DOI: 10.1534/genetics.111.126813

Phylogenetic and Functional Characterization of the *hAT* Transposon Superfamily

Peter Arensburger,* Robert H. Hice,* Liqin Zhou,† Ryan C. Smith,* Ariane C. Tom,* Jennifer A. Wright,* Joshua Knapp,* David A. O'Brochta,‡ Nancy L. Craig† and Peter W. Atkinson*,[1]

**Department of Entomology and Institute of Integrative Genomic Biology, University of California, Riverside, California 92521-0001,* †*Howard Hughes Medical Institute, Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205 and* ‡*Department of Entomology and Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, Maryland 20742*

Manuscript received January 14, 2011
Accepted for publication February 17, 2011

ABSTRACT

Transposons are found in virtually all organisms and play fundamental roles in genome evolution. They can also acquire new functions in the host organism and some have been developed as incisive genetic tools for transformation and mutagenesis. The *hAT* transposon superfamily contains members from the plant and animal kingdoms, some of which are active when introduced into new host organisms. We have identified two new active *hAT* transposons, *AeBuster1*, from the mosquito *Aedes aegypti* and *TcBuster* from the red flour beetle *Tribolium castaneum*. Activity of both transposons is illustrated by excision and transposition assays performed in *Drosophila melanogaster* and *Ae. aegypti* and by *in vitro* strand transfer assays. These two active insect transposons are more closely related to the *Buster* sequences identified in humans than they are to the previously identified active *hAT* transposons, *Ac*, *Tam3*, *Tol2*, *hobo*, and *Hermes*. We therefore reexamined the structural and functional relationships of *hAT* and *hAT*-like transposase sequences extracted from genome databases and found that the *hAT* superfamily is divided into at least two families. This division is supported by a difference in target-site selections generated by active transposons of each family. We name these families the *Ac* and *Buster* families after the first identified transposon or transposon-like sequence in each. We find that the recently discovered *SPIN* transposons of mammals are located within the family of *Buster* elements.

TRANSPOSONS are fundamental agents of genome evolution and can acquire functions independent of transposition in genomes (Britten 2005; Britten 2006; Volff 2006). The abundance of whole-genome sequence data has greatly increased our ability to identify and characterize the complete complement of transposons within genomes, leading to a deeper understanding of the origins of transposons and their dynamic relationships within the genomes they inhabit. Transposons are classified as either RNA transposons that transpose through an RNA intermediate or as DNA transposons that transpose via a DNA intermediate (Finnegan 1990).

DNA transposons in the *hAT* superfamily are widespread in plants and animals and include a number of active, well-studied elements such as the *Ac* transposon of *Zea mays*, the *hobo* transposon of *Drosophila melanogaster*, the *Hermes* transposon of the housefly *Musca domestica*, and the *Tol2* transposon of the Japanese Medaka fish, *Oryzias latipes* (McClintock 1950; Blackman *et al.* 1989; O'Brochta *et al.* 1996; Kawakami and Shima 1999). *hAT* transposons are also found in mammalian genomes including humans where they are the most abundant DNA transposons and comprise 1.55% (195 Mb) of the total genome (Lander *et al.* 2001). None of the *hAT* elements in the human genome, however, are known to have been active during the past 50 million years (Lander *et al.* 2001). In contrast, in the genome of the little brown bat, *Myotis lucifugus*, there has been a marked expansion of DNA transposons with *hAT* transposons being the most abundant and recent component of this expansion (Ray *et al.* 2008). No active *hAT* transposons have been recovered from the genome of *M. lucifugus*, but the abundance of these and other DNA transposons indicates that some mammals have retained the ability to support high DNA transposon activity (Ray *et al.* 2008).

SPIN (space invaders) transposons are *hAT* transposons that have been found in opossum, bushbaby, tenrec, and little brown bat as well as in the African clawed frog, the green anole lizard, murine rodents, and a

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.111.126813/DC1.

[1]*Corresponding author:* Institute of Integrative Genome Biology, University of California, Riverside, CA 92521.
E-mail: peter.atkinson@ucr.edu

DATA ACCESSION NUMBERS

Protein sequences used in the article were collected from several databases and one manuscript. These are listed below as well as sequence names as they appear in Figure 1. Accession numbers for the appropriate database are show in parentheses if they differ from the sequence name. The sequence SPIN_Ml_a is a variant of the SPIN_Ml element, and the sequence of SPIN_Ml_a is available upon request.

GENBANK (http://www.ncbi.nlm.nih.gov)

Ac-like (AAC46515), Ac (CAA29005), AeBuster1 (ABF20543), AeBuster2 (ABF20544), AmBuster1 (EFB22616), AmBuster2(EFB25016), AmBuster3 (EFB20710), AmBuster4 (EFB22020), BtBuster1 (ABF22695), BtBuster2 (ABF22700), BtBuster3 (ABF22697), CfBuster1 (ABF22696), CfBuster2 (ABF22701), CfBuster3 (XP_854762), CfBuster4 (XP_545451), CsBuster (ABF20548), Daysleeper (CAB68118), DrBuster1 (ABF20549), DrBuster2 (ABF20550), EcBuster1 (XP_001504971), EcBuster3(XP_001503499), EcBuster4 (XP_001504928), Hermes (AAC37217>), hermit (LCU22467), Herves (AAS21248), hobo (A39652), Homer (AAD03082), hopper-we (AAL93203), HsBuster1 (AAF18454), HsBuster2 (ABF22698), HsBuster3 (NP_071373), HsBuster4 (AAS01734), IpTip100 (BAA36225), MamBuster2 (XP_001108973), MamBuster3 (XP_001084430), MamBuster3 (XP_001084430), MamBuster4 (XP_001101327), MmBuster2 (AAF18453), PtBuster2 (ABF22699), PtBuster3 (XP_001142453), PtBuster4 (XP_527300), Restless (CAA93759), RnBuster2 (NP_001102151), SpBuster1 (ABF20546), SpBuster2 (ABF20547), SsBuster4 (XP_001929194), Tam3 (CAA38906), TcBuster (ABF20545), Tol2 (BAA87039), tramp (CAA76545), XtBuster (ABF20551)

ENSEMBL (http://www.ensembl.org)

PtBuster1 (ENSPTRG00000003364)

REPBASE (http://www.girinst.org)

Ac-like2 (hAT-7_DR), Ac-like1 (hAT-6_DR), hAT-5_DR, MlBuster1 (hAT-4_ML), Myotis-hAT1, SPIN_Et, SPIN_Ml, SPIN-Og

TEFam (tefam.biochem.vt.edu)

AeHermes1 (TF0013337), AeBuster3 (TF001186), AeBuster4 (TF001187), AeBuster5 (TF001188), AeBuster7 (TF001336), AeHermes2 (TF0013338), AeTip100-2 (TF000910), Cx-Kink2 (TF001637), Cx-Kink3 (TF001638), Cx-Kink4 (TF001639), Cx-Kink5 (TF001640), Cx-Kink7 (TF001636), Cx-Kink8 (TF001635)

Pace *et al.* (2008)

SPIN_Md, SPIN_Xt

triatomid bug (Pace *et al.* 2008; Gilbert *et al.* 2010). In contrast *SPIN* transposons are absent from cats, dogs, cattle, armadillo, treeshrew, and humans (Pace *et al.* 2008). Their discontinuous distribution combined with their sequence conservation and abundance in some species has led to the hypothesis that *SPIN* transposons have been repeatedly horizontally transferred within deuterostomes; however, as yet, no active native *SPIN* transposon has been found, nor has the likely vector of these horizontal transfers been identified, although host–parasite relationships have been invoked as possible platforms for enabling this to occur (Pace *et al.* 2008; Gilbert *et al.* 2010*et al.*). We show here that *SPINs* are closely related to likely exapted and highly conserved proteins present in some mammalian species.

The *hAT* superfamily of DNA transposons also provides a number of examples of elements being exapted to essential functions within the host genome, a process sometimes referred to as domestication. Sinzelle *et al.* (2009) list 10 examples of domesticated *hAT* transposons with 8 of these retaining both their DNA binding and catalytic domains. These include the DAYSLEEPER gene from *Arabidopsis thaliana*, the *GARY* genes from cereal grasses, the *DREF* gene from *D. melanogaster*, the *GON-14* and *LIN-15B* genes of *Caenorhabditis elegans*, and the *TRAMP* and *Buster* genes from *Homo sapiens* (Sinzelle *et al.* 2009). Feschotte and Pritham (2007) list 11 examples of *hAT* transposon domestication with 3 of these possibly also involving the codomestication of *P*-element-like sequences. *hAT* transposon sequences are present in the human genome with the *Charlie/MER1 hAT* elements accounting for approximately 90% of *hAT* sequences; however, an active *hAT* transposase consensus sequence for these has not been clearly identified, suggesting that these may be inactive, "dead" transposons (Smit and Riggs 1996; Lander *et al.* 2001). Similarities in the mechanism of excision and the generation of circular DNA products, or episomes, between the *Hermes hAT transposase* and the human RAG1 recombinase indicate that both share a common mechanism of transposition (Zhou *et al.* 2004; O'brochta *et al.* 2009). However, comparisons of the primary structures of the RAG1 transposase with all eukaryotic transposases suggests that these are more closely related to the *Transib* transposases (Zhou *et al.* 2004; Kapitonov and Jurka 2005; O'Brochta *et al.* 2009).

hAT transposons and related domesticated sequences are thus a very large superfamily both in their diversity and in their absolute numbers and there is an increasing interest in the role that this superfamily has played in the evolution of the plant and animal kingdoms. A handful of different active members of this superfamily have been discovered; indeed *Ac*, the first transposon to be identified, is a *hAT* transposon discovered by McClintock on the basis of its mutagenic ability in maize (McClintock 1950). Much remains to be learned about the activity and regulation of *hAT* elements, particularly when active forms are introduced into new hosts.

Employing a bioinformatic methodology that was previously successful in identifying the active *Herves hAT* transposon from the mosquito *Anopheles gambiae*

(Arensburger *et al.* 2005), we sought to determine whether new active *hAT* transposons were present in two recently sequenced insect genomes—those of the mosquito *Aedes aegypti* and the red flour beetle, *Tribolium castaneum* (Nene *et al.* 2007; Richards *et al.* 2008). We identified two new active *hAT* transposons, *AeBuster1* from *Ae. aegypti* and *TcBuster* from *T. castaneum*. Sequence comparison with existing *hAT* transposons and *hAT*-like sequences showed that both were surprisingly closely related to the *Buster* genes of deuterostomes yet were clearly active transposons. We demonstrate that both *AeBuster1* and *TcBuster* are active in interplasmid transposition assays performed in *D. melangaster* and *Ae. aegypti* embryos. Transposition of these transposons into target DNA generates 8-bp target-site duplications as is true of other *hAT* elements but with a consensus sequence consistently different to those generated by the *Hermes* and *hobo* transposons. These are the first *Buster* transposons shown to be active.

The activity of both *AeBuster1* and *TcBuster* led us to reexamine the *hAT* transposon superfamily phylogeny. Previous studies have focused only on those transposons and sequences closely related to the *Ac*, *hobo*, *Tol2*, and *Hermes* elements and concluded that, while there was evidence that this superfamily is very ancient, the grouping of these elements into clusters consistent with the evolution of their hosts suggested that horizontal transfer had not played a role in their evolution (Simmons 1992; Kempken and Windhofer 2001; Rubin *et al.* 2001). However, a more recent analysis of *hAT* elements detected in 12 Drosophila species concluded that four clades (or families) of *hAT* elements could be identified (Ortiz and Loreto 2009). Furthermore, this phylogeny was not completely congruent with the phylogeny of the host species, indicating a possible role for horizontal transfer in the distribution of some of these *hAT* transposons, a possibility that was originally conceived when the distribution of *hobo* elements was first analyzed by DNA sequencing (Ortiz and Loreto 2009). We find that, on the basis of the primary sequence of their transposases, the *hAT* superfamily is composed of at least two families that we classify as the *Ac* family and the *Buster* family, named after the first elements identified in each. The structural distinction between the *Ac* and *Buster* families is further supported by the functional difference in target-site selection of members of each upon insertion into target DNA. We find that the recently discovered *SPIN* elements are members of the *Buster* family.

MATERIALS AND METHODS

Discovery of novel transposable elements and mammalian exapted genes: We refined the bioinformatic methods previously used to identify the active *Herves hAT* element from the mosquito *A. gambiae* (Arensburger *et al.* 2005). In brief, whole-genome sequences from *Ae. aegypti* (Liverpool) and *T. castaneum* (Georgia 2) were searched for regions showing similarity to known *hAT* transposase sequences (transposase sequences collected from the following databases: Repbase http://www.girinst.org/, Tefam http://tefam.biochem.vt.edu/tefam/, and NCBI http://www.ncbi.nih.gov), using the TBLASTN program (cutoff value $e < 10^{-4}$) (Nene *et al.* 2007; Richards *et al.* 2008). Coordinates of all identified regions were used as input for a set of custom written PERL scripts for further analysis. These scripts searched each coordinate region and surrounding sequences for nucleotide regions with the following characteristics: (1) presence of an intact ORF, (2) 8-bp target-site duplications (TSDs), (3) terminal inverted repeats (TIRs) ≥10 bp in size immediately adjacent and internal to the TSDs, and (4) the same transposon flanked by different TSDs in the genome and so most likely present at multiple loci. The output of these scripts was manually reviewed. Scripts are available upon request.

In the course of this search a surprising similarity was observed between the *AeBuster1* and *TcBuster* transposase sequences and translated regions of the human genome. These similarities were further investigated using various BLAST programs, leading to the discovery of four putative exapted mammalian genes (described below).

Consensus TSD sequences: Target-site sequence preferences of 299 putatively full-length *hAT* transposons were examined, using the following criteria for inclusion into the analysis. The Repbase database was queried for all *hAT* repeat class entries that had a predicted transposase sequence and belonged to a species with an available whole-genome sequence; this resulted in the inclusion of 273 elements (supporting information, Table S1). Two additional *SPIN* elements not included in Repbase but previously described were also added (Pace *et al.* 2008). A total of 24 additional transposons discovered during the course of this analysis were also included.

All potential 8-bp TSD sequences were identified by repeating the following steps for each transposon. First, the genome of origin was searched using the BLAST algorithm (cutoff value $e < 10^{-4}$) for any region that matched the first 50 bp of the transposon. The genome was then searched again for matches to the last 50 bp of the transposon using the same method. Second, using these matching regions, the 8 bp immediately adjacent and outside of the transposon sequence was identified and stored. Third, to prevent overrepresentation of elements that were likely duplicated through a nontransposase-mediated mechanism, any duplicate 8-bp sequence was removed, leaving only a single representative sequence for each potential TSD. Finally, a 50% consensus sequence was generated.

Sequence selection, alignment, and phylogenetic analysis: Three amino acid transposase sequence data sets were created and used for phylogenetic analysis. The first database consisted of 50 transposons selected (on the basis of a survey of prior publications) to represent the phylogenetic diversity of the *hAT* superfamilly. Emphasis for selection was placed on full-length and potentially active transposons rather than on consensus sequences. The number of taxa included in this data set was limited to allow more in-depth phylogenetic analysis (see below). A second database of 78 sequences included the 50 previously identified sequences as well as the amino acid sequences of 28 exapted genes with known sequence similarity to *hAT* transposons (Table S2). These exapted gene sequences were identified from the literature and as part of this study (see below). The third database of 192 transposase sequences was created by using the transposons used for the consensus TSD analysis (see above), selecting those sequences that had at least 10 identified, unique TSD sequences.

The 50-sequence data set was aligned using the program M-Coffee, which computes a consensus alignment from several

multiple sequence alignment programs (PCMA, mafft, clustalw, dialign, poa, muscle, probsons, and T-Coffee) (Moretti *et al.* 2007). The remaining two data sets were aligned using the multiple sequence alignment program T-Coffee (Notredame 2010). Furthermore, this 50-sequence data set was used as input for the program ProtTest (Abascal *et al.* 2005) to identify the best-suited amino acid substitution matrix for these data, which was WAG (Whelan and Goldman 2001) with among-site rate variation accommodated by a gamma shape parameter and invariant sites. On the basis of these results, two phylogenetic trees were created using this data set and amino acid substitution matrix: (1) a tree based on the maximum-likelihood optimality criterion using the program TREE-PUZZLE v. 5.2 (Schmidt *et al.* 2002), and (2) a tree based on a Bayesian estimation of phylogeny using the program MRBAYES (Ronquist and Huelsenbeck 2003). Phylogenetic trees of the other two data sets were constructed using the program FastTree2 using default parameters (Price *et al.* 2010).

Clones and plasmids: *pBSAeBusterLR: AeBuster1* was amplified in sections from *Ae. aegypti* (Orlando). Genomic DNA was purified using a Wizard Genomic DNA Kit (Promega), and 360 ng of DNA was used as template in 50-μl PCR reactions using a TripleMaster PCR System (Eppendorf). The pGT-AeBusterL2 clone was made by PCR amplification using primers that encompassed the region from the TSD of one copy of *AeBuster1* to the sequence immediately upstream of the ATG of the *AeBuster* ORF. The primers used were AeBuster L2 For: 5′-AATGGTACCGCTTATGG**CATAGATTCCCAAAC T**GTG-3′ (TIR in boldface type), and AeBuster L Rev: 5′-GATCTCGAGATCTGAAATTATCAAATAATGAATCGCATA TTCTG-3′, with the following PCR program: 94° 2′, 4 × (94° 20″, 60° 20″, 72° 30″), 25 × (94° 20″, 69° 20″, 72° 30″), 72° 5′, 4°. Following amplification, the DNA was purified using a Qiaquick PCR purification kit (Qiagen), quantified on an agarose gel, and ligated into the pGEM-T Easy vector (Promega). Inserts were sequenced, and then clones were digested with *Kpn*I and *Xho*I (New England Biolabs). Gel-purified fragments (Zymoclean Gel DNA Recovery Kit, Zymo Research) were cloned into pBluescript SK+ digested with the same enzymes to make the clone pBSAeBL2. The right end of *AeBuster1* was amplified in a similar manner using the primers AeBuster R For: 5′-GATTCTAGATGCGCATCGAACAACATT TTTAGTGAG-3′ and AeBuster R2 Rev: 5′-AATGAGCTCCCA TAAGC**CATAGGTTCCCAAACT**TTTC-3′ (TIR in boldface type), which encompassed the region immediately 3′ of the stop codon through the target-site duplication. The PCR program was: 94° 2′, 4 × (94° 20″, 60° 20″, 72° 30″), 25 × (94° 20″, 72° 30″), 72° 5′, 4°. The right end PCR product was cloned as above, first into pGEM-T Easy and then into the left end clone following digestion with *Sac*I and *Xba*I (New England Biolabs, NEB) to yield the clone pBSAeBusterLR.

pBSAeBuster donor 1: *Sma*I-digested pBSAeBLR was ligated with *Xmn*I-digested pGENToriAlpha, which was derived from pK19 and in which the kanamycin^R gene was replaced with the gentamycin^R gene from pFastBac HTb (Invitrogen) to generate a donor element, which has a replication origin, a gentamycin^R gene, and a lacZ-alpha gene.

pKhsp70AeBuster1: For cloning the *AeBuster1* ORF, PCR was carried out as above, but with the primers AeBuster ORF For, 5′-GATGATATCAGATATGGATAAATGGTTGTTGAAGA AGC-3′, and Ae Buster ORF Rev, 5′-GATGATATCTTAGTGT GATGGATGGCTTGCTTG-3′. The PCR program was: 94° 2′, 3 × (94° 20″, 63° 20″, 72° 2′), 25 × (94° 20″, 68° 20″, 72° 2′), 72° 5′, 4°. DNA was purified as above, ligated into pGEM-T Easy, and sequenced. One clone, pGEMT-AeBuster clone 9 ORF, contained a sequence with two silent mutations, but no differences in protein sequence from *AeBuster1* transposase copies 1 and 2, and was used for the helper plasmid containing the *AeBuster1* transposase. pGEMT-AeBuster clone 9 ORF was digested with *Eco*RV (NEB) and cloned into the *Sma*I site of pKhsp70 to make the helper pKhsp70AeBuster1.

pBADAeBuster1: pBAD/Myc-His A (Invitrogen) digested with *Nco*I and *Apa*I (NEB) was ligated to a PCR product (digested with *Nco*I and *Apa*I) amplified as above using pGEMT-AeBuster clone 9 ORF as the template. The primers used were AeBuster coli F, 5′-GATCCATGGATAAATGGTTG TTGAAGAAGCCC-3′, and AeBuster coli R (AATGGGCCCGT GTGATGGATGGCTTTGCTTG). The same PCR program was used as for the ORF amplification described above. Purification of the His-tagged *AeBuster1* transposase was performed using the same protocol as described for the His-tagged *Hermes* transposase (Zhou *et al.* 2004).

pBSTcBuster LR1 and *pBSTcBusterLR2: TcBuster* was amplified from *T. castanaeum* Ga-2 DNA (obtained from Dr. Susan Brown, Kansas State University, Manhattan, KS). Approximately 30 ng of DNA was used as template in 50 μl PCR reactions using the TripleMaster PCR System (Eppendorf). The clone pBSTcBusterL was made by PCR amplification using primers that encompassed the region from the TSD of *TcBuster1* to the sequence immediately upstream of the ATG of the *TcBuster1* ORF. The primers used were *TcBuster1* L For (AATGGTACCCTTTAGGC**CAGTGTTCTTCAACCT**G (TIR in boldface type) and *TcBuster1* L Rev (CATCTCGAGAT TTCTGAACGATTCTAGGTTAGGATCAAAC) with the following PCR program: 94° 2 min, 3 × (94° for 20 sec, 62° for 20 sec, 72° for 30 sec), 26 × (94° for 20 sec, 70° for 20 sec, 72° for 20 sec), 72° for 5 min, 4°. Following amplification the DNA was purified using the QIAquick PCR Purification Kit (Qiagen), quantified on an agarose gel, digested with *Kpn*I and *Xho*I (NEB), repurified, and ligated into pBluescript SK+ (Stratagene), which had been digested with *Kpn*I and *Xho*I and gel purified (Zymoclean Gel DNA Recovery Kit, Zymo Research). Inserts were sequenced, and then clones were digested with *Sac*I and *Xba*I (NEB). Gel-purified vector was ligated to *TcBuster1* right end PCR product and digested with *Sac*I and *Xba*I.

The right end of *TcBuster1* was amplified in a similar manner using the primers *TcBuster1* R1 For (GATTCTAGAC CAAAGCACGGGCTCACCTTGTTC) or *TcBuster1* R2 for (GAT TCTAGACAACTGATCCATCCCGATATTGATAATTTGTGC) and *TcBuster1* R Rev (AATGAGCTCGTATAAGCAGTGTTTT CAACCTTTGCCATCC), using the PCR program described above. The R1 end contained only 145 bp of the transposon, while the R2 end contained 313 bp of transposon, and included some ORF sequence. Following amplification, the PCR products were purified, digested with *Sac*I and *Xba*I, quantified, and ligated to the left end clone to make clones pBSTcBusterLR1 and pBSTcBusterLR2.

pTcBDG2 and pTcBDG3: Donor plasmids pTcBDG2 and pTcBDG3 were made by ligation of *Xba*I-digested pBSTcBusterLR1 or pBSTcBusterLR2 with *Nhe*I-digested pGENToriAlpha.

pKhsp70TcBuster1: The *TcBuster* helper plasmid was made as follows: PCR was performed with the primers *TcBuster1* ORF For (AATGATATCAGAAATATGATGCTGAATTGGCTCAAAA GTGG) and *TcBuster1* ORF Rev (GATGATATCTTAATGAC TTTTTTGCGCTTGCTTATTATTGCAC). The PCR program was: 94° for 2 min, 3 × (94° for 20 sec, 67.5° for 20 sec, 72° for 2 min), 26 × (94° for 20 sec, 70° for 20 sec, 72° for 2 min), 72° for 5 min, 4°.

PCR product was purified as above, digested with *Eco*RV, cloned into the plasmid pKhsp70 (Arensburger *et al.* 2005), and digested with *Sma*I to generate pKHsp70TcBuster1. The sequence of the ORF was identical to the sequence from the bioinformatic analysis described above except for a single base change that resulted in a Ser–Thr substitution at aa 551. Thr

at this position was found in all of the clones sequenced for the *TcBuster1* ORF.

pGDV1: Is a *Bacillus subtilus* plasmid incapable of replicating in *Escherichia coli* and used as a target plasmid in plasmid-based transposition assays (Sarkar *et al.* 1997a).

Plasmid-based transposition and excision assays: A mixture of three plasmids, "donor" (pBSAeBuster-GenOriLacZ or pBSTcBuster-GenOriLacZ), "target" (pGDV1), and "helper" (pKhsp70AeBuster1 or pKhsp70TcBuster1) were introduced into insect cells either by transfection of Schneider 2 cells or by microinjection into *D. melanogaster* and *Ae. aegypti* preblastoderm embryos as previously described (Arensburger *et al.* 2005; Sarkar *et al.* 1997b). Transfection of donor, target, and helper (or negative control plasmids pK19 or Ac5cEGFP) into Schneider 2 cells was performed in six-well plates as previously described (Arensburger *et al.* 2005). For some experiments the quantity of donor was decreased and the helper increased (1.25 and 3.75 μg). Plasmids were recovered with the Wizard Genomic DNA kit and used to transform DH10 cells by electroporation. Transposition of donor elements into the target plasmids resulted in recombinant plasmids expressing resistance genes for gentamycin and chloramphenicol. To select for transformed DH10 cells with recombinant plasmids resulting from transposition-electroporated cells were plated on LB plates containing gentamycin (7 μg/ml) and chloramphenicol (10 μg/ml). After 3 days incubation at 37°, resistant colonies were picked and grown in LB media containing only gentamycin. Plasmid DNA was purified from these cells using the Wizard Plus miniprep kit (Promega). The presence of a recombinant plasmid arising from transposition was verified by digesting the plasmid DNA with *Eco*RV to check for a diagnostic pattern of bands (1.1, 1.5, and 3.2 kb). Plasmids passing this initial test were confirmed as transposition events by DNA sequencing.

Alternatively, A*eBuster1* excision assays were performed on plasmids recovered from transposition assays performed in *Ae. aegypti* embryos. Recovered plasmid DNA was digested with *Eco*RV and the resulting DNA was used to transform DH10 cells by electroporation followed by selection on LB plates containing ampicillin and X-gal (20 mg/liter). Because *Eco*RV cuts only within the *AeBuster1* donor element, plasmids arising as a result of excision are resistant to *Eco*RV linearization. Uncut excision products efficiently transform *E. coli* while linearized donor plasmids do not. Putative excision events were confirmed by restriction digestion and DNA sequencing. Ampicillin^R, LacZ$^-$ colonies were selected, mapped and then sequenced across the empty excision site using the primers 5′-CGTCCCATTCGCCATTCAGG-3′.

Cell-free transposition reactions: The methods used here are essentially identical to those described for a number of other elements (Zhou *et al.* 2004). Short (40 bp) DNA oligonucleotides containing the terminal sequences of the left and right ends of *AeBuster1* were radiolabeled using T4 kinase and [γ-P^{32}]ATP. Labeled oligonucleotides were annealed to their unlabeled complementary strands in 10 mM Tris-HCl, pH 7.5, by heating at 90° for 5 min and then cooling the reaction at room temperature. The annealed left and right ends were used directly in strand-transfer reactions. The *AeBuster1* L-40mers were 5′-**CATAGATTCCCAAACT**GTGGGTCGCGA CCCCCTGGGGGGT and 3′-**GTATCTAAGGGTTTGA**CAC CCAGCGCTGGGGGACCCCCCA. The *AeBuster1* R-40mers were 5′-**CATAGGTTCCCAAACT**TTTCGGATGCGCGACCCC CCTAGC and 3′-**GTATCCAAGGGTTTGA**AAAGCCTACGCG CTGGGGGGATCG. TIRs are in boldface type. Strand-transfer reactions contained 400 nM *AeBuster1* transposase and 10 nM *AeBuster1* ends in 25 mM MOPS pH 7.6, 5% (v/v) glycerol, 2 mM DTT, 10 mM $MgCl_2$, 100 μg/ml BSA, and 10 nM pUC19 plasmid as target DNA in a final volume of 20 μl. The reaction mixtures were incubated 30° for 2 hr, stopped by adding SDS and EDTA to 1% SDS and 20 mM EDTA and incubated for 1 hr at 37°. The products were displayed on 1% agarose/TAE gels.

RESULTS

The *hAT* superfamily consists of at least two families of transposons: We used an algorithm that searched the *Ae. aegypti* and *T. castaneum* genomes for *hAT* transposons on the basis of the presence of TIRs ≥10 bp in length and 8 bp TSDs flanking a sequence with at least one ORF that has similarity to *hAT* transposases (Arensburger *et al.* 2005). The conceptual ORF translation of these new elements, as well as selected previously described *hAT* transposase sequences, was used to construct an unrooted maximum-likelihood transposase tree that revealed two major clusters of related sequences shown in the gray regions of Figure 1 that we have designated the *Ac* family and the *Buster* family. The *hAT* transposase sequences used in this analysis were selected on the basis of the phylogenetic diversity of the superfamily and on their annotation as full-length sequences, thus increasing the likelihood that they are, or recently were, mobile in their host organism. A second tree using the same transposase sequences but using an optimality criterion based on Bayesian inference rather than maximum likelihood was consistent with the tree shown in Figure 1 (the Bayesian tree shown in Figure S1). The relationship between these *hAT* transposases and known related exapted genes is also shown (Feschotte and Pritham 2007; Sinzelle *et al.* 2009) (Figure S2).

The majority of the amino acid sequence variation contributing to this distribution was found to reside in two domains of these transposases. We used the crystal structure of the *Hermes* transposase and T-coffee alignments of *hAT* and *Buster* transposases to reexamine the domain structure of these transposases (Figure S3). We found that both families of *hAT* transposases have four domains, an N-terminal domain containing a BED zinc finger, a second DNA-binding domain that, at least in the *Hermes* transposase, also plays a role in oligomerization, a catalytic domain containing the first two carboxylates of the catalytic triad, and then a long insertion domain that contains numerous α-helices (Zhou *et al.* 2004; Hickman *et al.* 2005) (Figure S3 and Figure S4). Then follows an α-helix containing a glutamate residue that is in close proximity to the catalytic domain and so completes the catalytic DDE triad (Hickman *et al.* 2005). This final region has been termed the *hAT* domain on the basis of its conservation between *hAT* transposases and its role in oligomerization; however, the crystal structure of the *Hermes* transposase indicates that multiple regions of this protein contribute to oligomerization (Hickman *et al.* 2005). The most significant variations among members of the *Ac* and *Buster* families lie in the second DNA-binding domain and in the insertion domain.

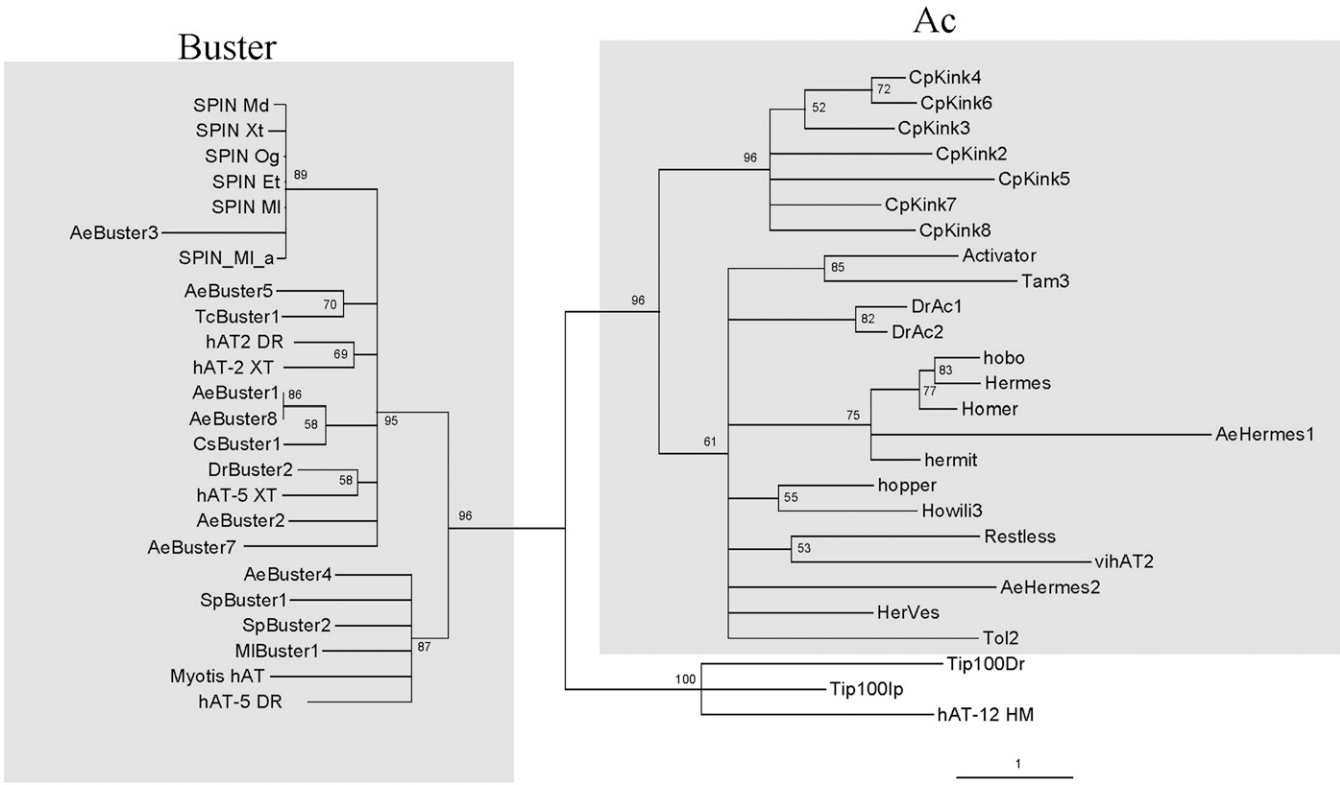

FIGURE 1.—Consensus phylogenetic tree showing the relationship of amino acid transposase sequences, between 50 selected full-length *hAT* elements based on a maximum-likelihood optimality criterion (50% majority rule consensus). Numbers next to most nodes show quartet puzzling reliability based on 10,000 puzzling steps, a measure of nodal support similar to bootstrapping, produced by the program TREE-PUZZLE. Nodal support inside the *SPIN*, *AeBuster3*, *SPIN_Ml_a* clade is not shown for purposes of clarity. The shaded areas indicate the proposed division of the *hAT* superfamily into the *Buster* and *Ac* families. The scale bar represents a phylogenetic distance of 1 amino acid substitution per site.

The *Buster* transposon family contains a large number of new transposon and transposon-like sequences, including the transposons *AeBuster1* from the mosquito *Ae. aegypti* and *TcBuster* from *T. castaneum* whose activity we describe below. The *Buster* family also includes its namesake, the *Buster/Charlie* nonfunctional transposon-like sequences originally described in humans (SMIT 1999) that we have found to be extremely highly conserved and consider more in detail below.

Buster transposons are widely distributed and were found in aquatic and terrestrial protostomes and deuterostomes. The *Buster* family contains the recently described transposons from the bat, *M. lucifgus* (RAY *et al.* 2006). *Myotis hAT* forms a clade with *Buster* transposase sequences from another bat transposase (MlBuster1), a mosquito (*Ae. aegypti*), a sea urchin *Strongylocentrotus purpuratus*, and zebrafish (*Danio rerio*). The newly discovered bat transposase *Myotis hAT SPIN_Ml_a* forms a clade with *AeBuster3* from *Ae. aegypti* and the recently discovered *SPIN* transposons in some deuterostomes, some of which have been proposed to have been horizontally transferred between species to best explain their distribution (PACE *et al.* 2008). The placement of the SPIN transposons within the *Buster* family reveals their relationship to the *Buster* transposase sequences found in deuterostomes such as the sea squirt, *Ciona savignyi*, as well as to those found in the mosquitoes, *Ae. aegypti* and *Culex quinquefasciatus*.

Family-specific target-site duplications: A further sequence difference separates transposon members of the *Ac* and *Buster* families although this is based on the mobility properties of the transposase rather than its structure. For the majority of members of the *Ac* family, the consensus TSD generated upon transposon integration is 5′-*n*T*nnnn*A*n*-3′. The *Tol2* transposon, which we classify as being a member of the *Ac* family, has a slightly altered TSD of 5′-*S*TTATAA*S*-3′ (where *S* stands for C or G), which still conforms to the preference of members of this family to insert at sites containing a T and A at the second and seventh positions (KONDRYCHYN *et al.* 2009). The *Buster* transposons in contrast have the TSD consensus of 5′-*nnn*TA*nnn*-3′, which we derived both from sequences flanking *Buster* transposons in their native genomes and from the transposition assays described below. This TSD was also reported flanking *hAT* sequences in *M. lucifugus* (RAY *et al.* 2006), which we now classify as being members of the *Buster* family. It remains to be determined which part of the *hAT* transposase mediates target-site recognition.

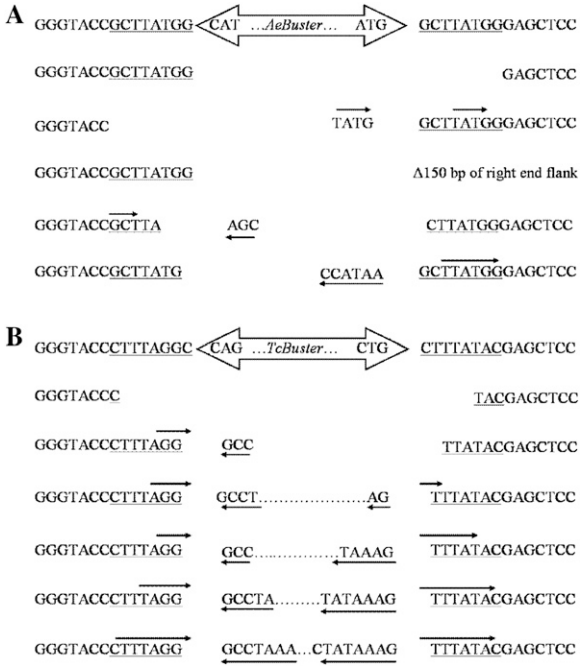

FIGURE 2.—Footprint sequences remaining at empty sites following excision of either *AeBuster1* (A) or *TcBuster* (B) from donor plasmids in *Ae. aegypti* embryos. TSDs are underlined and the transposon is shown within the block arrows. The sequences of five empty sites arising from the excision of *AeBuster1* and six empty sites arising from the excision of *TcBuster* are shown with additional DNA that is inserted into the empty sites shown between the TSDs. Smaller arrows show how this is related to sequences within the TSDs.

To determine if this difference in target-site selection was conserved between other members of the *Ac* and *Buster* families, we extracted 273 *hAT* sequences from RepBase using the criteria described in MATERIALS AND METHODS, combined them with two recently described *SPINs* (these two additional SPINs were not deposited into RepBase) and 24 *hAT* sequences from the current study, and examined their flanking TSDs (Table S1). Those *hAT* sequences that were annotated with a transposase sequence in RepBase or for which the transposase sequence was known (192 *hAT* sequences) were used to construct a more extensive tree and the distribution of the two types of TSDs assigned to each sequence (Figure S5). In this tree, members of the *Buster* family were found within a single clade as was expected from our previous results. Furthermore, 38/45 sequences assigned to this *Buster* clade were flanked by TSDs that contained the dinucleotide TA in the fourth and fifth position. Of the remaining seven, two (hAT-2 ET, hAT-41 SM) had the dinucleotide *N*A at the fourth and fifth positions (*N* indicating that a 50% consensus nucleotide sequence could not be identified for that position), two (SPIN Md, hAT1 MD) contained a nucleotide other than TA at the fourth and fifth positions so did not conform to the *Buster* target-site consensus, and the remaining three contained *NN* at these two positions. In contrast only 13/147 *hAT* sequences in the remaining clades had TSDs with TA at the fourth and fifth positions. Five of these (homo4, hAT-52 HM, hATm-3 HM, hATm-4 HM, hATm-56 HM) also contained T and A at the second and seventh positions, respectively, which, taken alone, are diagnostic of the *Ac* family TSD consensus sequence; three (hAT-53 HM, hATm34-HM, hATm26-HM) contained either T at the second position or A at the seventh position, while the remaining five did not show any similarities to the *Ac* family TSD consensus.

Further support for *SPINs* being members of the *Buster* family and generating TSDs containing TA at the fourth and fifth positions was obtained through the construction of a functional mammalian SPIN transposon on the basis of consensus sequence of inactive *SPINs* from mammals and their comparison with the active *Buster* transposons described here (X. LI, H. EWIS, R. H. HICE, N. PARKER, L. ZHOU, N. MALANI, F. BUSHMAN, C. FESCHOTTE, P. W. ATKINSON, N. L. CRAIG, unpublished results). Analysis of the TSDs generated by transposition of this resurrected *SPIN* in human cell revealed a consensus sequence of 5′-*NNN*TA*NNN*-3′ (X. LI, H. EWIS, R. H. HICE, N. PARKER, L. ZHOU, N. MALANI, F. BUSHMAN, C. FESCHOTTE, P. W. ATKINSON, N. L. CRAIG, unpublished results).

The three *Tip* transposons do not provide sufficient data to enable the placement of *Tip* transposons into either the *Ac* or *Buster* family or into a separate, third family of the *hAT* superfamily. Previously, sequence similarities between the *Tip100* ORFs and *Charlie* sequences were identified (ROBERTSON 2002). However, at the present time our small sample size does not allow further clarification of their position relative to other *hAT* transposons.

***TcBuster* and *AeBuster1* are active *Buster* transposons:** The distribution of *Buster* transposons across invertebrates and vertebrates led us to ask whether any were indeed active transposons. The *Buster* transposons from two insects, *Ae. aegypti* and *T. castaneum*, contain intact ORFs flanked by perfect TIRs and TSDs. The *TcBuster* transposon is 2489 bp in length, contains 18-bp TIRs, and has a 2110-bp ORF encoding a transposase 636 amino acids long. The *AeBuster1* transposon is 2459 bp in length, contains 15-bp TIRs, and has a 1919-bp ORF encoding a transposase 639 amino acids long. Both transposons thus possessed the structural attributes of active *hAT* transposases.

Both transposons were cloned from mosquito and beetle genomic DNA and were initially tested for activity in excision assays in *Ae. aegypti* embryos using established techniques (ATKINSON *et al.* 1993). Upon excision from their sites in the donor plasmid both *Buster* transposons left footprint sequences consistent with what has been observed for other active *hAT* transposons such as *hobo*, *Ac*, and *Tam3* (COEN *et al.* 1989; FEDEROFF 1989; ATKINSON *et al.* 1993) (Figure 2).

TABLE 1

***In vivo* transposition frequencies of *TcBuster* and *AeBuster1* in *D. melanogaster* and *Ae. aegypti* embryos**

Transposon/transposase	Species	No. of experiments	No. of donor plasmids	No. of independent transpositions	Transposition frequency ($\times 10^{-4}$)
TcBuster	*D. melanogaster*	1	134,600	35	2.6
0	*D. melanogaster*	2	145,000	0	0
TcBuster	*Ae. aegypti*	3	514,000	40	0.78
0	*Ae. aegypti*	3	382,200	0	0
AeBuster1	*D. melanogaster*	2	171,000	26	1.52
0	*D. melanogaster*	3	98,400	0	0
AeBuster1	*Ae. aegypti*	4	596,400	10	0.17
0	*Ae. aegypti*	3	268,000	0	0

Typically, additional nucleotides are found at the excision site and arise following the resolution of hairpin structures formed during excision followed by nonhomologous end joining (Zhou *et al.* 2004). In some cases the addition of DNA is associated with small deletions of the flanking DNA; in other cases no additional DNA remains and deletions into flanking DNA occur. These types of excision footprints were observed following excision of both *AeBuster1* and *TcBuster* in *Ae. aegypti* embryos; however, from the small sample size examined, it appeared as if more complete templated addition of extra DNA occurred following *TcBuster* excision.

We next examined the ability of both *Buster* transposons to transpose in *Ae. aegypti* and *D. melanogaster* embryos and in *D. melanogaster* S2 cells. Interplasmid transposition assays were performed according to previous protocols (Sarkar *et al.* 1997a). Both *Buster* transposons were capable of accurate cut-and-paste transposition in *D. melanogaster* and *Ae. aegypti* (Tables 1 and 2). As expected for *hAT* transposons, an 8-bp TSD was generated upon integration (Tables 3, 4, and 5) and, as seen when the endogenous TSDs of these transposons were sequenced, these *Busters* generated a TSD consensus sequence of 5′-*nnn*TA*nnn*-3′, which differs from TSDs generated by the *Ac* family members (Figure 3). The frequencies of transposition of both transposons in these two insects was within the same order of magnitude recorded for *Hermes* in both species (Sarkar *et al.* 1997a,b).

The *AeBuster1* transposase was His-tagged and purified from *E. coli* using protocols developed for the related *Hermes* transposase (Zhou *et al.* 2004). This catalyzed *in vitro* the strand transfer of a 40 mer containing either end of the *AeBuster1* element to a target molecule as was seen in similar experiments with the *Hermes* transposase (Zhou *et al.* 2004) (Figure 4). Strand transfer occurred with equal efficiency in the presence of Mg^{2+} or Mn^{2+}. These data show that at least one type of *Buster* transposase is amenable to purification from an *E. coli* expression system and that, *in vitro*, only 40 bp of the *AeBuster1* ends are necessary for recognition and binding of the transposase leading to strand transfer.

Conservation of *Buster* genes in mammals: The *Buster* family consists of both active transposons and very similar transposase-like sequences that have lost their TIRs, are highly conserved across species both in sequence and genomic location, and so appear to be important domesticated, or exapted, genes. As shown in Table S3A and Table S3B there is a high degree of conservation of *Buster* gene ORFs between mammalian species but clearly differences between the *Buster* genes themselves.

Buster1 genes retain a number of features important to active transposases such as the carboxylate DDE triad that forms the catalytic domain of active *hAT* transposases and the RW motif located downstream from the second aspartate of this triad. On the basis of the analysis of *Hermes* transposase this motif is believed to stabilize the penultimate, displaced nucleotide of flanking DNA upon hairpin formation during transposon excision (Zhou *et al.* 2004). The *TcBuster* and *AeBuster1* transposases contain the C*xx*H motif also found in the *Hermes* transposase and many other *hAT* transposases, and substitution of this histidine with alanine in *Hermes* severely limits *in vitro* transposition, indicating its requirement for function (Zhou *et al.* 2004).

The mammalian *Buster1* genes have the C*xx*H motif expanded to CLLYRH with histidine being replaced by

TABLE 2

***In vivo* transposition frequencies of *TcBuster* and *AeBuster1* in *D. melanogaster* S2 cells in the presence and absence of transposase**

Transposon/transposase	Species	No. of donor plasmids	No. of independent transpositions	Transposition frequency ($\times 10^{-4}$)
TcBuster	*D. melanogaster*/S2 cells	633,000	9	0.14
0	*D. melanogaster*/S2 cells	65,000	0	0
AeBuster1	*D. melanogaster*/S2 cells	2,413,000	7	0.03

TABLE 3

TSDs generated by the transposition of *TcBuster* into developing embryos of *D. melanogaster* and *Ae. aegypti*

Species	TSD	No. of Independent Events	Insertion position
D. melanogaster	CCT**TA**AAC	3	83(+)
	GTT**TA**AGG	3	83(−)
	TCA**TA**TTC	2	398(−)
	GAC**TA**CAT	1	509(+)
	CAG**TA**ATC	2	785(−)
	ATCAAAGC	1	801(−)
	GTG**TA**AAT	2	810(−)
	AAA**TA**AAC	1	869(−)
	TGT**TA**CTG	1	876(+)
	GAT**TA**AAG	1	943(+)
	CTT**TA**ATC	2	943(−)
	CGT**TA**AAT	2	966(−)
	ATC**TA**AAT	1	1045(−)
	CTC**TA**GCC	1	1965(−)
	CTC**TA**GAG	2	1993(+)
	ACA**TA**GCC	2	2207(−)
	GAA**TA**AAG	1	2325(+)
	CTT**TA**TTC	1	2325(−)
	GCT**TA**AAT	1	2445(+)
	GAT**TA**GAC	1	2470(+)
	GTC**TA**ATC	2	2470(−)
	TGT**TA**TAT	1	2493(+)
	ATA**TA**ACA	1	2493(−)
Ae. aegypti	ATT**TA**GAT	1	25(+)
	CCT**TA**AAC	1	83(+)
	GAA**TA**TGA	1	398(+)
	GGA**TA**GAC	2	427(+)
	GGT**TA**AAA	1	489(+)
	GAC**TA**CAT	2	509(+)
	GAT**TA**CTG	1	785(+)
	ATTGCTTT	1	798(+)
	GCTTTGAT	1	801(+)
	ATT**TA**CAC	2	810(+)
	GAT**TA**AAG	3	943(+)
	ATT**TA**ACG	1	966(+)
	CTA**TA**CGT	1	1268(−)
	GCA**TA**GGT	1	1925(+)
	GGC**TA**GAG	4	1965(+)
	CTC**TA**GAG	7	1993(+)
	CGT**TA**GCA	1	2050(+)
	AGA**TA**GAG	3	2146(+)
	GGC**TA**TGT	1	2207(+)
	ACA**TA**GCC	1	2207(−)
	GAA**TA**AAG	2	2325(+)
	CTT**TA**TTC	1	2325(−)
	TGT**TA**TAT	1	2493(+)

The 8-bp TSDs are shown with the conserved TA at positions 4 and 5 in boldface type. Insertion site into the pGDV1 target plasmid (2575 bp in length) and the orientation of insertion are shown.

TABLE 4

TSDs generated by the transposition of *AeBuster1* into developing embryos of *D. melanogaster* and *Ae. aegypti*

Species	TSD	No. of independent events	Insertion position
D. melanogaster	CATCAAGA	1	54(+)
	GTT**TA**AGG	1	83(−)
	ATTCAAAT	1	144(+)
	ATT**TA**GAC	1	168(+)
	ACT**TA**ACC	1	207(−)
	ATT**TA**CAC	1	810(+)
	GAT**TA**AAG	3	943(+)
	CTT**TA**ATC	1	943(−)
	CTC**TA**GAG	2	1993(+)
	ACT**TA**TAG	1	2160(+)
	CTA**TA**AGT	1	2160(−)
	ACA**TA**GCC	1	2207(−)
	AGC**TA**ACA	2	2214(−)
	TTA**TA**ACA	2	2381(+)
	GTT**TA**CCA	1	2425(+)
	TACCAGAT	1	2428(+)
	GCT**TA**AAT	2	2445(+)
	GAT**TA**GAC	1	2470(+)
	GTC**TA**ATC	1	2470(−)
Ae. aegypti	AAA**TA**AAC	1	460(+)
	ATT**TA**CAC	1	810(+)
	ATA**TA**AAT	1	925(+)
	CTT**TA**ATC	1	943(−)
	CTC**TA**GAG	3	1993(+)
	GGC**TA**TGT	1	2207(+)
	ACA**TA**GCC	1	2207(−)
	TGT**TA**GCG	1	2226(+)

The 8-bp TSDs are shown with the conserved TA at positions 4 and 5 in boldface type. Insertion site into the pGDV1 target plasmid (2575 bp in length) and the orientation of insertion are shown.

the aromatic amino acid tyrosine but in all other *Busters* the C*xx*H motif is conserved. In *TcBuster*, *AeBuster1* transposase, and the mammalian *Buster1* proteins, this C*xx*H or CLLYRH motif is part of a larger motif with the sequence C*xxx*C(28–29*x*)C*xx*H/CLLYRH, which, while similar to the THAP-type Zn and MYM-type Zn finger domains (Bessiere *et al.* 2008; Gocke and Yu 2008), appears to be a novel type of zinc finger that could, in principle, coordinate a divalent cation. The second C of this expanded motif is two residues upstream from the central D of the catalytic triad of these transposases, raising the question as to whether this region of the molecule can actually participate in two different chelations of a divalent cation. The conservation of these motifs in all these mammalian *Buster1* proteins suggests that *Buster1* remains functional in all these species and is supported by the low (d_N/d_S) values of *Buster1* across these species (Tables S3A and Tables S3B).

The *Buster2*, *3*, and *4* sequences are present in many eutherian species and are also highly conserved (Tables S3A, Table S3B, and Table S4). They are all found in humans, chimps, rhesus monkeys, dogs, and cattle with *Busters2* and *3* previously being identified in humans (Smit 1999). *Buster3*, also found in horses, is as conserved among these species as is *Buster1*; however, its DDE motif has been replaced by ADE that most likely renders it incapable of transposase-like catalytic activity. *Buster2* is also present in these four species and, in rats

TABLE 5

TSDs generated by the transposition of *AeBuster1* and *TcBuster* into *D. melanogaster* S2 cells

Transposon	TSD	No. of independent events	Insertion position
AeBuster1	GTT**TA**ATA	1	32(−)
	TAT**TA**GAA	1	679(+)
	GAT**TA**CGT	1	785(+)
	ATT**TA**CAC	1	810(+)
	GAT**TA**AAG	1	943(+)
	CTC**TA**GAG	1	1993(+)
	GAT**TA**GAC	1	2470(−)
TcBuster	ATT**TA**GAT	1	25(+)
	CCT**TA**AAC	1	83(+)
	TTT**TA**ACC	1	489(−)
	GAC**TA**CAT	1	509(+)
	GCTTTGAT	1	801(+)
	AAA**TA**AAC	1	869(−)
	GAT**TA**AAG	1	943(+)
	GTT**TA**CCA	1	2425(−)
	GCT**TA**AAT	1	2445(+)

The 8-bp TSDs are shown with the conserved TA at positions 4 and 5 in boldface type. Insertion site into the pGDV1 target plasmid (2575 bp in length) and the orientation of insertion are shown.

and mice, it is part of a fusion with a long N-terminal region with its own additional zinc finger motif. Rodent *Buster2* is highly conserved and is found in Rattus and Mus EST libraries. In humans, chimps, dogs, and cattle this additional N-terminal domain is present at the *Buster2* locus and conceptual translations indicate that the same protein could be synthesized in these species. The annotation of *Buster 4* in the human genome shows that it also contains an additional long N-terminal domain that contains a leucine-rich domain and an integrase domain. This has previously been deposited into GenBank as a SCAN domain containing protein 3 (NP_443155.1|), the carboxy end of which is identical to the *Charlie 10 hAT* element.

DISCUSSION

We have presented phylogenetic and integration site analyses that support the hypothesis that the *hAT* superfamily is composed of two distinct families of transposons. Previous analyses of *hAT* transposon phylogeny focused only on transposons within what we now define as the *Ac* branch of the *hAT* element superfamily (Kempken and Windhofer 2001; Rubin *et al.* 2001). The members of this branch known to be active are *Ac*, *Tam3*, *hobo*, *Hermes*, *Herves*, and *Tol2* with all also being active in new host species (Martin *et al.* 1989; Haring *et al.* 1991; O'brochta *et al.* 1994; Sarkar *et al.* 1997a; Kawakamai and Noda 2004; Arensburger *et al.* 2005; Everttts *et al.* 2007)

Our data show that members of the *Buster* family, which include the recently discovered *SPIN* transposons, are in fact capable of transposition upon introduction to new host species. Our data also illustrate the strong conservation of several domesticated *Buster* genes within several mammalian species including primates, dogs, horses, and cattle, strongly suggesting an essential, but as yet, unknown function for these genes. Most probably, especially for *Buster1*, this involves recognizing and binding to DNA and then possibly modifying it similar in principle to the example of how the RAG1 protein, which is related to the *Transib* transposase, rearranges specific chromosomal DNA as part of the process of V(D)J recombination (Kapitonov and

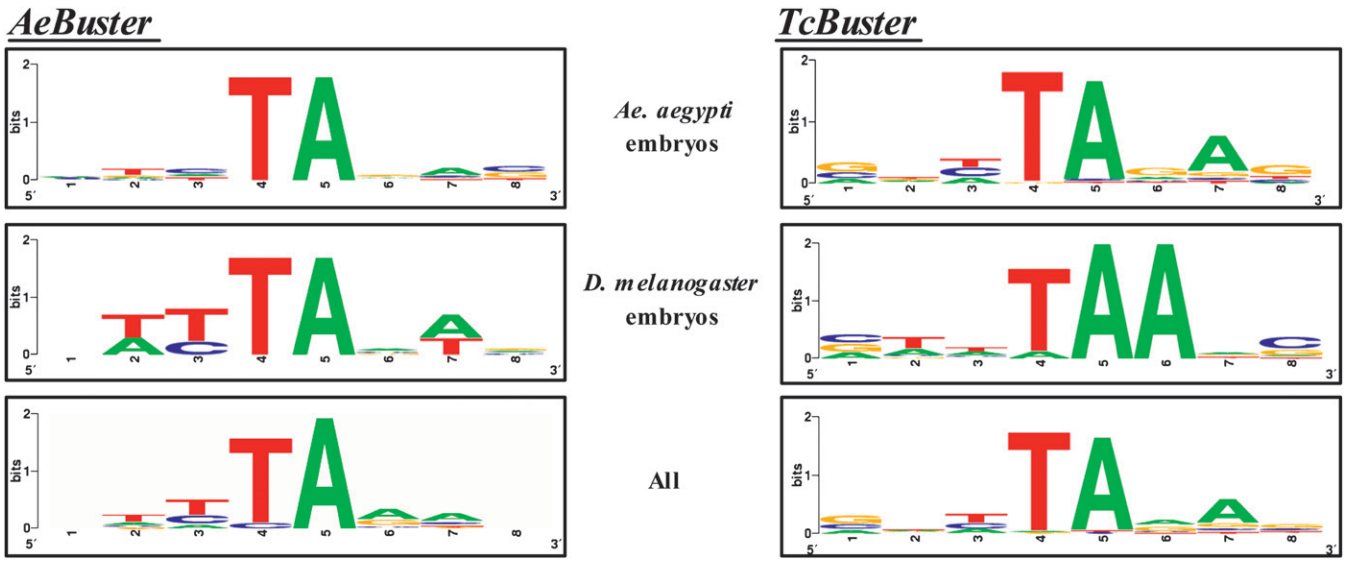

Figure 3.—WebLogo (http://weblogo.berkeley.edu/logo.cgi) TSDs generated by *AeBuster 1* and *TcBuster* in developing embryos of *Ae. aegypti* and *D. melanogaster*.

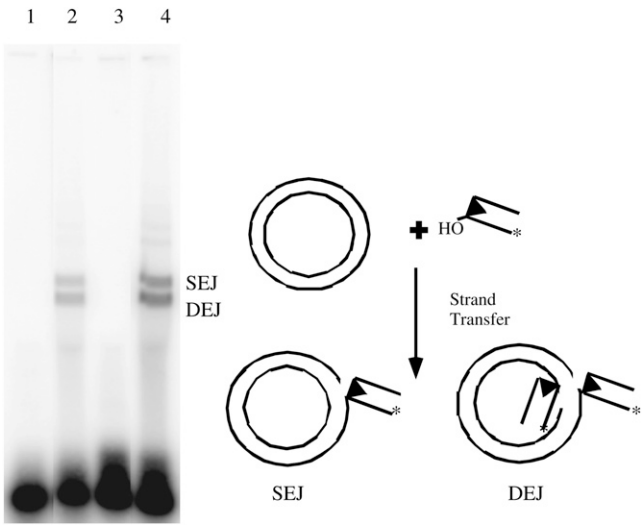

Figure 4.—Strand transfer reactions using precleaved left and right *AeBuster1* ends.5′-end-labeled 40mer oligonucleotides containing the L and R ends with their 16-bp TIRs were incubated with *AeBuster1* transposase and a pUC19 target DNA and then displayed on an agarose gel. In a single end join (SEJ), a single transposon end oligonucleotide is joined to the target plasmid; in a double end join (DEJ), to transposon ends oligonucleides join to the target DNA at the same position, linearizing the plasmid. Lane 1, left end–no transposase; lane 2, left end–plus transposase; lane 3, right end–no transposase; lane 4, right end–plus transposase. All the reactions shown were run on the same gel and are from the same gel image but have been cropped and arranged for easier viewing.

Jurka 2005). We speculate the true roles of these *Buster* proteins may therefore be closely related to their original roles as recombinases.

The organization of the *Buster* family is complex with close relationships between *Buster* sequences from mammals and *Buster* transposons from invertebrates. The *M. lucifugus Ml Buster1* and *Myotis-hAT* transposons are within a clade shared with *Buster* transposon sequences from sea urchin, zebrafish, and *Ae. aegypti*, while the *Myotis SPIN_MI_a* transposon is closely related to the *AeBuster3* transposon of *Ae. aegypti* and the newly discovered *SPIN* transposons from mammals and the pipid frog, *Xenopus tropicalis*. On separate branches are *Buster* transposons from two mosquitoes (*Ae. aegypti*, *Cx. quinquefasciatus*), a beetle (*T. castaneum*), a tunicate (*Ciona savigny*), and the highly conserved domesticated *Buster1*, *2*, *3*, and *4* sequences from mammals.

The breadth, yet discontinuity, in the distribution of *Buster* family sequences across vertebrates and invertebrates is intriguing (Figure 1). In cases where domestication has occurred in the mammals, the *Buster* sequences can be highly conserved (Table S3A, Table S3B, and Table S4). As noted by Pace *et al.* (2008), the *SPIN* transposons are also highly conserved among vertebrates as well as with an insect species, leading these authors to propose that they have been horizontally transferred between species (Gilbert *et al.* 2010). Given this hypothesis of horizonatal transfer of *SPIN* transposons between mammals, it is possible that the related *Buster* transposons may also be capable of transposition in vertebrates, including humans.

AeBuster1 is the first active DNA transposon to be isolated from the human disease vector mosquito *Ae. aegypti*. *TcBuster* is also an active transposon yet appears to be present in the sequenced *T. castaneum* genome only once. Given its activity in *D. melanogaster* and *Ae. aegypti*, this suggests that *TcBuster* may be a new introduction into *T. castaneum* or, alternatively, it has been subjected to fairly efficient negative regulation in this beetle.

Comparison of the TSDs generated by *AeBuster1* and *TcBuster* with those generated by *hAT* transposons of the *Ac* family show a clear difference in target-site specificity. In all cases 8-bp TSDs are generated but the consensus sequence between *TcBuster*, *AeBuster1*, and the other *hAT* transposons differs in that there is a strong preference for TA in the central positions of the 8-bp TSD. A more direct comparison between *AeBuster1*, *TcBuster*, and *Hermes* transposition into the pGDV1 target plasmid in transposition assays performed in *D. melanogaster* clearly illustrates this difference (Sarkar *et al.* 1997a). The hot spots of *Hermes* transposition into pGDV1 at sites 318 bp, 736 bp, 1254 bp, and 2303 bp are never targeted by *AeBuster1* or *TcBuster*, while the hot spots of both *AeBuster1* and *TcBuster* transposition into pGDV1 at sites 943 bp, 1993 bp, and 2470 bp are, with the exception of a single *Hermes* transposition into 1993 bp in *Ae. aegypti*, not targeted by *Hermes* (Sarkar *et al.* 1997a,b). *AeBuster1* and *TcBuster* therefore display a markedly different target-site preference from the *Hermes* transposon and other members of the *Ac* family.

The structural basis for this difference in TSD preference is not known but it is reasonable to speculate that this consistent difference between members of the *Ac* and *Buster* families resides in an as-yet unidentified region of the transposase that recognizes target DNA. The structure of the *Hermes* transposase reveals two DNA binding domains that, based on secondary structure alignments, are also present in the two active *Buster* transposases although it is not known if either plays a role in target-site recognition. Numerous amino acid differences exist between these regions. The difference in TSDs between the two active *Buster* transposases and the active members of the *Ac* family provides a means by which the target-site recognition sequence could be clearly identified through mutagenesis followed by assays that would reveal a change in TSD consensus from the *n*T*nnnn*A*n* form to the *nnn*TA*nnn* form or vice versa.

There are clear similarities between the TSDs generated by *AeBuster1* and *TcBuster* and those surrounding a subset of *hAT*-like *MER1* elements previously identified in the human genome (Smit and Riggs 1996). More recently, the insertion-site preferences of *MER1A*, *MER1B*, *MER20*, *MER30*, and *MER58A* sequences were

determined through the bioinformatic analysis of their positions within other transposons in the human genome and found to have the same preferences for T and A at the fourth and fifth positions of the 8-bp TSD (Levy *et al.* 2009). We propose that these *MER1* elements are actually remnants of once active *Buster* transposons, the only extant members of which are now the four domesticated *Buster* genes, which, on the basis of their conservation between species, clearly have important, but as-yet unknown, functions.

Our assignment of the *hAT* superfamily into at least two families on the basis of the sequence of their transposases and target-site selections is necessarily constrained by the number of completely sequenced genomes and our ability to locate full-length *hAT* transposases that are most likely functional. Similarly, the TIRs of *hAT* transposons can exhibit much variation in their sequence, making the identification of internally deleted *hAT* transposons and *hAT* MITEs problematic. Flanking 8-bp TSDs conforming to one of two consensus sequences described here are informative but not necessarily definitive since other transposons, for example the *P* element of *D. melanogaster*, also generate 8-bp TSDs upon insertion into DNA. Our data do not allow the assignment of *Tip100* transposons to one of the two families. We are also unable to assign *nDart hAT* elements from rice, which have been shown to generate GC-rich TSDs when they are mobilized through crossing with strains containing active *Dart* transposons (Takagi *et al.* 2010). The absence of an identified functional *Dart* transposase also currently prevents the true assignment of these transposons within the *hAT* superfamily.

The discovery of active *AeBuster1* and *TcBuster* transposons and their ability to transpose in new hosts offers support for the proposition that members of this family of transposons have been, at some stage, transferred between deuterostomes and protostomes. Indeed the presence of *Buster* sequences in a single clade composed of such diverse host species as mosquito, sea urchin, zebrafish, and bat illustrate the lack of concordance between the phylogenies of these sequences and their hosts.

The size and diversity of the *hAT* superfamily is somewhat unique among eukaryote transposons. On the basis of the structure and target-site recognition, it contains clearly distinguishable families with multiple active members from each available for modification and study both *in vitro* and *in vivo*. Active members from the *Ac* and *Buster* families are capable of high levels of transpositional activity in new hosts in different phyla and, within each family, clear examples of transpoase domestication are becoming more apparent as more genomes continue to be annotated. The presence and distribution of *Buster* transposons in mosquitoes, known vectors of pathogens between vertebrates and invertebrates, offers one conceivable route of horizontal transfer of these transposons and may, in part, explain the distribution of these transposons and domesticated sequences that is emerging. As such the ability to modify and exquisitely control their activity and target-site specificity should enable them to be become efficient genetic tools in medical and agricultural applications.

This research was supported by National Institutes of Health grants AI45741 to PWA and NLC and GM48102 to D.A.O. and P.W.A. N.L.C. is an Investigator at the Howard Hughes Medical Institute.

LITERATURE CITED

Abascal, F., R. Zardoya and D. Posada, 2005 ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21:** 2104–2105.

Arensburger, P., D. A. O'Brochta, J. Orsetti and P. W. Atkinson, 2005 *Herves*, a new active *hAT* transposable element from the African malaria mosquito *Anopheles gambiae*. Genetics **169:** 697–708.

Atkinson, P. W., W. D. Warren and D. A. O'Brochta, 1993 The *hobo* transposable element of *Drosophila* can be cross-mobilized in houseflies and excises like the *Ac* element of maize. Proc. Natl. Acad. Sci. USA **83:** 9693–9697.

Bessiere, D., C. Lacroix, S. Campagne, V. Ecochard, V. Guillet *et al.*, 2008 Structure-function analysis of the THAP zinc finger of THAP1, a large C2H2 DNA binding module linked to Rb/E2F pathways. J. Biol. Chem. **283:** 4352–4363.

Blackman, R. K., M. Macy, D. Koehler, R. Grimaila and W. M. Gelbart, 1989 Identification of a fully functional *hobo* transposable element and its use for germ line transformation of *Drosophila*. EMBO J. **8:** 211–217.

Britten, R., 2005 Transposable elements have contributed to thousands of human proteins. Proc. Natl. Acad. Sci. USA **103:** 1798–1803.

Britten, R., 2006 Almost all human genes resulted from ancient duplication. Proc. Natl. Acad. Sci. USA **193:** 19027–19032.

Coen, E. S., T. P. Robbins, J. Almeida, A. Hudson and R. Carpenter, 1989 Consequences and mechanisms of transposition in Antirrhinum majus., pp. 413–436 in Mobile DNA, edited by D. E. Berg, and M. M. Howe. American Society for Microbiology, Washington, DC.

Evertts, A. G., C. Plymire, N. L. Craig and H. L. Levin, 2007 The *Hermes* transposon of *Musca domestica* is an efficient tool for the mutagenesis of *Schizosaccharomyces pombe*. Genetics **177:** 2519–2523.

Federoff, N. V., 1989 Maize transposable elements, pp. 375–411 in Mobile DNA, edited by D. E. Berg and M. M. Howe. ASM Press, Washington, DC.

Feschotte, C., and E. J. Pritham, 2007 DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. **41:** 331–368.

Finnegan, D. J., 1990 Transposable elements and DNA transposition in eukaryotes. Curr. Op. Cell. Biol. **2:** 471–477.

Gilbert, C., S. Schaack, J. K. I. Pace, P. J. Brindley and C. Feshotte, 2010 A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature **464:** 1347–1350.

Gocke, C. B., and H. Yu, 2008 ZNF198 stabilizes the LSD1-coREST-HDCAC1 complex on chromatin through its MYM-type zinc fingers. PLoS One **3:** e3255.

Haring, M. A., M. J. Teeuwen-de Vroomen, H. J. Nijkamp and J. Hille, 1991 Trans-activation of an artificial *dTam3* transposable element in transgenic tobacco plants. Plant Mol. Biol. **16:** 39–47.

Hickman, A. B., Z. N. Perez, L. Zhou, P. Musingarimi, R. Ghirlando *et al.*, 2005 Molecular architecture of a eukaryotic transposase. Nat. Struct. Mol. Biol. **12:** 715–721.

Kapitonov, V. V., and J. Jurka, 2005 RAG1 core and V(D)J recombination signal sequnces were derived from *Transib* transposons. PLoS Biol. **3:** e181.

Kawakamai, K., and T. Noda, 2004 Transposition of the *Tol2* element, an *Ac*-like element from the Japanese medaka fish, *Oryzias latipes*, in mouse embryonic stem cells. Genetics **166:** 895–899.

Kawakami, K., and A. Shima, 1999 Identification of the *Tol2* transposase of the medaka fish *Oryzias latipes* that catalyzes excision of

a nonautonomous *Tol2* element in zebrafish *Danio rerio*. Gene **240:** 239–244.

Kempken, F., and F. Windhofer, 2001 The *hAT* family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma **110:** 1–9.

Kondrychyn, I., M. Garcia-Lecea, A. Emelyanov, S. Parinov and V. Korzh, 2009 Genome-wide analysis of *Tol2* transposon reintegration in zebrafish. BMC Genomics **10:** 418.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Levy, A., S. Schwartz and G. Ast, 2009 Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. Nucleic Acids Res. **38:** 1515–1530.

Martin, C., A. Prescott, C. Lister and S. MacKay, 1989 Activity of the transposon *Tam3* in *Antirrhinum* and tobacco: possible role of DNA methylation. EMBO J. **8:** 997–1004.

McClintock, B., 1950 The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA **36:** 344–355.

Moretti, S., F. Armougom, I. M. Wallace, D. G. Higgins, C. V. Jongeneel *et al.*, 2007 The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. Nucleic Acids Res. **35:** W645–W648.

Nene, V., J. R. Wortman, D. Lawson, B. Haas *et al.*, 2007 Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science **316:** 1718–1723.

Notredame, C., 2010 Computing mutiple sequence/structure alignments with the T-coffee package. Curr. Protoc. Bioinformatics **3.8:** 1–25.

O'Brochta, D. A., W. D. Warren, K. J. Saville and P. W. Atkinson 1994 Interplasmid transposition of *Drosophila hobo* elements in non-drosophilid insects. Mol. Gen. Genet. **244:** 9–14.

O'Brochta, D. A., W. D. Warren, K. J. Saville and P. W. Atkinson, 1996 *Hermes*, a functional non-drosophilid gene vector from *Musca domestica*. Genetics **142:** 907–914.

O'Brochta, D. A., C. D. Stosic, K. Pilitt, R. A. Subramanian, R. Hice *et al.*, 2009 Transpositionally active episomal *hAT* elements. BMC Mol. Biol. **14:** 108.

Ortiz, M., and E. L. S. Loreto, 2009 Characterization of new *hAT* transposable elements in 12 *Drosophila* genomes. Genetica **135:** 67–75.

Pace, J. K. I., C. Glibert, M. S. Clark and C. Feschotte, 2008 Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. USA **105:** 17023–17028.

Price, M. N., P. S. Dehal and A. P. Arkin, 2010 FastTree 2: approximately maximum-likelihood trees for large alignments. PLoS One **5:** e9490.

Ray, D. A., H. J. T. Pagan, M. L. Thompson and R. D. Stevens, 2006 Bats with *hATs*: Evidence for recent DNA transposon activity in genus *Myotis*. Mol. Biol. Evol. **24:** 632–639.

Ray, D. A., C. Feschotte, H. J. T. Pagan, J. D. Smith, E. J. Pritham *et al.*, 2008 Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome Research **18:** 717–728.

Richards, S., R. A. Gibbs, G. M. Weinstock, S. J. Brown, R. Dennell *et al.*, 2008 The genome of the modern beetle and pest *Tribolium castaneum*. Nature **452:** 949–955.

Robertson, H. M., 2002 Evolution of DNA transposons in eukaryotes, pp. 1093–1110 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. ASM Press, Washington, DC.

Ronquist, F., and J. P. Huelsenbeck, 2003 MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572–1574.

Rubin, E., G. Lithwick and A. A. Levy, 2001 Structure and evolution of the *hAT* transposon superfamily. Genetics **158:** 949–957.

Sarkar, A., C. J. Coates, S. Whyard, U. Willhoeft, P. W. Atkinson *et al.*, 1997a The *Hermes* element from *Musca domestica* can transpose in four families of cylorrhaphan flies. Genetica **99:** 15–29.

Sarkar, A., K. Yardley, P. W. Atkinson, A. A. James and D. A. O' Brochta, 1997b Transposition of the *Hermes* element in embryos of the vector mosquito, *Aedes aegypti*. Insect Biochem. Mol. Biol. **27:** 359–363.

Schmidt, H. A., K. Strimmer, M. Vingron and A. von Haeseler, 2002 TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:** 502–504.

Simmons, G. M., 1992 Horizontal transfer of *hobo* transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. Mol. Biol. Evol. **9:** 1050–1060.

Sinzelle, L., Z. Izsvak and Z. Ivics, 2009 Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell. Mol. Life Sci. **66:** 1073–1093.

Smit, A. F. A., 1999 Interspersed repeats and other mementos of transposable elements in Mamm. Genomes. Curr. Opinion Genet. Dev. **9:** 657–663.

Smit, A. F. A., and A. D. Riggs, 1996 *Tiggers* and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93:** 1443–1448.

Takagi, K., M. Maekawa, K. Tsugane and S. Iida, 2010 Transposition and traget preferences of an active nonautonmous DNA transposon *nDart1* and its relatives belonging to the *hAT* superfamily in rice. Mol. Genet. Genomics **284:** 343–355.

Volff, J.-N., 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays **28:** 913–922.

Whelan, S., and N. Goldman, 2001 A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. **18:** 691–699.

Zhou, L., R. Mitra, P. W. Atkinson, A. B. Hickman, F. Dyda *et al.*, 2004 Transposition of *hAT* elements links transposable elements and V(D)J recombination. Nature **432:** 995–1001.

Communicating editor: E. Alani

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.111.126813/DC1

Phylogenetic and Functional Characterization of the *hAT* Transposon Superfamily

Peter Arensburger, Robert H. Hice, Liqin Zhou, Ryan C. Smith, Ariane C. Tom, Jennifer A. Wright, Joshua Knapp, David A. O'Brochta, Nancy L. Craig and Peter W. Atkinson

DOI: 10.1534/genetics.111.126813

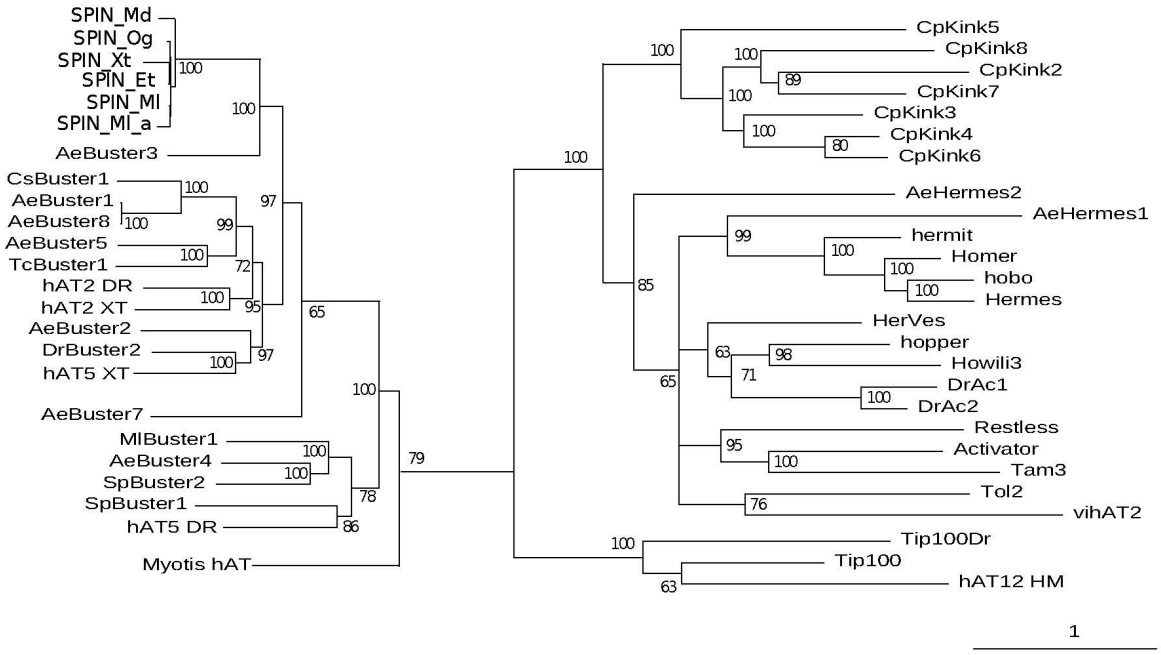

FIGURE S1.—Consensus phylogenetic tree showing the relationship of amino acid transposase sequences, between 50 selected full-length *hAT* elements based on Bayesian inference (50% majority rule consensus). Numbers next to most nodes show Bayesian posterior probability values produced by the program MRBayes. Nodal support inside the *SPIN*, *MlBuster2* clade is not shown for purposes of clarity. The scale bar represents a phylogenetic distance of 1 amino acid substitutions per site.

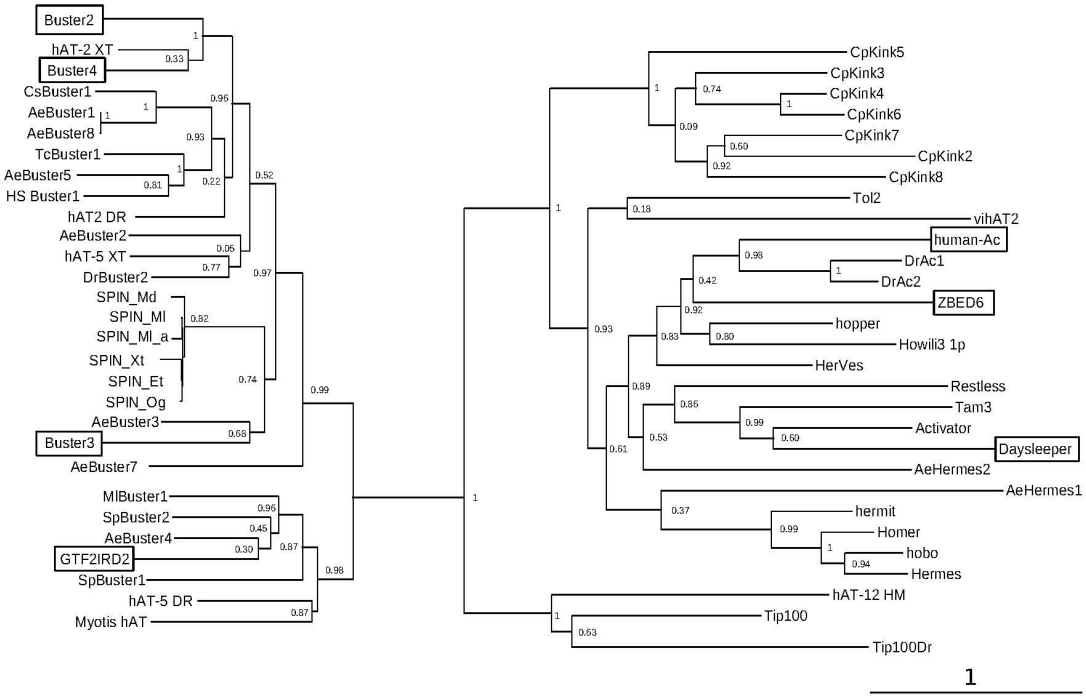

FIGURE S2.—Phylogenetic tree showing the relationships of 50 selected full-length *hAT* elements and 28 exapted genes with sequence similarity to *hAT* transposases. Numbers next to each node indicate local support values based on the Shimodaira-Hasegawa test as implemented in the program FastTree2. Support values are not shown for several nodes inside the *SPIN*, *SPIN_Ml_a* clade to maintain legibility. Exapted gene sequence names are shown in boxes. For purposes of clarity highly similar exapted gene sequences (see text) were grouped under the following labels: *Buster1* = *AmBuster1*, *BtBuster1*, *CfBuster1*, *EcBuster1*, *HsBuster1*, *PtBuster1; Buster2* = *AmBuster2*, *BtBuster2*, *CfBuster2*, *HsBuster2*, *MamBuster2*, *MmBuster2*, *PtBuster2*, *RnBuster2; Buster3* = *AmBuster3*, *BtBuster3*, *CfBuster3*, *EcBuster3*, *HsBuster3*, *MamBuster3*, *PtBuster3; Buster4* = *AmBuster4*, *CfBuster4*, *EcBuster4*, *HsBuster4*, *MamBuster4*, *PtBuster4*, *SsBuster4*. The scale bar represents a phylogenetic distance of 1 amino acid substitutions per site.

FIGURE S3.—Consensus alignment of 50 *hAT* transposons using the program M-Coffee using multiple sequence alignment programs (see text for details). Colors indicate area of high (red) and low (blue) agreement between the different alignment programs. Top numbers indicate average consistency of each sequence on a scale of 100.

Figure S3 is available for download as a PDF file at
http://www.genetics.org/cgi/content/full/genetics.111.126813/DC1.

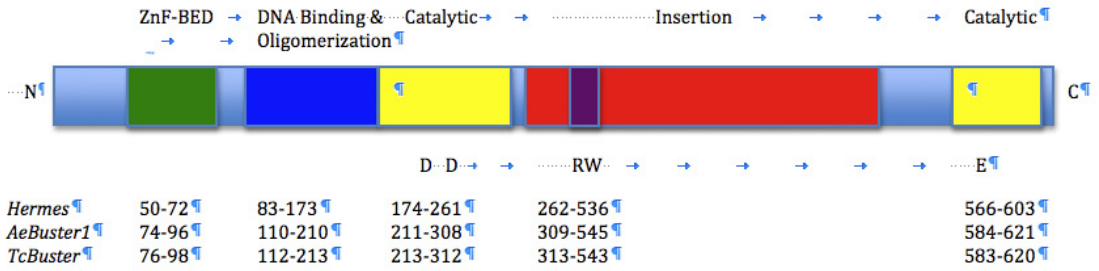

FIGURE S4.—The four domains shared between the *Hermes*, *AeBuster1* and *TcBuster hAT* transposases based on Hickman *et al.*, (2004) and our current studies. The figure is modified from Li *et al.* (in preparation). The approximate aa coordinates of each domain in each of the three transposases are shown below the diagram. The positions of the DDE motif in the catalytic domain are shown, as is the position of the RW motif located towards the amino end of the insertion domain.


hAT-3 AC
hAT-4 SM
hAT-34 SM
hAT-4 HM
SPIN_Md
SPIN_Og
SPIN_Et
SPIN_Ml
SPIN_Ml_a
hAT-5 XT
hAT-16 SM
CsBuster1
AeBuster1
TcBuster
hAT-14 SM
hAT-39 SM
hAT-35 SM
AeBuster5
hAT-2 DR
hAT-11 SM
SpBuster1
hAT-46 SM
hAT-19 HM
hAT-1 XT
hAT-3 XT
HAT6 CI
MlBuster1
HAT2 MD
hAT-2 MD
hAT-2 AC
hAT-2 ET
hAT-2 TS
hAT-12 SM
hAT-1 SM
AeBuster4
SpBuster2
hAT-5 SM
Myotis hAT
HAT1 MD
hAT-3 DR
hAT-11 XT
hAT-42 SM
hAT-41 SM
hAT-5 DR
hAT-75 HM
hAT-29 HM
hAT6-1 NV
hAT-28 SM
hAT-18 HM
hAT-35 HM
hAT-22 SM
hAT-37 SM
hAT-27 SM
hAT-65 HM
hAT-7 HM
hAT-41 HM
hAT-58 HM
hAT-67 HM
hAT-69 HM
hAT-16 HM
hAT-6 HM
hAT-29 SM
hAT-21 HM
hAT-55 HM
hAT-8 HM
hAT-5 HM
hAT-17 HM
hAT-56 HM
hAT-21 SM
hAT-12 HM
hAT-22 HM
hAT-34 HM
hAT-23 HM
hAT-40 HM
hAT-24 HM
hAT-47 HM
hAT-15 HM
hAT-4 NV
hAT-30 SM
hAT-25 HM
hAT-42 HM
hAT-43 HM
hAT-54 HM
hAT-61 HM
hAT-50 HM
hAT-2 NV
hAT-33 SM
hAT-47 SM
hAT-20 HM
hAT-68 HM
hAT-73 HM
hAT-76 HM
hAT-74 HM
hAT-25 SM
hAT-1 HM
hAT-44 HM
hAT-2 HM
hAT-28 HM
hAT-27 HM
hAT-26 HM
hAT-37 HM
hAT-53 HM
1

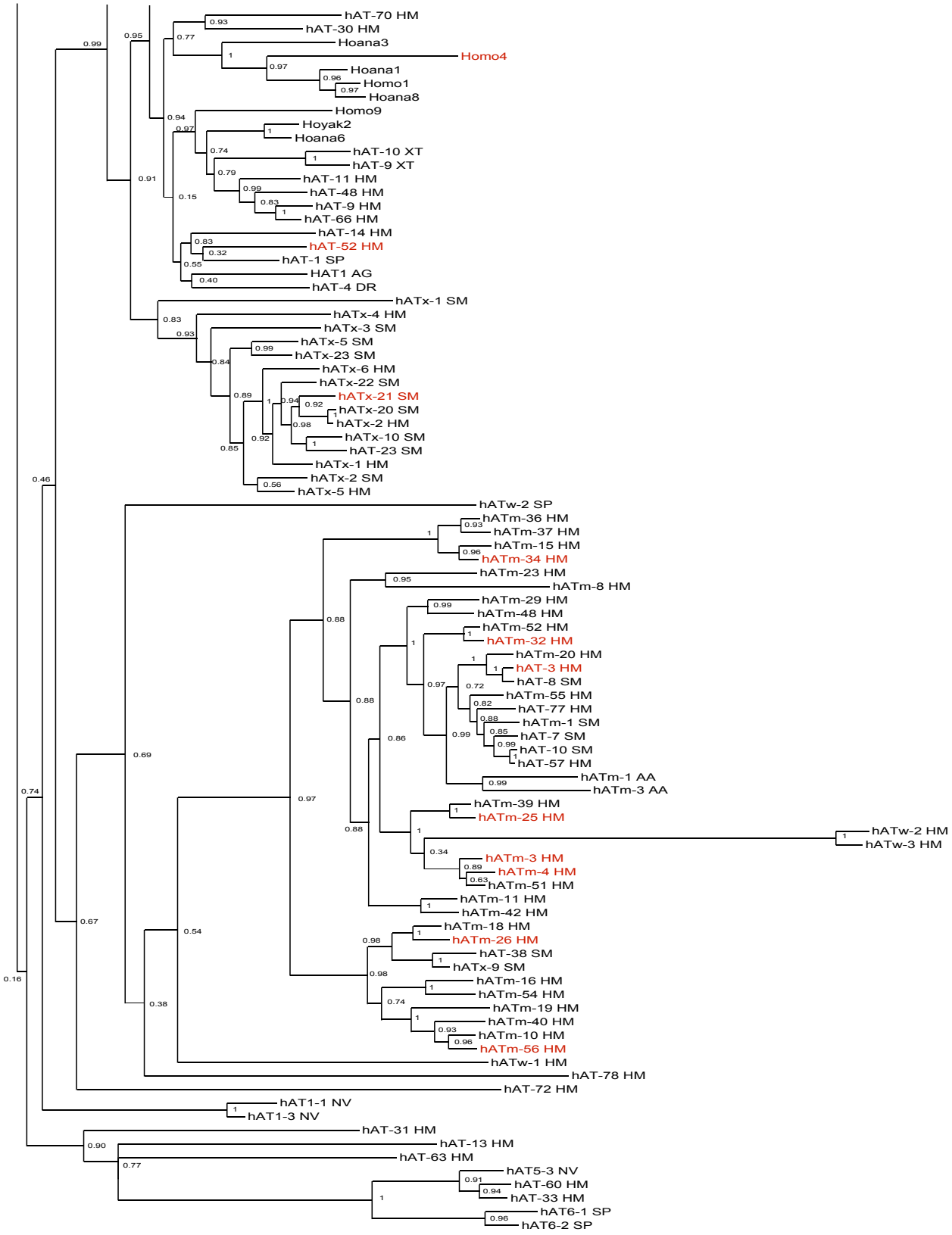
hAT-70 HM
hAT-30 HM
Hoana3
Homo4
Hoana1
Homo1
Hoana8
Homo9
Hoyak2
Hoana6
hAT-10 XT
hAT-9 XT
hAT-11 HM
hAT-48 HM
hAT-9 HM
hAT-66 HM
hAT-14 HM
hAT-52 HM
hAT-1 SP
HAT1 AG
hAT-4 DR
hATx-1 SM
hATx-4 HM
hATx-3 SM
hATx-5 SM
hATx-23 SM
hATx-6 HM
hATx-22 SM
hATx-21 SM
hATx-20 SM
hATx-2 HM
hATx-10 SM
hAT-23 SM
hATx-1 HM
hATx-2 SM
hATx-5 HM
hATw-2 SP
hATm-36 HM
hATm-37 HM
hATm-15 HM
hATm-34 HM
hATm-23 HM
hATm-8 HM
hATm-29 HM
hATm-48 HM
hATm-52 HM
hATm-32 HM
hATm-20 HM
hAT-3 HM
hAT-8 SM
hATm-55 HM
hAT-77 HM
hATm-1 SM
hAT-7 SM
hAT-10 SM
hAT-57 HM
hATm-1 AA
hATm-3 AA
hATm-39 HM
hATm-25 HM
hATw-2 HM
hATw-3 HM
hATm-3 HM
hATm-4 HM
hATm-51 HM
hATm-11 HM
hATm-42 HM
hATm-18 HM
hATm-26 HM
hAT-38 SM
hATx-9 SM
hATm-16 HM
hATm-54 HM
hATm-19 HM
hATm-40 HM
hATm-10 HM
hATm-56 HM
hATw-1 HM
hAT-78 HM
hAT-72 HM
hAT1-1 NV
hAT1-3 NV
hAT-31 HM
hAT-13 HM
hAT-63 HM
hAT5-3 NV
hAT-60 HM
hAT-33 HM
hAT6-1 SP
hAT6-2 SP
1

FIGURE S5.—Phylogenetic tree of selected transposons from animals from Repbase, two *SPIN* elements, and this study (see text for details). Taxa identified as having base pairs "A" and "T" at TSD positions 4 and 5 respectively are shown in red (see text for details). Numbers next to most nodes indicates local support values based on the Shimodaira-Hasegawa test as implemented in the program FastTree2. Nodal support was not shown for several clades with very small branches to help legibility. The scale bar represents a phylogenetic distance of 1 amino acid substitutions per site.

TABLE S1

Fifty percent consensus TSD sequence of animal transposons from RepBase, Pace *et al.* (2008), and the current study.

See text for how TE sequences were selected. "N/A" in the "number of unique 8 bp sequences examined" column indicates that no 8 bp putative TSDs were found using the search criteria described in the text.

TABLE S2

Database reference and species of origin of 28 presumed exapted mammalian genes

Tables S1 and S2 are available for download as Excel files at http://www.genetics.org/cgi/content/full/genetics.111.126813/DC1.

TABLE S3A

d_N, d_S, and d_N/d_S ratios estimate using the method of Nei and Gojobori (1986)for pairwise comparisons of *Buster1* (above diagonal) and *Buster2* (below diagonal) nucleotide ORF sequences in six mammalian genomes.

	1	2	3	4	5	6
1. *H. sapiens*	-	0.069[a] (0.001[b] 0.021[c])	0.090 (0.032 0.357)	0.093 (0.026 0.276)	NE[d]	NE
2. *P. troglodytes*	0.554 (0.002 0.003)	-	0.098 (0.034 0.347)	0.097 (0.026 0.267)	NE	NE
3. *B. taurus*	0.259 (0.062 0.238)	0.255 (0.062 0.241)	-		NE	NE
4. *C. familiaris*	0.281 (0.058 0.207)	0.276 (0.058 0.210)	0.165 (0.043 0.258)	-	NE	NE
5. *M. musculus*	0.237 (0.139 0.587)	0.238 (0.140 0.592)	0.206 (0.138 0.671)	0.199 (0.123 0.617)	-	NE
6. *R. norvegicus*	0.262 (0.143 0.548)	0.263 (0.145 0.553)	0.248 (0.144 0.579)	0.216 (0.129 0.599)	0.216 (0.049 0.227)	-

TABLE S3B

d_N, d_S, and d_N/d_S ratios estimates for pairwise comparisons of *Buster3* (above diagonal) and *Buster4* (below diagonal) nucleotide ORF sequences in four mammalian genomes.

	1	2	3	4
1. *H. sapiens*	-	0.221 (0.002 0.009)	0.051 (0.011 0.210)	0.021 (0.007 0.265)
2. *P. troglodytes*	0.260 (0.005 0.018)	-	0.048 (0.010 0.201)	0.028 (0.008 0.265)
3. *B. taurus*	0.215 (0.061 0.286)	0.224 (0.064 0.286)	-	0.051 (0.011 0.221)
4. *C. familiaris*	0.169 (0.080 0.475)	0.174 (0.083 0.478)	0.166 (0.076 0.458)	-

[a] d_N/d_S ratio
[b] d_N
[c] d_S
[d] Not Evaluated

Nei, M and T. Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-426.

TABLE S4

The distribution of highly conserved Buster sequences in eutherian mammals

SPECIES	BUSTER GENES				
	1	2a	2b	3	4
Homo sapiens (Human)	X	X		X	X
Pan troglodytes (Chimpanzee)	X	X		X	X
Macaca mulatta (Macaque monkey)		X		X	X
Ailuropoda melanoleuca (Giant panda)	X	X		X	X
Canis familiaris (Domestic dog)	X	X		X	X
Equus caballus (Horse)	X			X	X
Sus scrofa (Domestic pig)					X
Bos taurus (Cow)	X	X		X	
Mus musculus (Mouse)			X		
Rattus norvegicus (Rat)			X		

(+) denotes presence, (-) absence, in a species. The Buster 2a and 2b genes are related to each other but in rodents have a large addition at their amino end which includes an addition zinc finger domain.