some proportion of observed *mbs*, in any lineage, have not been conserved throughout anthropoid evolution but have, at some point harbored 2 substitutions, subsequently undergoing reversion at 1 of these positions, thereby restoring their status as *mbs*, and theoretically, any in vivo function which may have been compromised during the interim. Therefore, correcting for reversion in the estimation of probabilities, but not in the count of observed *mbs* would have introduced a bias, increasing the probability of erroneously rejecting the null hypothesis of neutral evolution (Type I error). Having obtained estimated substitution rates, we calculated the probabilities of *mbs* conservation as follows.

The probability of observed conservation over the interval is dependent upon whether the sequence harbored 0 or 1 substitution at time 1. Therefore, we estimated the proportion of *mbs* in each of these categories at each of these divergences. We inferred the number of substitutions at time 1 by pairwise alignment of the sequences. If a substitution observed in the human sequence was also observed in the macaque, we assumed that the substitution was present in the most recent common ancestor of hominoids and cercopithecoids, and likewise

regarding the marmoset data. For loci at which we observed no shared substitution, we assumed the ancestral *mbs* to have been pristine. So, the probability that an *mbs* which harbored no substitutions at time 1 exhibits fewer than 2 substitutions at time 2 is the probability that it incurs no substitution during the interval plus the probability that it incurs 1 substitution at either position within the CG dinucleotide plus the probability of a substitution at any of 13 non-CG positions at which a substitution is tolerated. The resultant probability is then multiplied by the probability that no substitution occurs at either intolerant position, described above. This probability is given in equation 1. The probability that an *mbs* which harbored 1 substitution at time 1 exhibits fewer than 2 substitutions at time 2 is the probability that it incurs no substitutions during the interval. If the single substitution inferred at time 1 is at a non-CG position, this probability is equal to the probability that, over the interval, it incurs no substitution at any of 14 non-CG positions plus the probability that it incurs no substitution at either position of the CG dinucleotide and is given in equation 2. If the inferred substitution at time 1 resides at either position of the former CG dinucleotide, the probability of observing an mbs at time 2 is the probability of no substitution occurring at any of 16 non-CG positions, as there is no longer a CG dinucleotide present in the sequence. This probability is given in equation 3. The probability, then, that an *mbs* observed in human and thereby inferred at time 1 is observed at time 2 is the sum of the above probabilities after each has been multiplied by the inferred proportion of each class at time 1 and is given by equation 4. Equations 1a – 3a are presented in general form. Equations 1b – 3b are based on the nucleotide composition of mbs.

Equation 1a. $P(m|0) = [P(1\ C)\ P(0\ n) + P(0\ C)\ P(1\ n) + P(0\ C)\ P(0\ n)]\ P(0\ i)$

Equation 2a. $P(m|n) = P(0\ C)\ P(0\ n)\ P(0\ i)$

Equation 3a. $P(m|C) = P(0\ n)\ P(0\ i)$

Equation 1b. $P(m|0) = 2\ P(C)\ P(1 - P(C))\ (1 - P(n))^{15} + (1 - P(C))^2\ (13)\ P(n)\ (1 - P(n))^{14}$

$+ (1 - P(C))\ (2)\ (1 - P(n))^{15}$

Equation 2b. $P(m|n) = (1 - P(C))^2\ (1 - P(n))^{14}$

Equation 3b. $(1 - P(n))^{16}$

Equation 4. P(m) = proportion (m 0) P(m|0) + proportion (m n) P(m|n) + proportion (m C) P(m|C)

2.2.4. Control

To test the model, and to provide a control for any error in estimation of rates, or biases introduced by simplification, we performed a parallel phylogenetic shadowing procedure on an arbitrarily defined sequence occurring within the *made1* consensus but outside the *mbs*. The sequence is indicated in bold face below (Figure 2.3). The control sequence, like *mbs*, is composed of 2 CG and 15 non-CG positions. (We note that the only CG dinucleotides in the *made1* consensus are those within the mbs. Therefore, we included these in the control.) We designated the nucleotide positions indicated in bold to represent the 2 intolerant positions in *mbs*. We then assessed the phylogenetic shadowing of this control sequence, using the same alignments we utilized in our estimation of neutral rates, each of which exhibit *mbs* in human, macaque and marmoset. We assumed that, in so restricting the data, any effect of local heterogeneity in the neutral rate of evolution which might bias the observed conservation of *mbs*, would exert a similar bias on the conservation of the control. We applied the same approach as described above for *mbs*.

```
human                CAGGCATCTATAGGTTGGTGCAAAAGTTATTGCGGTTTTTGCCATTGAAA
macaque              CA-GTATCTATAGGTTGGTGCAAAAGTTATTGCGGTTTTTGCCATGGTAA
marmoset             CA-GTATTCATGGGTTGGTGCAAAAGTTATTGCAGTTTTTGCCATGGAAA
control                      TTAGGTT----------------CG-TTTTTGCC
```

Figure 2.3 Phylogenetic shadowing control. The control sequence, in bold, flanks the *mbs* and includes the CG dinucleotide within *mbs*. Positions underlined represent the 2 intolerant positions, as described for *mbs* in methods. Mismatches to the consensus are in red. In this example, the control sequence, conserved in human, is also conserved in macaque but is degenerate in marmoset.

2.2.5. Additional Shadowing

After obtaining sequence data for additional species, as described above, we performed a similar estimation of substitution rate, calculation of expected *mbs* conservation, and shadowing control for the 5 NWM species in this dataset. For simplicity, and to avoid fragmenting a small dataset, we treated the 5 NWM as a composite, calculating a single mean

substitution rate for this group over the past 20 myr since the platyrrhine diversification. Using this rate, we calculated the conditional probability of observed mbs conservation, given the observation of mbs in marmoset.

2.3 Results

2.3.1 Identification of Orthologous Mbs

Based on the definition of *mbs* given above, our interrogation of the human genome yielded ~1200 *mbs*. Interrogation of the macaque genome with these queries returned orthologs of 842 *mbs* conserved in human. We evaluated the macaque sequences and observed 505 / 842 (59.98%) human *mbs* to be conserved at the orthologous loci in macaque.

We then used the human sequences from those loci at which *mbs* was also conserved in macaque as queries, and interrogated the marmoset genome. We retrieved 294 orthologs, of which 120 (40.81%) were conserved, as previously defined.

2.3.2 Excess of Mbs Observed in Marmoset

To arrive at an expected proportion of *mbs* observed in human which should be conserved, by chance (assuming neutrality) in macaque, we applied a standard conditional probability, as described in methods, based on an estimated rate of neutral evolution in the macaque genome. By evaluating nucleotide sequences flanking conserved *mbs*, we calculated a mean divergence of ~3% at non-CG positions in macaque, relative to the inferred sequence for the common ancestor of human and macaque (~25 mya). CG positions exhibited ~44% divergence from consensus. Dividing this figure by the upper bound for the age of *Hsmar1/made1* (58 myr) yielded an estimated rate of ~8.0 x 10^{-9} mutations/site/yr, or ~20% since divergence of cercopithecoid and hominoid lineages. This is consistent with estimates in the literature of a CG mutation rate up to 10 times greater than the non-CG rate in primates (Labuda and Striker 1989, Batzer et al. 1990). As described in methods, the probability that an *mbs* inferred at a node is conserved at the end of a branch depends upon whether the ancestral *mbs* had 0 or 1 mutations, and if 1, whether it occurred at a CG dinucleotide. Therefore, we examined human-macaque alignments for all orthologs in order to infer the ancestral state of

each *mbs*. By our observation, 60% of *mbs* exhibited no mutation common to human and macaque, and were assumed to be pristine in the common ancestor. 27% exhibited a shared non-CG mutation, and 12% shared a CG mutation. Substituting these values into the equations given in methods, we calculated an expected proportion of *mbs* conserved in macaque, given occurrence in human, to be 0.6490, or approximately 546 out of 842. We note that this figure is significantly greater than the observed proportion of 0.5998 (505 / 842) ($X^2 = 9.35$, $p < 0.005$) (This deficit of observed *mbs* is likely to be an artifact of the model, as we will discuss below.) The lack of an observable excess of conserved *mbs* in macaque suggests that, if some *mbs* are evolving under constraint, they comprise a small subset of loci, within a neutrally evolving population of sites.

In the marmoset analysis, we estimated a mean divergence of ~9.6%, at non-CG positions, relative to the inferred sequence for the common ancestor of human and marmoset. CG positions showed ~48% divergence from consensus, from which we estimated ~32% divergence since the catarrhine-platyrrhine split (~40 mya). By our observation, 80% of observed *mbs* exhibited no mutation common to human and marmoset, and were assumed to be pristine when these lineages diverged. 14% exhibited a shared non-CG mutation, and 5.5% shared a CG mutation. Substituting these values into the equations given in methods, we calculated an expected proportion of *mbs* conserved in human also conserved in marmoset to be 0.2971, or approximately 87 out of 294. Our observation of 120 / 294 (0.4081) was significantly greater than the predicted frequency ($X^2 = 17.8$, $p < 0.00005$) (Figure 2.4).

2.3.3 Testing the Model on an Arbitrary Control

To test the reliability of the model, and to control for error in our estimations of rates, we conducted a parallel phylogenetic shadowing of a presumably neutrally evolving sequence also occurring within *made1* transposons and immediately adjacent to the *mbs*. As described in methods, we defined a sequence according to the *made1* consensus and comprised of the same number of non-CG positions plus a CG dinucleotide. We applied the same definition of a

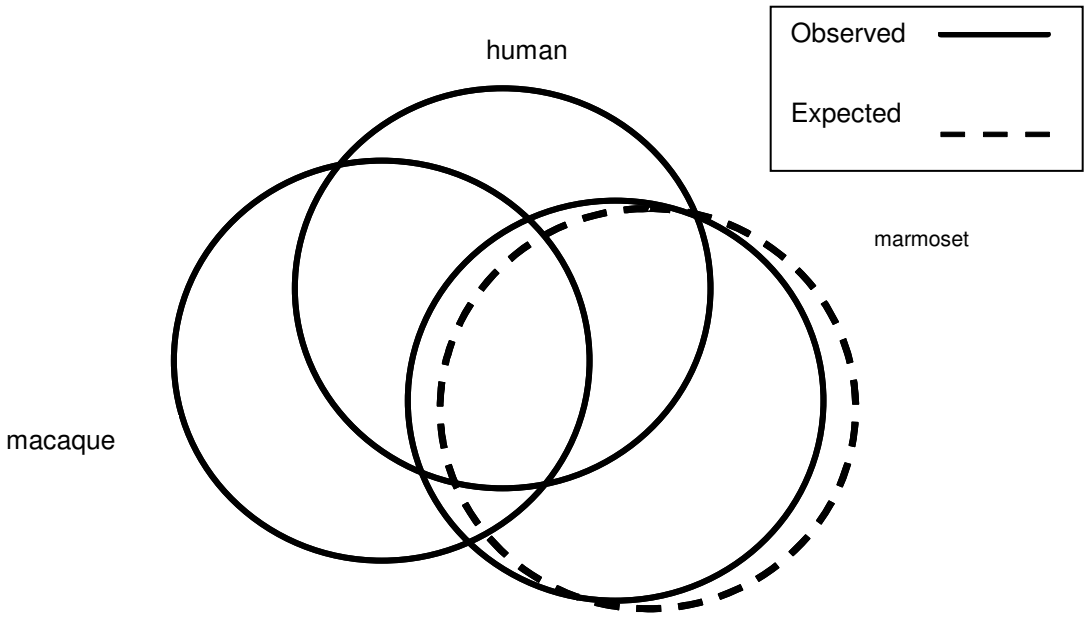


Figure 2.4 Phylogenetic shadowing in a sequential culling process, with a significant excess of observed *mbs* in marmoset.

conserved sequence as having no more than 1 substitution versus the consensus, and we designated 2 position to be intolerant of substitution, as in our *mbs* model. We evaluated only those loci for which we observed *mbs* in human, macaque, and marmoset, as in the estimation of neutral rates. We note that this shadow control sequence has a somewhat lower GC content than *mbs* and should, therefore be expected to exhibit a slightly lower rate of neutral evolution, increasing the observed level of conservation and making it a sufficiently conservative control model. In this analysis, we observed 51.2% conservation in macaque and 32.4% conservation in marmoset. We note that, for the marmoset, this value is very nearly identical to the 29.7% predicted by the model ($X^2 = 0.13$, $p > 0.7$). For the macaque, we observed a greater discrepancy between the predicted conservation (64.9%) and that observed in the control. We note, however, that the difference is not significant at the sample size used for the control ($X^2 = 3.37$, $p > 0.05$). We take these results as evidence that the equations described in methods

produce a reasonably reliable null model. We will consider possible explanations for the apparently looser fit to the macaque data in the Discussion, below.

2.3.4 Additional Phylogenetic Shadowing

Next, we evaluated multiple alignments for the 120 loci conserved in human, macaque, and marmoset, for potential amenability as PCR targets in additional anthropoid genomes. We selected 44 loci, for which we designed primers and performed PCR using genomic DNA from 2 apes (chimp and gorilla), 1 Old World monkey (green monkey), and 5 New World monkeys (red-bellied tamarin, squirrel monkey, owl monkey, dusky titi, woolly monkey). We had dual objectives. First, we wished to further pare down the list of loci potentially subject to purifying selection, continuing to eliminate sites lacking broad conservation, so that we might reduce our dataset to a manageable number of candidate sites for in vivo analysis. Second, we sought further statistical support for our hypothesis of purifying selection acting on a subset of *mbs*. The genomes being evaluated, especially the New World monkeys, which diversified ~20 mya are sufficiently divergent as to limit our success in obtaining PCR products from these genomes. We have thus far obtained and sequenced PCR products for nearly all of the ape and Old World monkey targets and the majority of the New World monkey sequences. To estimate expected proportions of conserved *mbs* in these lineages, we applied the same approach described above for macaque and marmoset. We analyzed the multiple alignments and arrived at an estimated mean observed nucleotide substitution rate of ~2.5% at non-CG positions over all of these branches, treated as a composite. We applied a CG rate of 16%, as described above. Substituting these values into the equations described in methods, we arrived at an expected proportion of ~0.7156 conserved, given conservation in marmoset. In our shadow control for the NWM data, we observed 70% conservation, suggesting that the model fits the data very well. We estimated divergence in the green monkey, since divergence from macaque, to be ~1.3% and 8.8% at non-CG and CG sites, respectively. Substituting these values into the equations described in methods gives an expected *mbs* conservation in green monkey of ~86%, given conservation in macaque. The observed levels of *mbs* conservation were slightly lower, but

similar to these values (66.7% and 81.5%, in NWM and green monkey, respectively.) We are continuing to generate data for these lineages, but we have not yet identified evidence of purifying selection on these sequences over the past 20 myr in either OWM or NWM.

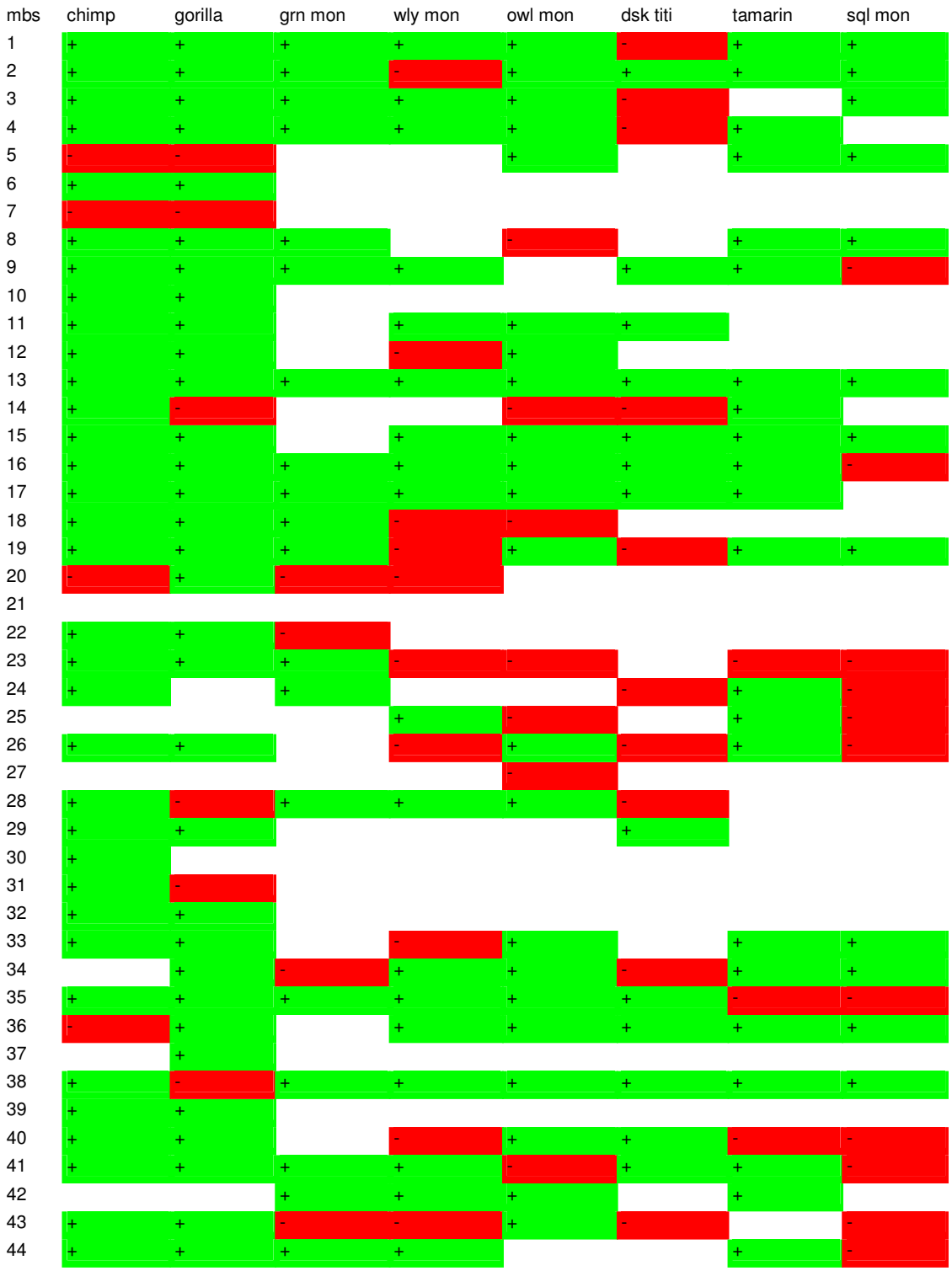
We have, however, identified several loci which exhibit broad conservation and are promising targets for in vivo analysis of binding by SETMAR (shown in Table 2.3, with "+' indicating *mbs* and "-" indicating degenerate TIRs). These include 14 loci at which we have identified conserved *mbs* in at least 5 anthropoids (in addition to human, macaque, and marmoset) including at least 3 NWM, while exhibiting no more than 1 loss. This group includes 7 loci with conserved *mbs* in at least 7 anthropoid genomes (in addition to human, macaque, and marmoset) including at least 4 NWM, while exhibiting no more than 1 loss. We have identified 3 loci with conserved *mbs* in at least 7 anthropoid genomes (in addition to human, macaque, and marmoset) including all 5 NWM, while exhibiting no more than 1 loss. Finally, within this group, we have identified a single ubiquitously conserved locus, with a conserved *mbs* in all species screened (Figure 2.5). We performed a binomial analysis of these observed levels of conservation, applying the probabilities described above. The results do no not differ from the pattern of conservation expected by chance.

Using the annotations in the UCSC Genome Browser, we have investigated the genomic context of these 14 broadly conserved *mbs*. Our preliminary results betray no obvious pattern, with 4 residing within introns, 2 downstream of a proximal gene, and nearly half residing in gene deserts (> 50 kb from the nearest annotated gene) (Table 2.4). So far, we have only scant gene ontology data.

2.4 Discussion

We have shown that a simple model of conditional probabilities can accurately predict levels of conservation of a defined motif among taxa. The fact that the model very accurately predicted conservation of an arbitrarily defined and presumably neutrally evolving control sequence, in both marmoset and the composite NWM datasets demonstrates the efficacy of the

Table 2.3 Additional Phylogenetic Shadowing of *Mbs.*



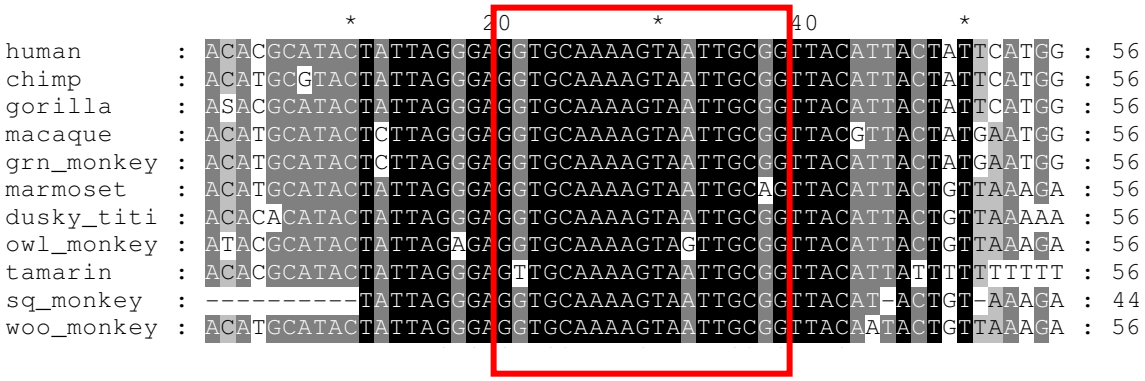


Figure 2.5 *Mbs* exhibiting orthologous conservation in all genomes examined. The human sequence in this example is identical to the consensus *mbs*. The sequence resides at chr5:163315209-163315227 in the human genome (UCSC Genome Browser, May 2004).

Table 2.4 Preliminary Observations of Genomic Environments of Broadly Conserved *Mbs*

Hs coordinates	genomic environment	ontology
chr1:39030260-39032338	0.2 kb upstream of BM558697	non-coding RNA
chr1:61657140-61659185	24 kb downstream of NFIA	cellular transcription factor; adenovirus replication factor
chr1:176410047-176412126	19 kb downstream of TDRD5	RNA-binding protein (provisional)
chr1:181526182-181528227	intron of C1orf24	unknown
chr4:101407286-101409365	gene desert	n/a
chr5:71350747-71352826	gene desert	n/a
chr5:163314202-163316274	gene desert	n/a
chr6:105061769-105063849	gene desert	n/a
chr6:111955049-111957102	multiple ESTs	unknown
chr6:130990858-130992912	gene desert	n/a
chr11:116326386-116328464	intron of KIAA0999	protein kinase
chr15:45036030-45038103	gene desert	n/a
chr16:23931067-23933183	intron of PRKCB1	cell signaling
chr18:21133587-21135666	intron of ZNF521	zinc-finger, nucleic acid binding

model. These results suggest that the same general model may be effectively applied to other conserved motif searches. The fact that the model provided a somewhat looser fit to the macaque control may be the result of sampling error in the estimation of neutral rates, caused by a small sample size and relatively low sequence divergence. In the marmoset analysis, the problem of small sample size is mitigated by a relatively high level of sequence divergence. In the composite NWM dataset, the problem of low divergence is offset by increased sample size by inclusion of sequences from multiple genomes at most loci. In the macaque, we confront both problems, a small sample and low divergence. Future work should include sequencing a number of loci from several OWM genomes, for shadowing against the macaque, in search of evidence of purifying selection on a subset of sequences in OWM.

Likewise, future work should focus on increasing resolution of the NWM data. The preliminary results for conservation in marmoset from a subset of *mbs* conserved in both human and macaque demonstrate the potential power of phylogenetic shadowing applied in a sequential culling process to facilitate the identification of a small subset of sequences which may be conserved by purifying selection within a much larger population of sequences which exhibit comparable levels of nucleotide conservation, but are evolving neutrally.

The observation of a greater frequency of *mbs* conserved in marmoset than expected, by chance, is consistent with the hypothesis that a subset of these sequences have been subject to purifying selection during anthropoid evolution. Our results suggest that most *mbs* in the human genome are evolving neutrally, with only a small proportion subject to constraint for a beneficial cellular function. The obvious question is: how many? The excess of *mbs* provisionally inferred based on the marmoset data is 33 (120 observed, 87 expected). Obviously, this estimation is influenced by modeling error and may be artificially high or low, although the results for the control data suggest that it is a reasonable estimate. Additionally, we note that our interrogation of the macaque genome only yielded orthologs for ~70% of queries, and the subsequent interrogation of the marmoset genome only produced hits for just under 60% of queries. (This may be partly due to the fact that neither of these genomes were

completely sequenced at the time of these interrogations.) Even if we assume that orthologs exist, in both macaque and marmoset, for all 1200 human *mbs*, and we extrapolate accordingly, we would hypothesize that the human genome harbors only about 80 *mbs* subject to purifying selection, out of ~1200. Given these proportions, we are not surprised that we have yet to detect an excess of *mbs* in our extension of phylogenetic shadowing into additional primate genomes. Among these 44 loci, there may be a dozen which are subject to purifying selection. There may be fewer. For a majority of loci in this subset, we have yet to obtain sufficient data, either to exclude them or to regard them as exhibiting especially broad conservation. The next phase of the *mbs* project should involve another round of PCR, using degenerate primers to increase output from NWM, especially. The addition of 2-3 more NWM species would also be useful. A larger and more complete dataset might also allow for a more subtle analysis, with expected levels of conservation calculated individually for each NWM branch, rather than in composite form as in this initial work. For simplicity, and to mitigate the problem of small samples, we have treated NWM in composite. NWM, however, are a diverse family, with substantial variation in generation times and, consequently, are likely to show variation in rates of neutral evolution (Kim et al. 2006).

The statistical analysis of molecular evolution of *mbs* is complementary to an experimental analysis of the putative function of *mbs* in vivo, as a binding site for SETMAR. Our preliminary work has already identified 14 sites as candidates for in vivo analysis, based on conservation in at least five anthropoids, including at least 3 NWM (not counting our initial 3-species screen), while exhibiting no more than 1 observed loss. Perhaps just as importantly, our screening process has eliminated 511 sites from consideration, including 7 loci from a reduced list of 44 candidates. (These 7, we excluded on the basis of 3 or more observed losses.) We note that recent findings from the ENCODE project suggest that a considerable proportion of protein-DNA interactions, even binding by transcription factors, are likely to produce no phenotype (Birney et al. 2007). Therefore, not only is the paring down of our dataset by phylogenetic shadowing necessary from a logistical standpoint, it provides a basis for inferring

purifying selection, and therefore biological significance, in a subset of *mbs* serving as in vivo binding sites for SETMAR. Filling in the remaining gaps in our dataset is likely to generate additional attractive subjects for in vivo analysis, as well as increased statistical power.

Furthermore, as we continue to refine our list of broadly conserved *mbs*, it may become possible to discover patterns in the genomic distribution of this subset of elements, with regard to various features of the genomic ecology, e.g., gene density, ontology of proximal genes, local recombination rates, transcription level and tissue specificity, etc. Such patterns, if observed, might suggest a concerted function exercised by the binding of SETMAR at multiple loci, perhaps as a conserved gene regulatory network.

The work described herein provides preliminary evidence for the activity of purifying selection to maintain a subset of mariner binding sites in at least some anthropoid lineages. In addition, it provides a rudimentary framework for an ongoing, and increasingly sophisticated, elucidation of the evolutionary history and, perhaps, contemporary function of *mbs* and SETMAR in the natural history of anthropoid primates.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990)
Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL, Deininger PL (1990) Structure and variability of recently inserted Alu family members. Nucleic Acids Res18:6793–6798

Baum M, Clarke L (2000) Fission yeast homologs of human CENP-B have redundant functions affecting cell growth and chromosome segregation. Mol Cell Biol 20:2852–2864.

Bejarano LA, Valdivia MM (1996) Molecular cloning of an intronless gene for the hamster centromere antigen CENP-B. Biochem Biophys Acta 1307:21–25.

Bird A. (1980) DNA methylation and the frequency of CpG in animal DNA.
Nucleic Acids Res 8: 1499–1504.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo´ R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET et al. (2007) Identification and analysis of functionalelements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-4

Brosius J, Gould SJ. (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci USA. 89:10706–10710.

Bulazel K, Metcalfe C, Ferreri GC, Yu J, Eldridge MD, O'Neill RJ. (2006) Cytogenetic and molecular evaluation of centromere-associated DNA sequences from a marsupial (Macropodidae: Macropus rufogriseus) X chromosome. Genetics 172:1129–1137.

Burkin DJ, Jones C, Burkin HR, McGrew JA, Broad TE (1996) Sheep CENPB and CENPC genes show a high level of sequence similarity and conserved synteny with their human homologs. Cytogenet Cell Genet 74:86–89.

Capy P, Bazin C, Higuet D, Langin T. Dynamics and evolution of transposable elements (1998) Austin (TX): Springer-Verlag.

Casola C, Hucks D, Feschotte C. (2008) Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Molecular Biology and Evolution 25: 29-41.

Charlesworth B, Sniegowski P, Stephan W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 212-220

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500.

Coelho PA, Nurminsky D, Hartl D, Sunkel CE (1996) Identification of Porto-1, a new repeated sequence that localises close to the centromere of chromosome 2 of Drosophila melanogaster. Chromosoma 105:211–222.

Comeron JM (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics 15:763–764.

Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc. Natl. Acad. Sci. USA 103:7941-42

Craig NL, Craigie R, Gellert M, Lambowitz,A M (2002) Mobile DNA II (Am. Soc. Microbiol, Washington, DC).

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator, Genome Research, 14:1188-1190

Doolittle WF, Sapienza C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603

Dou T, Gu S, Zhou Z, Ji C, Zeng L, Ye X, Xu J, Ying K, Xie Y, Mao Y (2004) Isolation and characterization of a Jerky and JRK/JH8 like gene, tigger transposable element derived 7, TIGD7. Biochem Genet 42:279–285.

Edwards NS, Murray AW (2005) Identification of xenopus CENP-A and an associated centromeric DNA repeat. Mol Biol Cell 16:1800–1810.

Feschotte C, Mouchès C (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol 17:730–737.

Feschotte C, Pritham EJ (2007) DNA transposons and the evoluiton of eukaryotic genomes. Annual Review of Genetics 41:331-68

Fowler KJ, Hudson DF, Salamonsen LA, Edmondson SR, Earle E, Sibson MC, Choo KH (2000) Uterine dysfunction and genetic modifiers in centromere protein B-deficient mice. Genome Res 10:30–41.

Goldberg IG, Sawhney H, Pluta AF, Warburton PE, Earnshaw WC (1996) Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA: implications for CENP-B function at centromeres. Mol Cell Biol 16:5156–5168.

Gould SJ, Vrba ES (1982) Exaptation- a missing term in the science of form. Paleobiology 8:4-15.

Haaf T, Mater AG, Wienberg J, Ward DC (1995) Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. J Mol Evol 41:487–491.

Hall TA (1999) BioEdit: a user-friendly biological alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98.

Halverson D, Baum M, Stryker J, Carbon J, Clarke L. (1997) A centromere DNA-binding protein from fission yeast affectschromosome segregation and has homology to human CENPB. J Cell Biol. 136:487–500

Hartl DL, Lozovskaya ER, Lawrence JG. (1992) Nonautonomous transposable elements and prokaryotes and eukaryotes. Genetica 86:47-53.

Henikoff S, Ahmad K, Malik HS. (2001) The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science* 293:1098 - 1102

Heslop-Harrison JS, Murata M, Ogura Y, Schwarzacher T, Motoyoshi F (1999) Polymorphisms and genomic organization of repetitive DNA from centromeric regions of Arabidopsis chromosomes. Plant Cell 11:31–42.

Hickey DA (1982) Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101:519-531

Hudson DF, Fowler KJ, Earle E, et al, (15 co-authors) (1998) Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol 141:309–319.

Irelan JT, Gutkin GI, Clarke L (2001) Functional redundancies, distinct localizations and interactions among three fission yeast homologs of centromere protein-B. Genetics 157:1191–1203.

Kapitonov VV, Jurka J (1999) Molecular paleontology of transposable elements from Arabidopsis thaliana. Genetica 107:27–37.

Kapoor M, Montes de Oca Luna R, Liu G, Lozano G, Cummings C, Mancini M, Ouspenski I, Brinkley BR, May GS (1998) The cenpB gene is not essential in mice. Chromosoma 107:570–576.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The UCSC Genome Browser Database. Nucl. Acids Res 31(1), 51-54.

Karolchik,D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D and Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucl. Acids Res. 32 (Suppl 1), D493-D496.

Kent WJ (2002) BLAT - The BLAST-Like Alignment Tool. Genome Res. 12(4), 656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler A M, Haussler D (2002) The Human Genome Browser at UCSC. Genome Res. 12(6), 996-1006.

Kim SH, Elango N, Warden C, Vigoda E, Yi SV (2006) Heterogeneous Genomic Molecular Clocks in Primates. PloS Genetics 2:1527-34

Kipling D, Warburton PE (1997) Centromeres, CENP-B and Tigger too. Trends Genet 13:141–145.

Kumar S (2005) Molecular clocks: Four decades of evolution. Nat Rev Genet 6: 654–662.

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A 99: 803–808.

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163.

Labuda D, Striker G (1989) Sequence conservation in Alu evolution. Nucleic Acids Res 17:2477–2491

Laird CD, McConaughy BL, McCarthy BJ (1969) Rate of fixation of nucleotide substitutions in evolution. Nature 224: 149–154.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Lee JK, Huberman JA, Hurwitz J (1997) Purification and characterization of a CENP-B homologue protein that binds to the centromeric K-type repeat DNA of Schizosaccharomyces pombe. Proc Natl Acad Sci USA. 94:8427–8432.

Lee, S.-H., Oshige, M., Durant, S. T., Rasila, K. K., Williamson, E. A., Ramsey, H., Kwan, L., Nickoloff, J. A. & Hromas, R. (2005) The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair Proc. Natl. Acad. Sci. USA 102, 18075–18080.

Li WH (1997) Molecular Evolution. Sunderland (Massachusetts): Sinauer.

LI, WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36: 96–99

Li WH, Ellsworth DL, Krushkal J, Chang BHJ, Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5: 182–187.

Lopez CC, Edstrom JE (1998) Interspersed centromeric element with a CENP-B box-like motif in Chironomus pallidivittatus. Nucleic Acids Res 26:4168–4172.

Lorite P, Carrillo JA, Tinaut A, Palomeque T (2004) Evolutionary dynamics of satellite DNA in species of the genus Formica (Hymenoptera, Formicidae). Gene 332:159–168.

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol 109:1963–1973.

Miller WJ, McDonald JF, Nouaud D, Anxolabehere D (1999) Molecular domestication—more than a sporadic episode in evolution. Genetica 107:197–207.

Mravinac B, Plohl M, Ugarkovic D (2004) Conserved patterns in the evolution of Tribolium satellite DNAs. Gene 332:169–177.

Murakami Y, Huberman JA, Hurwitz J (1996) Identification, purification, and molecular cloning of autonomously replicating sequence-binding protein 1 from fission yeast Schizosaccharomyces pombe. Proc Natl Acad Sci USA. 93:502–507.

Nakagawa H, Lee JK, Hurwitz J, Allshire RC, Nakayama J, Grewal SI, Tanaka K, Murakami Y (2002) Fission yeast CENP-B homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications. Genes Dev 16:1766–1778.

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. New York: Oxford University Press.

Ngan VK, Clarke L. (1997) The centromere enhancer mediates centromere activation in Schizosaccharomyces pombe. MolCell Biol. 17:3305–3314

Nicholas KB, Nicholas HB, Jr. (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors. http://www.psc.edu/biomed/genedoc

Nonomura KI, Kurata N (1999) Organization of the 1.9-kb repeat unit RCE1 in the centromeric region of rice chromosomes. Mol Gen Genet 261:1–10.

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284: 604-607

Pace JK II, Feschotte C (2007) The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Research 17:422-32

Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281.

Perez-Castro AV, Shamanski FL, Meneses JJ, Lovato TL, Vogel KG, Moyzis RK, Pedersen R (1998) Centromeric protein B null mice are viable with no apparent abnormalities. Dev Biol 201:135–143.

Plasterk RHA, Izsvák Z, Ivics Z (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 15:326–332.

Robertson HM (1996) Members of the pogo superfamily of DNA-mediated transposons in the human genome. Mol Gen Genet 252:761–766.

Robertson HM, Zumpano KL (1997) Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. Gene 205, 203–217.

Schneider TD, Stephens RM (1990) Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18:6097-6100

Smit AFA, Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA. 93:1443–1448.

Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, Iwahara J, Okazaki T, Yokoyama S (2001) Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J 20:6612–6618.

Tawaramoto MS, Park SY, Tanaka Y, Nureki O, Kurumizaka H, Yokoyama S (2003) Crystal Structure of the Human Centromere Protein B (CENP-B) Dimerization Domain at 1.65-Å Resolution. J. Biol. Chem., Vol. 278, Issue 51, 51454-51461

Toth M, Grimsby J, Buzsaki G, Donovan GP (1995) Epileptic seizures caused by inactivation of a novel gene, jerky, related to centromere binding protein-B in transgenic mice. Nat Genet 11:71–75.

Weide R, Hontelez J, van Kammen A, Koornneef M, Zabel P (1998) Paracentromeric sequences on tomato chromosome 6 show homology to human satellite III and to the mammalian CENP-B binding box. Mol Gen Genet 259:190–197.

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

Yi S, Ellsworth DL, Li WH (2002) Slow molecular clocks in Old World monkeys, apes, and humans. Mol Biol Evol 19: 2191–2198.

Yoda K, Kitagawa K, Matsumoto H, Muro Y, Okazaki T (1992) A human centromere protein, CENP-B, has a DNA binding domain containing four potential alpha helices at the NH2 terminus, which is separable from dimerizing activity. J Cell Biol 119:1413–1427.

Yoda K, Nakamura T, Masumoto H, Suzuki N, Kitagawa K, Nakano M, Shinjo A, Okazaki T (1996) Centromere protein B of African green monkey cells: gene structure, cellular expression, and centromeric localization. Mol Cell Biol 16:5169–5177.

Zeng Z, Kyaw H, Gakenheimer KR, Augustus M, Fan P, Zhang X, Su K, Carter KC, Li Y (1997) Cloning, mapping, and tissue distribution of a human homologue of the mouse jerky gene product. Biochem Biophys Res Commun 236:389–395.

BIOGRAPHICAL INFORMATION

Donald Hucks received a B.S. in Biology from the University of Texas at Arlington in 2005.