Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals

Antoine Molaro,[1] Janet M. Young,[1] and Harmit S. Malik[1,2]

[1]Division of Basic Sciences, [2]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

Eukaryotic genomes must accomplish both compact packaging for genome stability and inheritance, as well as accessibility for gene expression. They do so using post-translational modifications of four ancient canonical histone proteins (H2A, H2B, H3, and H4) and by deploying histone variants with specialized chromatin functions. Some histone variants are conserved across all eukaryotes, whereas others are lineage-specific. Here, we performed detailed phylogenomic analyses of "short H2A histone" variants found in mammalian genomes. We discovered a previously undescribed typically-sized H2A variant in monotremes and marsupials, *H2A.R*, which may represent the common ancestor of the short H2As. We also discovered a novel class of short H2A histone variants in eutherian mammals, *H2A.Q*. We show that short H2A variants arose on the X Chromosome in the common ancestor of all eutherian mammals and diverged into four evolutionarily distinct clades: *H2A.B*, *H2A.L*, *H2A.P*, and *H2A.Q*. However, the repertoires of short histone H2A variants vary extensively among eutherian mammals due to lineage-specific gains and losses. Finally, we show that all four short H2As are subject to accelerated rates of protein evolution relative to both canonical and other variant H2A proteins including H2A.R. Our analyses reveal that short H2As are a unique class of testis-restricted histone variants displaying an unprecedented evolutionary dynamism. Based on their X-Chromosomal localization, genetic turnover, and testis-specific expression, we hypothesize that short H2A variants may participate in genetic conflicts involving sex chromosomes during reproduction.

[Supplemental material is available for this article.]

Nucleosomes are the basic unit of chromatin in most eukaryotes. A typical nucleosome particle wraps 150 bp of DNA around an octamer of four histone proteins: H3, H4, H2A, and H2B, each present in two copies (Kornberg 1974; Kornberg and Thomas 1974; Malik and Henikoff 2003). These four canonical "core" histone proteins are among the most conserved proteins in eukaryotes (Malik and Henikoff 2003; Gonzalez-Romero et al. 2010; Draizen et al. 2016) and are thought to share ancestry with histones found in some archaea (Sandman and Reeve 2006) and Marseilliviridae (Erives 2017). Canonical histones are usually expressed from multiple gene copies in genomes, which can be arranged in multigene clusters (Marzluff et al. 2008). Histone proteins have a stereotypical structure characterized by alpha-helices comprising a histone fold domain (HFD), which is flanked by largely unstructured N- and C-terminal tails (Luger et al. 1997). Post-translational modifications of histone tails can regulate their association with linker DNA (between neighboring nucleosomes) or recruitment of chromatin-modifying proteins and provide another layer of regulation to overall chromatin structure.

In combination with post-translational modification of histones, eukaryotic genomes also establish diverse chromatin states using histone variants, encoded by stand-alone single-copy genes, which replace canonical histones in the nucleosome (Weber and Henikoff 2014; Talbert and Henikoff 2017). Although histone variants are closely related to canonical histones, they differ at several key amino acid residues or contain additional domains that dictate their noncanonical functions (Malik and Henikoff 2003; Talbert and Henikoff 2010; Draizen et al. 2016). For example, the H2A.Z variant is preferentially incorporated at the transcriptional start sites of active genes, the CENPA (also known as CenH3) variant localizes to centromeres and triggers the assembly of the kinetochore, and the H2A.X variant assists with DNA repair at sites of double-strand breaks (Weber and Henikoff 2014; Talbert and Henikoff 2017).

Some histone variants, such as H2A.Z and CENPA, diverged from canonical histones early in eukaryotic evolution and are present in nearly all eukaryotic lineages. Other variants originated more recently and are found in just one lineage, e.g., macroH2A in Filozoans (a lineage that includes animals and choanoflagellates) (Malik and Henikoff 2003; Talbert and Henikoff 2010; Rivera-Casas et al. 2016). Most histone variants are highly conserved and evolve under strong purifying selection (Piontkivska et al. 2002; Rooney et al. 2002; Talbert et al. 2012). One exception is CENPA, which evolves rapidly in animals and plants. This rapid evolution is hypothesized to be a result of centromere drive or chromosomal competition during female meiosis (Henikoff et al. 2001; Malik and Henikoff 2009).

Most core and most variant histones are expressed ubiquitously. However, one class of histone variants, known as short H2A variants, is expressed almost exclusively during mammalian male germ cell development (Govin et al. 2007; Boussouar et al. 2008; Ferguson et al. 2009), before the nearly complete replacement of histones by protamines in sperm nuclei (Oliva and Dixon 1991; Hammoud et al. 2009). Three classes of short H2A variants, H2A.B, H2A.L, and H2A.P, have been described so far; each exhibits <50% amino acid identity with canonical H2A (Shaytan et al. 2015; Draizen et al. 2016). H2A.L and H2A.B are each known

Corresponding author: hsmalik@fhcrc.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.229799.117.

28:460–473 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/18; www.genome.org

to be encoded by several closely related genes in the mouse and human genomes (Ferguson et al. 2009; Ishibashi et al. 2010). Mouse H2A.B is first detected in spermatocyte nuclei at the onset of meiosis but disappears during spermiogenesis (Govin et al. 2007; Soboleva et al. 2012). In contrast, H2A.L accumulates in spermatid nuclei until the end of spermatogenesis and remains in mature sperm chromatin even after protamine exchange in mouse (Govin et al. 2007; Baker et al. 2008; Ishibashi et al. 2010), eventually disappearing from the paternal pronucleus following fertilization (Wu et al. 2008). In contrast, human sperm lacks H2A.L protein and instead retains H2A.B (Baker et al. 2007). There have been only a few studies investigating the in vivo functions of short histone variants. These have shown that H2A.B may be required for post-meiotic gene expression in mouse (Barral et al. 2017; Soboleva et al. 2017), whereas a specific *H2a.l* gene is essential for initiating the replacement of histones with protamines and male fertility (Barral et al. 2017; Soboleva et al. 2017).

In vitro studies have revealed that short H2A variants have divergent features that influence chromatin structure and function. These features (summarized in Supplemental Fig. S1) include a truncated C-terminal "docking domain" (a key feature of H2A that maintains the native structural properties of the nucleosome), loss of N-terminal lysine residues, and a weakened acidic patch, which normally stabilizes nucleosome-nucleosome interactions during chromatin compaction for other H2As (Luger et al. 1997; Bonisch and Hake 2012; Soboleva et al. 2012; Shaytan et al. 2015; Buschbeck and Hake 2017). Consequently, nucleosomes containing short H2A variants wrap shorter stretches of DNA (~120–130 bp) than canonical H2A-containing nucleosomes (~150 bp) and form more loosely packed chromatin. Indeed, short H2A variants localize to sites of open chromatin and potentiate DNA synthesis, transcription, and splicing (Chadwick and Willard 2001; Angelov et al. 2004; Bao et al. 2004; Gautier et al. 2004; Doyen et al. 2006; Govin et al. 2007; Syed et al. 2009; Soboleva et al. 2012, 2017; Tolstorukov et al. 2012; Arimura et al. 2013; Sansoni et al. 2014; Barral et al. 2017).

Evolutionary analyses of short histone variants have chiefly focused on *H2A.B* genes (Eirin-Lopez et al. 2008; Ishibashi et al. 2010) or mouse *H2a.l* genes (Ferguson et al. 2009). However, we still lack a clear understanding of the origins, orthology, and evolutionary trajectories of short H2A variants and how this has affected their function in mammalian genomes. Lack of an evolutionary framework has created confusion in nomenclature and in the interpretation of functional studies in mouse and human. For instance, *H2A.P* was initially believed to be an allele of *H2A.L* (Govin et al. 2007) but was recently confirmed as a distinct histone variant (Marino-Ramirez et al. 2006; Draizen et al. 2016; El Kennani et al. 2017). Finally, it remains unclear if expression during male gametogenesis is an ancestral feature of all extant short H2As and what specific structural features discriminate short H2A families from each another. Here, we performed detailed phylogenomic analyses to gain further insight into the function and diversification of mammalian male germ cell–specific short H2A histone variants.

Results

Age and orthology of short histone H2A variants in eutherian mammals

To gain a comprehensive understanding of the origins and evolution of mammalian short H2A variants, we scanned genome sequences from representative mammals (and chicken as an outgroup) using BLASTN and TBLASTN searches with human H2A.B, human H2A.P, and rat H2a.l sequences as queries (Supplemental Table S1). We used PhyML to perform phylogenetic analyses of these candidate short H2A variants, comparing them to canonical and variant H2As (H2A.Z, H2A.X and macroH2A) from representative species of mammals, using chicken and *Xenopus* as outgroups (Fig. 1A; Supplemental Data S1; Guindon and Gascuel 2003; Guindon et al. 2010). We also used shared synteny of each candidate short H2A variant to identify genes that had been present at the same genomic location in ancestral mammalian species (Fig. 1B; Supplemental Fig. S2).

Although most candidate short H2A variants grouped with previously characterized *H2A.B*, *H2A.P*, or *H2A.L* genes, we uncovered two previously undescribed and phylogenetically distinct clades of H2A variants. We tentatively named the first distinct clade *H2A.Q,* abiding by recently established histone nomenclature guidelines (Talbert et al. 2012). *H2A.Q* is a short H2A variant present in many eutherian mammalian genomes at a distinct syntenic location (Fig. 1B). We found a second distinct clade of H2A variants in marsupial and platypus genomes, which lack orthologs of any of the four short histone H2A variants found in eutherian mammals. We named this variant *H2A.R* both to highlight its novelty and to be compliant with the proposed guidelines for histone nomenclature (Talbert et al. 2012). While the H2A.R variant shares some characteristics with the short H2As (H2A.B, H2A.L, H2A.P, and H2A.Q), its features more closely resemble canonical H2A (see below). In contrast, no novel histone H2A variants were recovered in any bird genomes that we queried (chicken, zebra finch, and turkey). This implies that *H2A.Rs* and short histone H2A variants evolved exclusively in the mammalian lineage, after the split from birds. Like canonical *H2A*, *H2A.R*, *H2A.B*, *H2A.L*, *H2A.P*, and *H2A.Q* sequences all have intronless open reading frames.

The four clades of eutherian mammal short H2A variants (H2A.B, H2A.L, H2A.P, and H2A.Q) emerged from a single, well-supported monophyletic clade, confirming their common ancestry. Furthermore, each of the short H2A variants, except H2A.L, is found in well-supported monophyletic clades. This indicates that these clades have diverged independently for at least 100 million years. All four eutherian short histone variants group with extremely high support with H2A.R variants, paralleling the mammalian species tree (Bininda-Emonds et al. 2007). Based on this phylogenetic analysis, we conclude that a precursor histone variant diversified to the typically sized H2A.R in monotremes and marsupials and to four classes of short histone H2A variants found only in eutherian mammals (Fig. 1A).

We used the distribution of short H2A genes along the accepted mammalian phylogeny (Bininda-Emonds et al. 2007), analysis of shared synteny (Fig. 1B), and parsimony criteria to conclude that the common ancestor of eutherian mammals encoded two or three *H2A.B* genes, three *H2A.L* genes, a single *H2A.P* gene, and a single *H2A.Q* gene in six distinct X-linked loci. For clarity and so that paralogs can be distinguished from one another, we arbitrarily numbered these ancestral loci relative to their position on each arm of the X Chromosome from telomere to centromere. *H2A.B.1.1* and *H2A.B.1.2* are so named since they seem to have arisen from an ancestral duplication within a single X Chromosome locus, distinct from *H2A.B.2*, although the incomplete status of the elephant and armadillo genome assemblies precludes us from precisely dating this event. We also uncovered additional paralogs of these short histone variants in nonsyntenic locations (Supplemental Fig. S2), likely representing additional lineage-restricted duplications. The most dramatic example of such

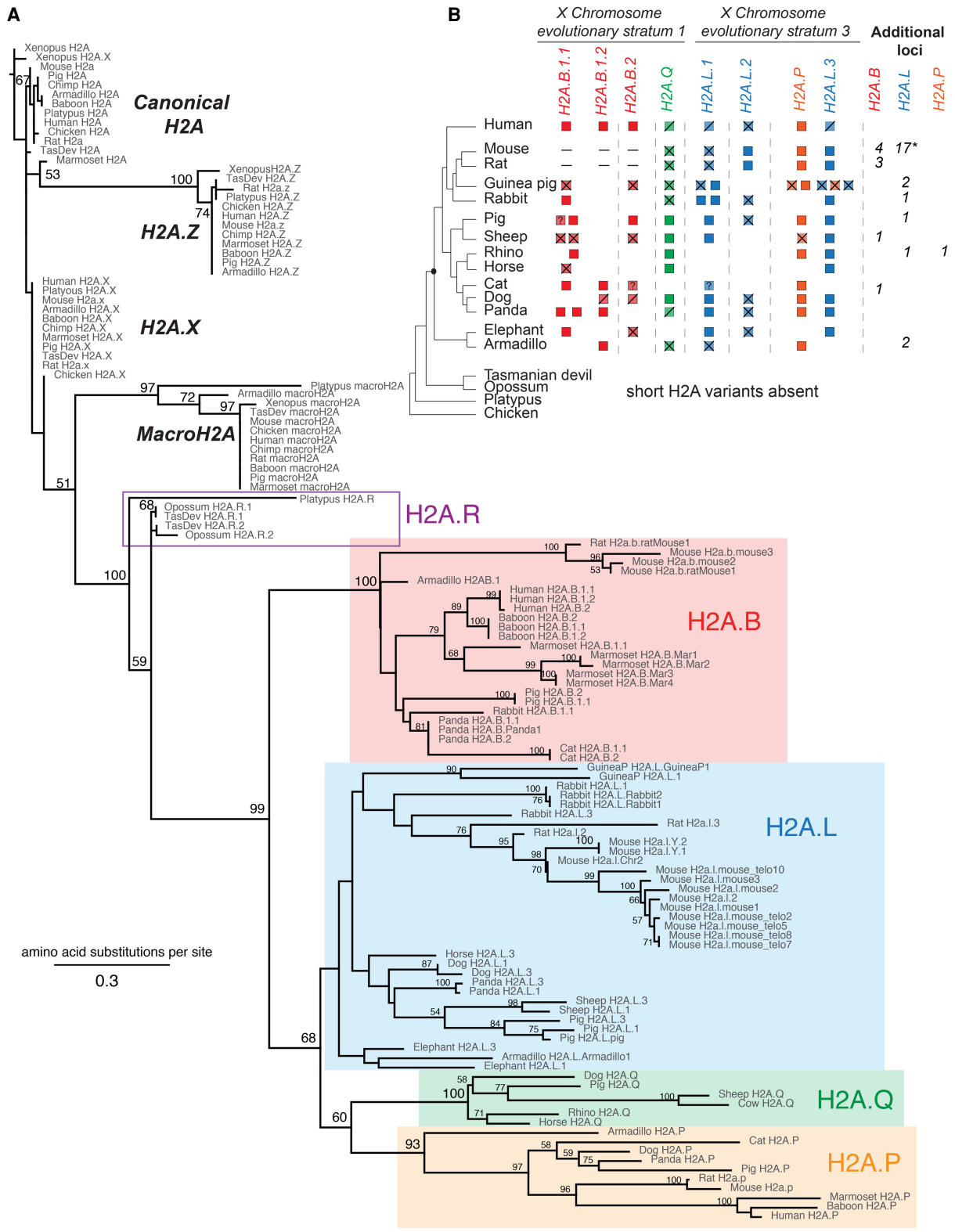

Figure 1. Mammalian short H2A repertoires and phylogeny. (*A*) Maximum-likelihood protein phylogeny of the histone fold domain of canonical H2A and H2A variants from *Xenopus*, chicken, and representative mammalian species. Bootstrap values are shown at all nodes that have >50% bootstrap support. The tree is rooted using *Xenopus* H2A. Short H2A variant clades are highlighted: H2A.B (red box), H2A.P (orange box), H2A.Q (green box), and H2A.L (blue box). (*B*) Schematic representation of the shared syntenic locations encoding short H2A variants. Short H2As are indicated to the *right* of a species cladogram (Bininda-Emonds et al. 2007). The origin of the eutherian mammals is highlighted by a black dot. Intact open reading frames are shown with filled boxes. Disrupted open reading frames are shown with crossed boxes (confirmed pseudogenes) or diagonal lines (inferred pseudogenes) (see Results). A dash (–) indicates absence at the syntenic location, a "?" refers to incomplete sequence information, and an empty white space indicates a missing gene at that syntenic location—we have not attempted to distinguish true gene loss from absence due to assembly gaps; see Supplemental Figure S2 for details. X Chromosome evolutionary strata are indicated at the *top* (Lahn and Page 1999; Sandstedt and Tucker 2004; Ross et al. 2005). For the additional *H2a.l* loci in mouse, "*" denotes the presence of two Y-linked and one autosomal copies.

lineage-specific expansion and diversification has been previously described in mouse (Ferguson et al. 2009), which encodes a total of 20 *H2a.l* loci (including two Y-linked and one autosomal copy) (Supplemental Table S1), whereas most other mammals encode between zero and five intact *H2A.L* genes. The duplications that gave rise to numerous short H2A loci could have arisen via segmental duplications, or retroduplication. Regardless of the mechanism of duplication, we provide evidence below that many of these loci encode functional protein-coding genes.

It is important to note that our analysis is, by necessity, based on mammalian genome assemblies at different stages of completion. However, the general conclusions are robust to caveats of imperfect assemblies, and the numbers of short H2As we identify in each genome should be viewed as lower bound estimates.

H2A.Q and H2A.R are novel testis-restricted H2A variants

Previously, only mouse and human tissues have been closely investigated for expression of short histone variants, including by detailed proteomic analysis of testes (El Kennani et al. 2017). However, both these genomes lack intact copies of the newly identified *H2A.Q* and *H2A.R* variants. In order to explore the possible function of *H2A.Q* and *H2A.R*, we analyzed publicly available RNA-seq data sets from a diverse set of mammals (dog, pig, human, platypus, and opossum) to determine whether and where these novel short histone variants are transcribed (Fig. 2A, B; Supplemental Table S3). We included *H2A.B*, *H2A.L*, and *H2A.P* genes for comparison (Fig. 2A), as well as selected "housekeeping" genes (Supplemental Fig. S3). Although short-read sequencing cannot distinguish expression levels of very similar paralogs and will partially "average" expression levels for similar genes (e.g., the three *H2A.B* genes of pig or human), this issue does not affect our analysis of *H2A.P*, *H2A.Q*, or *H2A.R*, which do not encode multiple similar paralogs in each genome.

Like other intact short H2A genes, dog and pig *H2A.Q* genes are transcribed in testes but are undetectable in other tissues examined, including ovary (Fig. 2A,B). Expression levels are robust and exceed those of most other genes in the genome. Indeed,

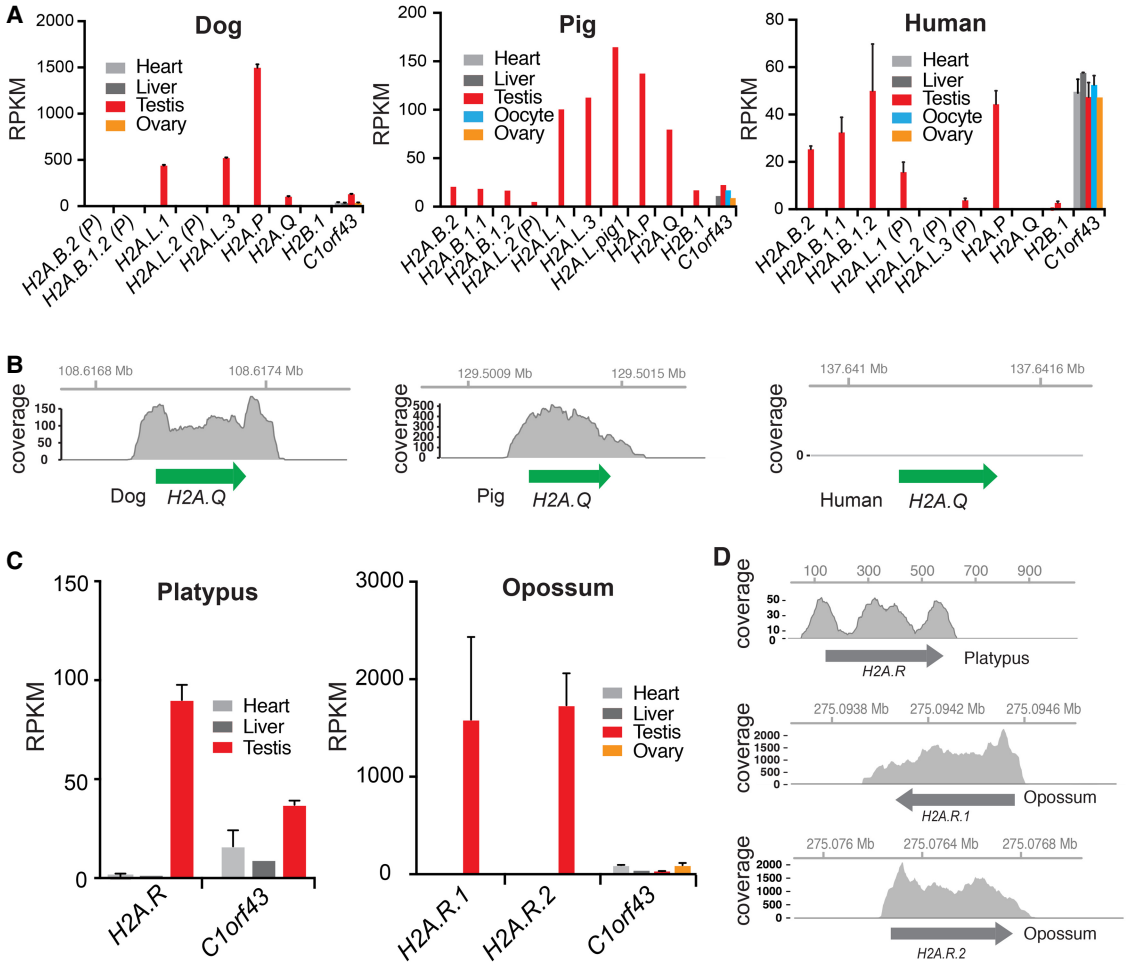

Figure 2. Expression of short histone variants. (*A*) Analyses of short H2A expression using public RNA-seq data from selected tissues of dog, pig, and human. Error bars show standard deviation of biological replicates, where available. Expression of the germ cell–specific H2B.1 variant and a housekeeping gene *C1orf43* (Eisenberg and Levanon 2013) are shown for comparison (more controls are shown in Supplemental Fig. S3). Mapped reads are shown in reads per kilobase per million mapped (RPKM). (*B*) RNA-seq coverage over the *H2A.Q* locus in dog, pig, and human. Coordinates along the assembled X Chromosome are indicated at the *top*. (*C*) RNA-seq analyses of platypus and opossum *H2A.Rs*, and one housekeeping gene (as in *A*), across somatic and reproductive tissues. Error bars as in *A*. (*D*) RNA-seq coverage as in *B*. Coordinates along the annotated contig or chromosome (Chr 6 for opossum *H2A.Rs* and contig111576 for platypus) are shown at the *top*.

expression levels of short histone variants are comparable to various "housekeeping" genes (Supplemental Fig. S3). This tissue-specific expression adds functional support to the idea that H2A.Q is a bona fide short H2A variant. However, we were unable to detect *H2A.Q* expression in any tissue in humans (Fig. 2A,B) or several other primates, suggesting that *H2A.Q* might be nonfunctional in primates. Furthermore, we found that platypus and opossum *H2A.R* variants are expressed in the testis but not in other tissues we examined, including opossum ovary (Fig. 2C,D). Our expression analysis allows us to infer that testis-specific expression preceded the divergence of *H2A.R* and eutherian short histone H2A variants and is a common feature of all these genes.

Recurrent losses of short histone H2A variants

Previous studies have suggested that the human genome contains one or more intact *H2A.L* genes (Draizen et al. 2016). However, we noticed atypical features that suggested some of these *H2A.L* copies may not be functional (Supplemental Fig. S4). We found that one of the three ancestral *H2A.L* loci (*H2A.L.2*) contained a pseudogene in all primates, having acquired inactivating mutations early in primate evolution. In contrast, many primate *H2A.L.1* and *H2A.L.3* genes had long 3′ extensions of different sizes but did not bear any stop codons or frameshifts within their coding region. We therefore evaluated primate *H2A.L.1* and *H2A.L.3* genes for signatures of purifying selection (Supplemental Fig. S4; see Methods). We estimated the rates of nonsynonymous (d_N) and synonymous (d_S) substitutions along codon alignments of these variants. The ratio of those rates (d_N/d_S) reflects the selective pressures acting on the aligned sequences; a ratio close to 1 reflects neutral evolution (or genetic drift, expected for a pseudogene), while a ratio close to 0 reflects very strong purifying selection. Using a phylogeny and sequence alignment, we tested the relative likelihoods of a model where the sequences evolve neutrally (model 0 with d_N/d_S fixed at 1) versus a model where the sequences evolve nonneutrally (model 0 with d_N/d_S estimated) (see Methods). These tests cannot reject a model of neutral evolution for Old World monkey and hominoid *H2A.L* genes ($d_N/d_S = 0.79$, $P = 0.49$ in a test to reject neutral evolution) (see Methods) but provide marginally significant evidence for purifying selection in the New World monkey *H2A.L* genes that lack C-terminal extensions ($d_N/d_S = 0.62$, $P = 0.054$). We therefore suggest that *H2A.L* function was lost in Old World monkeys and hominoids but retained in New World monkeys. Our findings are congruent with extensive proteomic studies that failed to detect H2A.L protein in human sperm or other tissues, even though H2A.L is readily apparent in mouse sperm (Baker et al. 2007, 2008; El Kennani et al. 2017).

We used similar analyses to deduce that primate *H2A.Q* sequences are also pseudogenes, even though no obvious defects in their coding regions were found (Supplemental Fig. S5). First, we found that simian primate *H2A.Q* genes have a deletion removing seven amino acids spanning Loop 1 and the alpha 2 helix (Supplemental Fig. S5). Second, we cannot reject a model of neutral evolution for primate *H2A.Q* orthologs ($d_N/d_S = 0.66$, $P = 0.27$), whereas there is robust evidence of purifying selection in other mammals ($d_N/d_S = 0.54$, $P < 0.0001$) (Supplemental Fig. S5). Together with our finding that *H2A.Q* is not expressed in humans (Fig. 2), these evolutionary observations allow us to suggest that its function has been lost in primates.

Surprisingly, none of the extant mammalian genomes we examined harbors the same repertoire of short histone H2A variants (Fig. 1B). While incomplete genome assemblies prevent us from knowing the exact complement of short H2A variants in each species, we find several instances where short histone variant orthologs in syntenic locations were present but clearly pseudogenized by early stop codons or frameshifts (Fig. 1B; Supplemental Fig. S2), strongly suggesting complete functional loss of one or more of the four classes of short H2A histone variants. For instance, all rodents we examined encode a pseudogenized version of *H2A.Q*, the sheep genome only encodes pseudogenes for *H2A.P* and *H2A.B*, and dog *H2A.B* genes have missing start codons. Our findings reveal an unprecedented degree of turnover for histone variant genes compared to other histone genes. Since none of the original eutherian short histone H2A gene clades appears to be strictly retained in all mammals, it suggests that neither H2A.L, H2A.B, H2A.P, nor H2A.Q performs a universally essential function in eutherian mammals or that these might have interchangeable functions in at least some species (see Discussion).

X-Chromosomal location of short histone H2A variants

In genome assemblies where sequences have been assigned to chromosomes (e.g., primates, rodents, and Laurasiatheria [pig, horse, dog, etc.]), we find that the six syntenic eutherian short H2A variant loci are all on the X Chromosome. Therefore, short H2A variants must have resided on the X Chromosome at least since the common ancestor of primates, rodents, and Laurasiatheria. Furthermore, assigning short H2A locations to previously identified X Chromosome evolutionary strata confirms that *H2A.B*, *H2A.L*, *H2A.P*, and *H2A.Q* genes originated on portions of the X Chromosome that are ancestral to all eutherian mammals (Lahn and Page 1999; Sandstedt and Tucker 2004; Ross et al. 2005). The marsupial *H2A.R* loci are autosomal (Chr 6 in opossum, Chr 2 in Tasmanian devil), and the platypus locus is on a short unmapped scaffold. Therefore, it is unclear whether the common ancestor of *H2A.R* and the short H2As was autosomal or sex chromosome-linked.

Rodents are the only exception to the X Chromosomal location of short histone variants in eutherian mammals. Two mouse *H2a.l* genes have duplicated/relocated onto the Y Chromosome and one onto Chromosome 2 (Ferguson et al. 2009). Moreover, the ancestral loci encoding *H2A.B.1.1* and *H2A.B.2* relocated away from the X Chromosome to autosomes in mouse (Chr 3 and Chr 16) and rat genomes (Chr 20 and Chr 18) as determined by flanking genes, and the encoded *H2A.B* genes have now been deleted or have decayed beyond recognition (Supplemental Fig. S2). However, the rat-mouse common ancestor acquired an intact *H2A.B* gene in a new X-linked locus (*H2a.b.ratMouse1*, also historically named H2A.B.3 [Soboleva et al. 2017]). The loss of autosomal *H2a.b* and retention of X-linked *H2a.b* suggests the intriguing possibility that X-linkage is evolutionarily preferred. However, it is also possible that X-linkage of short histone variants is simply explained by their evolutionary origins on this chromosome.

Taken together, our results show that a repertoire of seven or eight short H2A variants was acquired on the X Chromosome in the common ancestor of all eutherian mammals following the split with metatherians (marsupials) over 100 million years ago. Following this initial event, our results further show that short H2A variants have experienced a series of lineage-specific events of pseudogenization and duplication. Finally, our analysis indicates that functional H2A short variants have remained largely X-linked during the radiation of eutherian mammals.

Distinguishing structural features of short H2A variant proteins

Our phylogenetic analysis allowed us for the first time to study the amino acid features that constrained short H2A evolution. We compared protein sequences of canonical H2A and three short H2A variants (H2A.B, H2A.L, and H2A.P) from a set of eight diverse eutherian mammals spanning the same evolutionary divergence (Methods; Supplemental Data S2; Fig. 3). Because many clades have lost functional H2A.Q, we used a different, less divergent set of species for this variant. We aligned these sequences to each other and compared them to each other via Logo plots. We then used the nucleosome crystal structure (PDB:1AOI) (Luger et al. 1997) to annotate the histone fold domain as well as DNA-proximal and H2B-contacting residues.

We found several features that distinguish all short H2A variants from canonical H2A. We highlighted these features using a Jensen-Shannon distance (JSD) metric (Fig. 3 bar graphs; see Methods). Many residues within the alpha-helices and the loops of the HFD differ between short histone variants and canonical H2A (high JSD values in the upper bar graphs) but are conserved across short H2As (low JSD values in the lower bar graphs). We highlight one of these residues, in helix alpha 2, with an asterisk to illustrate this dichotomy. Such residues likely reflect residues whose evolution has been highly constrained since the common origin of all short histone variants.

As previously noted, short H2A histone variants have a shorter C-terminal tail and lack the acidic patch present in canonical H2A (Supplemental Fig. S1; Luger et al. 1997; Chakravarthy et al. 2004; Gautier et al. 2004; Eirin-Lopez et al. 2008; Gonzalez-Romero et al. 2008; Ishibashi et al. 2010; Bonisch and Hake 2012; Talbert et al. 2012; Rivera-Casas et al. 2016; Buschbeck and Hake 2017). These changes, as well as the high divergence of Loop 1, make it less likely that stable heterodimers will form between short histone H2A variants and canonical H2A. Our analyses also identified eight charge-altering positions within the HFD that appear to distinguish short histone H2A proteins from canonical H2A (Fig. 3, highlighted in green in JSD plots). These charge-altering residues are mostly acidic in the short H2A variants and are clustered around H2A::H2A contact sites in proximity to DNA (Supplemental Fig. S6; Luger et al. 1997). These charge-altering changes may further explain the lower stability and packing density of nucleosomes comprised of short histone H2A variants (Bao et al. 2004; Doyen et al. 2006; Syed et al. 2009; Barral et al. 2017).

Using the JSD metric, we also highlighted residues that distinguish short histone H2A variants from each other (high values for JSD H2A.B vs. H2A.L vs. H2A.P vs. H2A.Q) (Fig. 3). We found that functional H2A.Q proteins have a unique 7-aa deletion spanning Loop 2 and alpha-helix 3 (Supplemental Fig. S5), which, together with their short tails, imply that they are among the shortest eukaryotic histones to be described to date. Furthermore, H2A.P and H2A.Q have lost two key conserved arginine (R) residues in Loop 1 and 2 that contact the DNA minor-groove. Loss of these arginine residues is predicted to further reduce the stability of H2A.P- and H2A.Q-containing nucleosomes. Furthermore, we found that H2A.P specifically acquired many additional acidic changes (blue dots, Fig. 3), including sites predicted to contact DNA and H2B. In fact, charge-altering changes in H2A.P have been so extensive that the theoretical isoelectric point (pI) of full-length H2A.P averages 4.7 across the eight species used, which

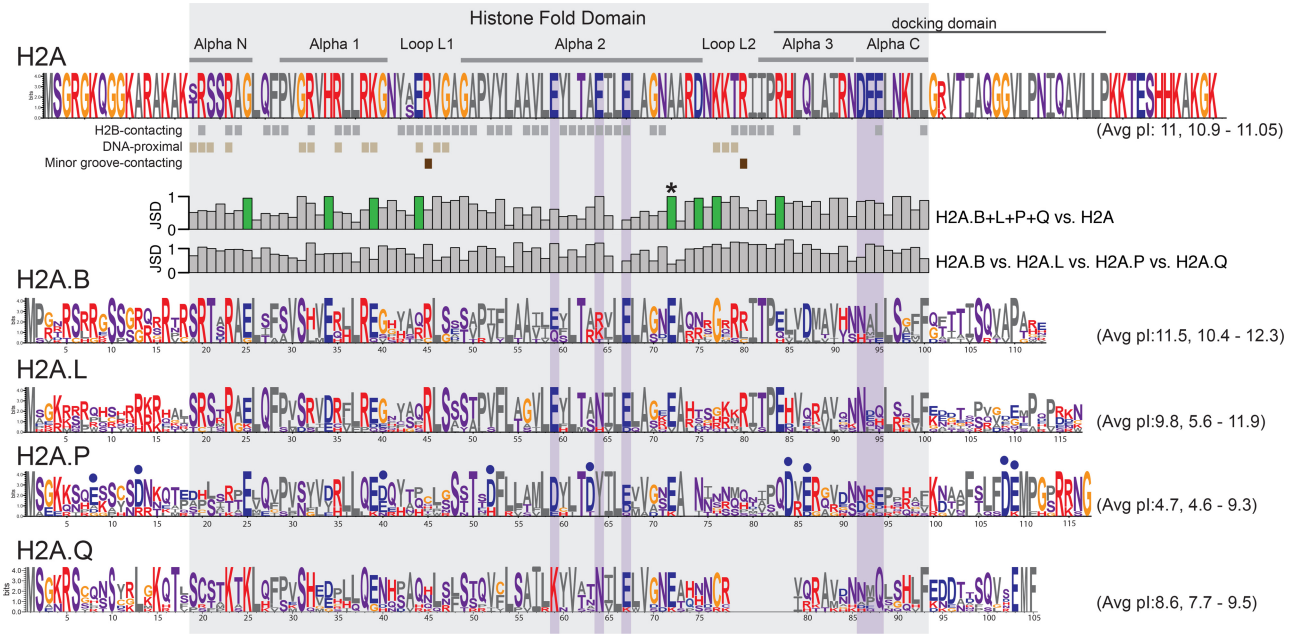

Figure 3. Short H2A variants have distinct protein features. Logo plots of protein alignments of canonical and short H2As across an identical set of representative eutherian mammals (except for H2A.Q) (see Methods). Residues are colored based on biochemical proprieties; hydrophobic residues in gray; positively charged in red; negatively charged in blue; polar in purple, and others (Gly and Cys) in orange. Residues in proximity to H2B (gray squares, at least 4 Å), DNA (light brown, 5 Å), or buried in the minor groove (dark brown) are annotated *below* the H2A logo. The histone fold domain and the six dispersed residues of the acidic patch are boxed in gray and purple, respectively. Jensen-Shannon distances (JSD) measured at each amino acid position between combinations of logos are indicated as bar graphs (low values meaning very similar, high values meaning very different). Charge-altering changes that are conserved among short H2As but different from canonical H2A are highlighted using green-colored bars. Residues uniquely altered to acidic residues in H2A.P are highlighted using blue dots. The average and range of isoelectric points (pI) across the selected mammals are shown in parentheses.

is dramatically different from the pI of canonical H2A (11.0) as well as other short histone H2A variants (Fig. 3). Such an acidic pI is highly unexpected for a DNA-packaging protein and would be predicted to result in destabilized histone-DNA interactions for H2A.P. Finally, more conserved features distinguishing H2A.B, H2A.L, and H2A.P from one another appear in their N- and C-terminal tails than in the HFD, e.g., the last 14 aa of H2A.P are more constrained than the rest of the protein. These distinguishing features suggest that each of the short histone H2A variants interact with different nonhistone proteins via their tails.

We also generated a Logo plot for the five representatives of the H2A.R clade. As suggested by our phylogeny (Fig. 1A), the Logo plot confirmed H2A.R variants' status as intermediate between the eutherian short histone H2A variants and canonical H2A (Supplemental Fig. S6). Like canonical H2As, H2A.R variants have long C-termini and a conserved docking domain. However, H2A.R variants also resemble short histone H2A variants in key residues of the HFD and have mutations in their acidic patch (Supplemental Fig. S6).

Concerted evolution of *H2A.B* and *H2A.L* gene duplicates

Canonical histone loci dispersed across mammalian genomes are highly homogeneous in sequence within species (Coen et al. 1982; Rooney et al. 2000; Marzluff et al. 2002, 2008; Nei and Rooney 2005). Their sequence homogeneity is maintained by a process of gene duplication and loss (birth-and-death) combined with strong purifying selection at the protein sequence level (Rooney et al. 2002). Our protein sequence phylogeny combined with our analysis of synteny (Fig. 1) revealed signatures of sequence homogenization between short H2As in the *H2A.B* and *H2A.L* genes that have continued to reside in the same syntenic genomic location for over 100 million years. We find that all copies of a variant from the same species cluster together in the phylogeny with remarkably high sequence identity (Fig. 1A). For example, panda H2A.B.1.1 and H2A.B.2 share 100% amino acid identity but only 75% identity with their respective syntenic orthologs in cat. These observations suggest the action of sequence homogenization occurring between dispersed histone variant genes within each genome.

To further explore this observation, we performed phylogenetic analyses based on nucleotide sequences, starting with primate *H2A.B* genes (Fig. 4A; Supplementary Data S4). We collected and aligned all *H2A.B*-related DNA sequences from the genomes of six primates representing roughly 40 My of evolution. In most cases, synteny allowed us to assign homologs to one of the three ancestral *H2A.B* loci (*H2A.B.1.1*, *H2A.B.1.2*, or *H2A.B.2*). First, we performed GARD analysis (Kosakovsky Pond et al. 2006) to search for internal recombination breakpoints but found none. However a maximum-likelihood phylogeny showed all *H2A.B* duplicates grouped by species rather than by syntenic location with nearly 100% identity (Fig. 4A), suggesting complete gene conversion.

However, it is also possible that unusual selective forces or extreme codon biases could result in such a phylogeny. To rule out selection or codon bias as explanations for such high within-species similarity, we compared genomic sequences surrounding the putatively converted copies using VISTA plots (Fig. 4B; Frazer et al. 2004). We found near identity between 50-kb regions encoding human *H2A.B.1.1* and human *H2A.B.1.2*, and a shorter 10-kb region of identity between genomic regions encoding human *H2A.B.1.1* and *H2A.B.2*, even though these regions are 576 kb apart in the genome. We find similar patterns in several other mammalian genomes, indicating that *H2A.B* duplicates have indeed been subject to recurrent gene conversion.

While we observed similar signatures of gene conversion between carnivore *H2A.L* genes (Supplemental Fig. S7; Supplementary Data S4), pseudogene *H2A.L* loci do not participate in this concerted evolution. For example, *H2A.L.2* has been a pseudogene since the common ancestor of ferret, walrus, seal, and panda and appears to have been exempt from sequence homogenization in those species. Similarly, primate *H2A.L.2* has

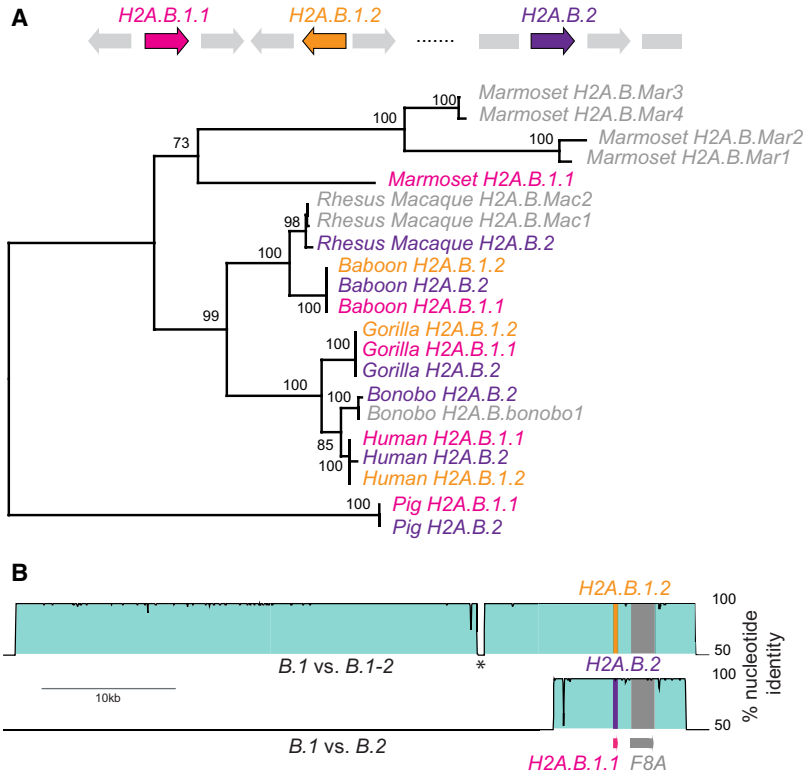

Figure 4. Concerted evolution by gene conversion among primate *H2A.B* genes. (*A*) Maximum-likelihood nucleotide phylogeny of representative primate *H2A.B* genes rooted using pig as an outgroup. Bootstrap values are shown for nodes with >50% support. A cartoon shows the loci containing ancestral duplications *H2A.B.1.1* (pink), *H2A.B.1.2* (orange), and *H2A.B.2* (purple), and gene names are colored accordingly in the tree. We also include some genes that are either in nonsyntenic loci (representing additional duplicates) or were on short scaffolds so could not be assigned to a locus (four from marmoset, two from rhesus macaque, and one from bonobo, denoted *.Mar1* to *.Mar4*; *.Mac1*, *.Mac2*; and *.bonobo1*, respectively). (*B*) VISTA plots showing very high nucleotide identity across large genomic regions flanking the three human *H2A.B* genes (proximal to *F8A*), comparing *H2A.B.1.1* genomic sequence as a reference to *H2A.B.1.2* (*upper* plot) and *H2A.B.2* (*lower* plot). An interruption in high identity (asterisk) is due to a recent transposon insertion.

been a pseudogene since before primate divergence and has been exempt from gene conversion since that time (Supplemental Fig. S4). In addition, consistent with our analysis of selective pressures of *H2A.L.1* and *H2A.L.3* loci from Old World monkeys and hominoids, these putative pseudogenes are no longer subject to gene conversion, whereas the (presumed functional) New World monkey *H2A.L* genes are (Supplemental Fig. S4; Supplementary Data S4). However, we do not observe gene conversion among the three ancestral *H2A.L* loci in rodents, although most are apparently functional (Supplemental Fig. S8). Examination of closely related *Mus* species also suggests that at least some recently expanded *H2a.l.2*-like genes in mouse have also not been subject to recent gene conversion. In these specific cases, we propose that lack of concerted evolution reflects ongoing diversification and functional specialization of subgroups of rodent *H2A.L* genes. Indeed, knockout of a single mouse *H2a.l* gene (*H2a.l.Chr2*) renders male mice infertile (Barral et al. 2017), showing that other *H2a.l* duplicates cannot compensate for its loss.

Overall, we conclude that recurrent gene conversion shaped the evolution of active *H2A.B* and *H2A.L* gene duplicates in most eutherian genomes. This mode of evolution appears distinct from canonical histone multigene evolution in many animal genomes (Piontkivska et al. 2002; Rooney et al. 2002).

Accelerated evolution and diversifying selection of short H2A variants

The long branch-lengths in the H2A phylogeny (Fig. 1A) suggest that short histone H2A variant clades appear to be evolving more rapidly than canonical or other H2A histones. This observation is consistent with previous analyses of *H2A.B* (Malik and Henikoff 2003; Eirin-Lopez et al. 2008; Ishibashi et al. 2010). We adopted two additional analyses to more rigorously explore the high divergence of short histone H2A variants. First, we estimated d_N/d_S across the species sets used for our previous logo plot analyses (Fig. 3; Methods). A d_N/d_S ratio close to 0 reflects very strong evolutionary constraints, whereas higher ratios reflect less severe constraints. We used PAML's codeml with NSsites model 0 (Yang 1997) (that assumes a single evolutionary rate across all sites and species) to obtain overall d_N/d_S values of 0.25 for *H2A.B*, 0.32 for *H2A.L* and 0.58 for *H2A.P* (Supplemental Table S2). All of these values are much higher than the d_N/d_S of 0.003 exhibited by canonical *H2A* and are comparable to or higher than the overall d_N/d_S of 0.30 for CENPA, the only mammalian histone known to be evolving under diversifying selection. In all cases, we can strongly reject neutral evolution ($d_N/d_S = 1$) for all short histone variant genes (Supplemental Table S2; see Methods), supporting the idea that they are functional protein-coding variants that contribute to mammalian fitness on some level. This is especially interesting in the case of *H2A.P* for which RNA expression but no protein expression has been observed (Baker et al. 2007, 2008; El Kennani et al. 2017). Our d_N/d_S analysis allows us to conclusively reject the model of neutral evolution ($P < 10^{-4}$), suggesting that *H2A.P* indeed must encode a protein.

We also compared divergence levels between different histone H2A variants (the four short H2A variants, H2A.R, canonical H2A, H2A.X, H2A.Z, macroH2A, and testis-specific H2A.1), the CENPA variant that is rapidly evolving (Malik and Henikoff 2001; Talbert et al. 2002), and a testis-specific H2B variant, H2B.1. We performed whole protein pairwise comparisons of each human protein to increasingly diverged orthologs, allowing us to examine the rate of protein evolution at different ages of divergence (Supplemental Table S4). Since human *H2A.L* and *H2A.Q* are pseudogenes, we relied on other pairwise comparisons for this analysis. We plotted amino acid identity as a function of species divergence time according to TimeTree (Fig. 5A; Hedges et al. 2015). Our analysis revealed very high conservation of canonical H2A and most other variant histones, including testis-specific H2A.1 and H2B.1, as well as H2A.R variants. In contrast, we found a much higher rate of divergence for CENPA, and for each of the short H2A variants, whose divergence rivals or exceeds CENPA.

The accelerated divergence we observe in *H2A.B*, *H2A.L*, *H2A.P*, and *H2A.Q* could have two possible explanations. Like CENPA (Schueler et al. 2010), the short H2A variants might be evolving under diversifying (positive) selection. Alternatively, elevated d_N/d_S rates could simply be the result of relaxed constraints. To distinguish these explanations, we collected simian primate *H2A.B* and *H2A.P* sequences from databases as well as by PCR amplification and Sanger sequencing. For *H2A.B*, we used a single representative from each species, because within-species gene conversion results in near-identity between paralogs in each genome (Fig. 4). We used maximum-likelihood approaches implemented in PAML (Yang 1997) and HyPhy (Pond et al. 2005) packages to ask whether a subset of codons evolves under positive selection. We obtained similar results whether we used phylogenies inferred from each alignment or the published species tree (Perelman et al. 2011).

From our analysis of 21 *H2A.P* orthologs, including two obtained by PCR and sequencing (Methods; Supplemental Data S3), we found a strong signature of diversifying selection (M7 vs. M8, P-value < 0.001; $d_N/d_S > 1$ for 26% of the sites, with average d_N/d_S of 2.9) (Fig. 5B). Five sites (codon 5, 34, 97, 98, and 114) had high posterior probabilities of evolving under diversifying selection (>0.9, M8 BEB) with several others just under that threshold (Fig. 5B). However, four of these sites also overlap CpG dinucleotides in one or more of the aligned species. CpG dinucleotides have a much higher mutation rate (Duncan and Miller 1980; Ehrlich et al. 1990), yet current methodologies (PAML or HyPhy) do not explicitly account for this. Although the influence of CpG sites is a strong caveat to the interpretation of diversifying selection for any gene, we note that the evidence for diversifying selection on *H2A.P* is comparable to that previously obtained for *CENPA* orthologs in primates (Schueler et al. 2010). Furthermore, FUBAR analysis (HyPhy package [Murrell et al. 2013]) additionally identifies codons 17, 43, 79, and 116 as having evolved under diversifying selection (posterior probability > 0.9); these signatures cannot all be explained by CpG hypermutation.

The evidence for diversifying selection is more equivocal for *H2A.B* (Fig. 5C). Our analysis used a total of 21 *H2A.B* sequences, including 14 from assembled genomes and seven obtained by PCR and sequencing (Methods; Supplemental Data S3). Although one set of starting parameters in PAML finds evidence for positive selection in primate H2A.B, this finding is not robust to the use of alternative codon models. However, FUBAR finds a single site, codon 38, with strong evidence for diversifying selection (posterior probability > 0.9) (Fig. 5C); this site also overlaps a CpG site. This site, found in alpha-helix 1 of the HFD, is also among those highlighted by the FEL, REL, and MEME algorithms as being under positive selection.

Despite high overall d_N/d_S values (Supplemental Table S2), we found no evidence of positive selection recurrently acting on any codon in *H2A.Q* or *H2A.L* in any clade we examined, although we note that functional loss precluded analysis of these genes in primates, a clade where we have excellent power to detect diversifying

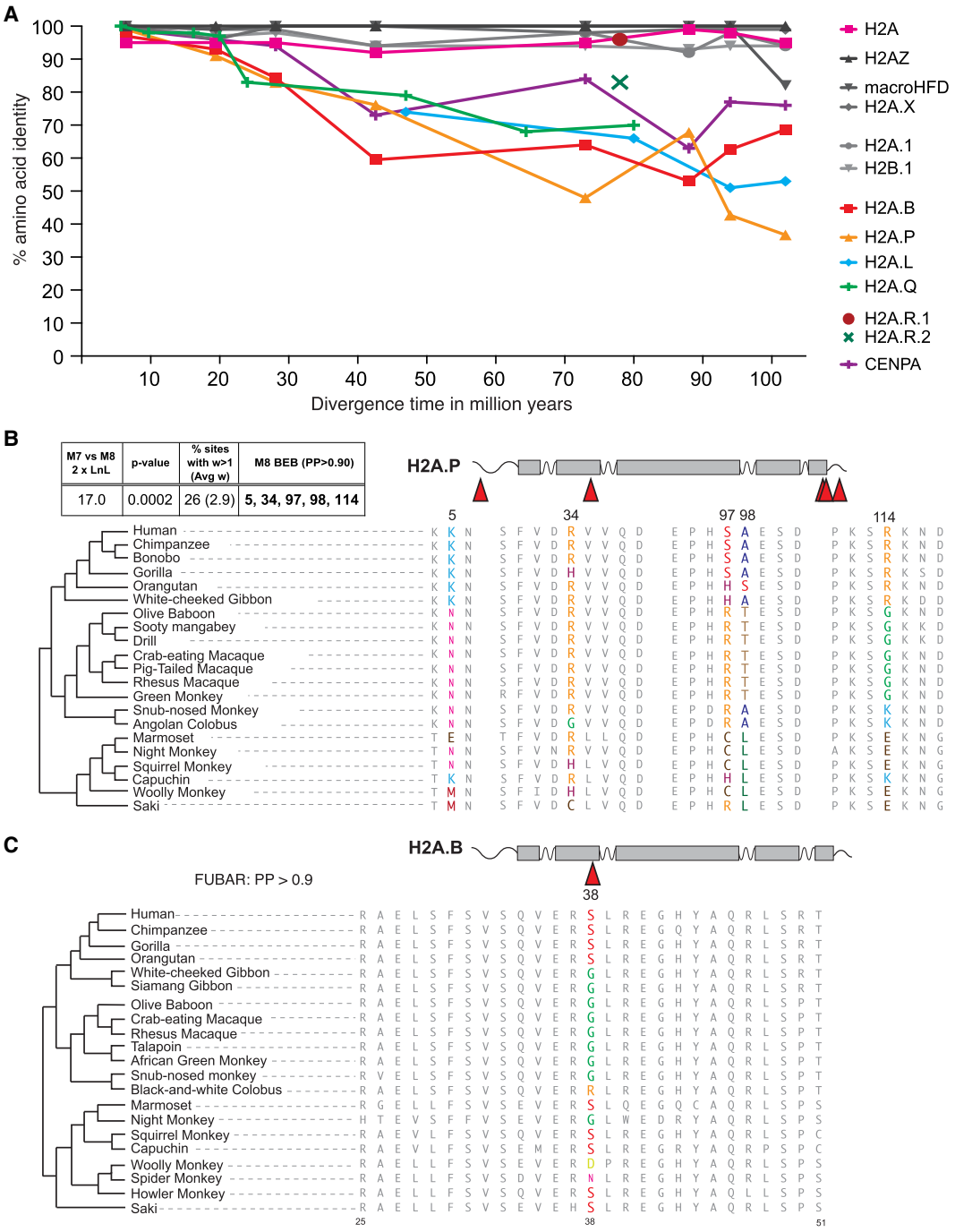

Figure 5. Rapid evolution of short H2As. (*A*) Pairwise amino acid identities (*y*-axis) between pairs of mammalian H2A proteins as a function of species divergence time (*x*-axis) (see Methods). Marsupial H2A.R proteins 1 and 2 are shown as single points, since each is represented by only two species diverged by 78 million years. (*B*) Portions of alignments surrounding positively selected sites in simian primate *H2A.P* genes (PAML, *top left* box). (*C*) A portion of the alignment of simian primate H2A.B proteins showing a positively selected codon at site 38 (colored amino acids), identified by FUBAR.

selection due to dense species sampling. Similarly, there are too few sequences available for *H2A.R* to allow well-powered analysis of selective pressures. Overall, our results indicate that *H2A.P* and possibly *H2A.B* have been subject to diversifying selection in simian primates, which could partly explain the greater divergence of short H2A histone variants compared to other H2A histones in mammals (Fig. 5A). However, much of the increased divergence of short histone H2A variants may be better explained by relaxed purifying selection.

Discussion

In this study, we have traced the origin and the evolutionary relationships of four short histone H2A variant families with unique

evolutionary trajectories in mammals. We infer that the last common eutherian mammal ancestor encoded seven or eight short H2A variant genes on the X Chromosome: two to three *H2A.B* genes, three *H2A.L* genes, a single *H2A.P* gene, and a single *H2A.Q* gene, which we have identified for the first time here. These variants were derived from a newly identified lineage of "canonically-sized" *H2A.R* variants, present in basally branching mammals. However, the repertoire of short H2A histone variants has since been remarkably plastic in different lineages, with some species entirely losing functional *H2A.B*, *H2A.P*, *H2A.Q,* or *H2A.L* genes, and other lineages experiencing additional duplications. The finding that no single short H2A variant is universally indispensable to all mammals might suggest that they are functionally nonessential or redundant with each other. Alternatively, the short histone variants might have variable functions between different mammals, whereby different short histone variants may have taken on similar functions in different lineages. Indeed, whereas mouse sperm chromatin contains H2A.L, human sperm contains H2A.B instead (Baker et al. 2007, 2008; El Kennani et al. 2017).

Our analyses also clearly distinguish active protein-coding genes from noncoding or pseudogenes. For example, although H2A.P protein has never been detected in mouse or human germ cells (Baker et al. 2007, 2008; El Kennani et al. 2017), our analysis strongly argues that *H2A.P* is indeed a functional protein-coding gene, as is *H2A.Q* in nonprimate mammals. Failure to detect H2A.P protein in testis samples may be due to expression in a limited subset of cells, or perhaps even because *H2A.P* mRNAs, which accumulate during spermatogenesis, may actually be delivered to the embryo and translated following fertilization. We note, however, our analysis cannot determine whether short H2A mRNAs, like those from *H2A.P*, might have noncoding functions in addition to their protein-coding capacity: this remains an open question.

We find that the four short H2A variants have amino acid features consistent with their "short-wrapping" properties (Fig. 3; Supplemental Fig. S1; Luger et al. 1997; Chakravarthy et al. 2004). Additionally, we find that all short histone variants differ from canonical H2A at some key positions, many of which are in or near Loop L1 of the HFD (Fig. 3). Such changes may therefore affect H2A::H2A interactions within the nucleosome and preclude formation of heterotypic nucleosomes that contain both canonical and short variant H2A proteins. These changes also suggest that short H2As can destabilize histone–histone and histone–DNA interactions. We find very few conserved residues in the HFD of short histone H2A variants that distinguish them from each other; instead, most of their specialization may stem from changes in the N- and C-terminal tails of these variants, especially in H2A.P and H2A.B. Such functional specialization due to expanded N-terminal tails has been seen previously; for example, H2B and linker histone H1 variants in sea urchins have SPKK motifs in their N-terminal tails that interact with the minor groove of linker DNA (Suzuki 1989; Poccia and Green 1992).

The structural evolution of short H2As raises the important question of their compatibility with other histone proteins within the nucleosome. There are several H2B and H3 variants that might interact with short H2As during male germ cell development. However, none of these variants display similar evolutionary trajectories to the short H2As. In addition to the high conservation of testis-specific H2B.1 presented here, there is no evidence that primate H2B.W (Churikov et al. 2004), mammalian subH2B (Aul and Oko 2001), H3.5, and H3t (Witt et al. 1996; Schenk et al. 2011) are subject to diversifying selection or to the degree of genetic turnover we have observed for short histone H2A variants in the present study. Although previous studies in mouse have shown that H2A.L-containing nucleosomes also contain H2B.1 (Govin et al. 2007), biochemical studies will be required to determine the nature of all short H2A-containing nucleosomes. It also seems likely that short H2A-containing nucleosome composition could vary across species, given the differences observed between species in their evolutionary retention (Fig. 1B).

Our analysis also revealed that genomically distant *H2A.B* and *H2A.L* paralogs are subject to recurrent gene conversion in multiple mammalian genomes. This sequence homogenization process suggests that, in many instances, the multiple copies of the same short histone variants within a species (e.g., *H2a.b* in mouse) are unlikely to perform different functions (Soboleva et al. 2017).

We find that short H2A variants show greater evolutionary divergence between species than even CENPA, the fastest-evolving histone variant examined to date (Malik and Henikoff 2001; Talbert et al. 2002). While this increased rate could be partly due to diversifying selection, we conclude that rapid divergence of short histone variants may primarily stem from their testis specificity; many male germ cell-specific proteins, especially those involved in DNA-packaging, are among the fastest-evolving proteins in mammalian genomes (Retief and Dixon 1993; Wyckoff et al. 2000; Torgerson et al. 2002; Martin-Coello et al. 2009). Since the only function demonstrated for short H2A variants is in protamine deposition (Barral et al. 2017), we speculate that the functional constraints acting on the short histone variants may be similar to those that act on protamines or sperm transition proteins that help deposit protamines.

The single autosomal *H2a.l.Chr2* gene in mouse is expressed at much higher levels than the sex-linked copies; knockout of this single gene is sufficient to cause male sterility (Barral et al. 2017). The X Chromosome is an interesting location for most short histone H2A variant genes that are expressed during male gametogenesis because most X-linked genes are silenced via a process known as meiotic sex chromosome inactivation during male meiosis (Turner 2015). However, a subset of genes, including short histone variants and genes located in the pseudoautosomal region of sex chromosomes, escape this silencing (Govin et al. 2007; Soboleva et al. 2012; Turner 2015). In addition, recent transcriptome analyses clearly show that hundreds of X-linked genes are re-activated in post-meiotic germ cells (Margolin et al. 2014). How this reactivation occurs and how expression levels compare to those of autosomal genes remain areas of active investigation. X-linked genes that do escape inactivation to be expressed in male meiosis are fully exposed to selection in the hemizygous state, and male-beneficial mutations can reach fixation even if they are harmful to females (Rice 1984; Ellegren and Parsch 2007). This mode of selection can drive the X Chromosome to accumulate and amplify sex-biased genes engaged in sexual antagonism (i.e., beneficial to one sex and detrimental or neutral to the other) (Ross et al. 2005; Ellegren 2011).

Their rapid divergence, X-linkage, and strong testis-specific expression suggest that short histone H2A variants in eutherian mammals may be engaged in a form of genetic conflict, either sexual antagonism or post-meiotic drive in spermatogenesis. For example, short histone variants could specifically package the chromatin of X Chromosome–containing germ cells to protect against the action of a Y-linked factor. Such a mechanism has been invoked to explain the recent, dramatic acquisition of chromatin genes on both the mouse X and Y Chromosomes

(Bachtrog 2014; Soh et al. 2014; Moretti et al. 2017). Although the nature of this genetic conflict is just emerging, our evolutionary studies firmly establish the evolutionary novelty in the origin and diversification of short histone H2A variants and highlight their unprecedented evolutionary signatures as impetus for their functional characterization.

Methods

Identification of short H2A variant genes

We iteratively queried the assembled genomes of 25 mammals using TBLASTN (Altschul et al. 1990, 1997) starting with human H2A.B.1.2 (NP_542451.1), rat H2a.l.2 (XP_002730244.1), and human H2A.P (NP_036406.1) as queries. In later iterations, we used additional closely related or better-conserved sequences (e.g., horse H2A.L.3) as queries. We retrieved nucleotide hits, recorded assembly coordinates, and identified syntenic locations using either the UCSC Genome Browser (Kent et al. 2002) or via BLASTN analyses (Altschul et al. 1990, 1997). We aligned all hits to other H2A sequences either manually or using MUSCLE or MAFFT aligners (Katoh et al. 2002; Edgar 2004). We used a combination of shared synteny and phylogenetic placement to classify ORFs and pseudogenes into one of the four short H2A families: *H2A.B*, *H2A.L*, *H2A.P*, or *H2A.Q*, and named them according to the ancestral syntenic locus in which they were found (e.g., 1, 1.1, 1.2, 2, or 3). If they were found outside syntenic loci or could not be reliably assigned there due to short contig size, we named the genes with an extension reflecting the species name and an arbitrary number (e.g., *H2A.L.cheetah1*).

Phylogenetic analyses

We used ClustalW (Larkin et al. 2007) to align predicted HFD protein sequences of canonical H2A, H2A.Z, H2A.X, macroH2A, and short H2As. We estimated a phylogenetic tree using maximum-likelihood methods implemented in PhyML (Guindon and Gascuel 2003; Guindon et al. 2010) with the Jones-Taylor-Thornton substitution model (Jones et al. 1992), 100 bootstrap replicates, and optimizing tree topology, length, and substitution rate.

For *H2A.B*, *H2A.L*, *H2A.P*, and *H2A.Q*, we aligned gene and pseudogene nucleotide sequences using MUSCLE (Edgar 2004) and built phylogenetic trees using maximum-likelihood methods (PhyML) (Guindon et al. 2010) using the HKY85 substitution model with 100 bootstrap replicates. Sequences were analyzed for evidence of recombination using the GARD algorithm implemented at datamonkey.org (Kosakovsky Pond et al. 2006).

Logos and nucleosome structure

Logo plots were generated using WebLogo (weblogo.berkeley.edu; Crooks et al. 2004) using H2A.B, H2A.L, H2A.P, and canonical H2A protein sequences from each of the following species: mouse, rat, Chinese hamster, pig, panda, leopard, rhinoceros, and armadillo. For H2A.Q, we used a less diverse panel of carnivore H2A.Q sequences. The isoelectric point for each of the proteins was computed using the ExPASy portal (Artimo et al. 2012) and then averaged. We displayed the published nucleosome structure (PDB:1AOI) (Luger et al. 1997) with some residues highlighted using the Chimera software (Pettersen et al. 2004).

To obtain a quantitative view of conservation between different H2A variants, we calculated a two-way Jensen-Shannon distance metric at each amino acid position in the histone fold domain as previously described (Doud et al. 2015). We also adapted the calculation to allow a four-way comparison (H2A.B vs. H2A.L vs. H2A.P vs. H2A.Q), calculating distance as follows:

$$\mathrm{JSD}_{4-\mathrm{way}} = \sqrt{(\mathrm{Term1} - \mathrm{Term2})} \text{ where}$$

$$\mathrm{Term1} = H((\mathrm{freq}_{\mathrm{H2AB}} + \mathrm{freq}_{\mathrm{H2AL}} + \mathrm{freq}_{\mathrm{H2AP}} + \mathrm{freq}_{\mathrm{H2AQ}})/4) \text{ and}$$

$$\mathrm{Term2} = (\mathrm{H}(\mathrm{freq}_{\mathrm{H2AB}}) + \mathrm{H}(\mathrm{freq}_{\mathrm{H2AL}}) + \mathrm{H}(\mathrm{freq}_{\mathrm{H2AP}}) + \mathrm{H}(\mathrm{freq}_{\mathrm{H2AQ}}))/4$$

Here, "freq" denotes a vector of length 20 representing amino acid frequencies at each position, and H denotes the Shannon entropy of such a vector. The possible range of a JSD for a two-way comparison is 0–1, whereas for a four-way comparison, the upper bound is 1.41, or $\sqrt{(\log_2(4))}$.

RNA-seq analysis

In order to examine expression of the short H2A variants, we analyzed publicly available transcriptome data from human, opossum, and platypus (Brawand et al. 2011), dog (Broad Institute), and pig (Wageningen University's FAANG project) (Yang et al. 2017). SRA identifiers are listed in Supplemental Table S3. We downloaded FASTQ files using NCBI's SRA toolkit (https://www.ncbi.nlm.nih.gov/books/NBK158900), and mapped reads to same-species genome assemblies using TopHat2 (Kim et al. 2013), using the "--max-multihits 1" option so that multiply-mapping reads were assigned randomly to a single location. We also obtained BED files of all annotated genes via the UCSC browser's table function (Kent et al. 2002), using RefSeq genes for human and Ensembl genes for dog, pig, opossum, and platypus. We then used BEDTools multicov (Quinlan and Hall 2010) and the coordinates given in Supplemental Table S1 to count reads overlapping each short H2A ORF and used R (R Core Team 2015) to divide those counts by the total number of mapped reads in each sample in millions, followed by the size of each transcript in kb to obtain RPKM values. For short H2A genes where the transcript extent is unclear, we used ORF coordinates to count reads and to calculate gene size for normalization. A handful of previously published housekeeping genes (Eisenberg and Levanon 2013) were selected for comparison (Supplemental Fig. S3), and orthologs in other species were identified using Ensembl gene trees (Zerbino et al. 2017).

Vista plots

To search for longer stretches of similarity in the loci flanking short H2A variant genes, we used dotter (Sonnhammer and Durbin 1995) to roughly define the extent of homology and mVISTA plots (Frazer et al. 2004) to examine genomic homology in more detail.

Calculating a rate of protein divergence for canonical and variant histones

We calculated pairwise identities between H2A variants and CENPA from alignments of full-length protein sequences to either human, dog, or Tasmanian devil reference sequences (Fig. 5A; Supplemental Table S4). Because macroH2A contains a large added C-terminal domain, we included only the HFD for better comparison with other H2As. We obtained species divergence times from the TimeTree database (www.timetree.org), which provides estimates based on median values from numerous published studies (Hedges et al. 2015).

PCR amplification and sequencing of primate *H2A.B* and *H2A.P*

To obtain additional primate *H2A.B* and *H2A.P* sequences, we amplified these genes from genomic DNA extracted from primate cell lines (Supplemental Table S5). All PCR products were TOPO-TA

cloned into the PCR4 vector from Invitrogen, according to the manufacturer's instructions. Purified plasmids were Sanger-sequenced using M13 primers.

Analysis of evolutionary selective pressures

We used the codeml algorithm from the PAML suite (Yang 1997) to test for positive selection on *H2A.B*, *H2A.Q*, and *H2A.P* in simian primates, and on *H2A.Q* in Cetartiodactyla. Codon alignments were generated using the online software PAL2NAL (Suyama et al. 2006) and analyzed in codeml using either the accepted species tree (Perelman et al. 2011) or a tree generated from the alignment using maximum-likelihood methods (PhyML) (Guindon et al. 2010) with the HKY85 substitution model. We compared "NSsites" evolutionary models that do not allow d_N/d_S to exceed 1 (M7 or M8a) to a model that does (M8). We tested for statistical significance using a χ^2 test of twice the difference in log-likelihoods between M8 and matched null model M7 or M8a, with the degrees of freedom reflecting the difference in number of parameters between the two models compared (Yang 1997). Positively selected sites were classified as those sites with M8 Bayes Empirical Bayes posterior probability >90%. The results we present are from codeml runs using the F3×4 codon frequency model and initial omega 0.4. Analyses were robust to use of different starting parameters (codon frequency model F61; starting omega 1.5) unless otherwise stated. We also used the FUBAR program implemented at datamonkey.org (Delport et al. 2010).

We also used codeml's model 0 to test various alignments for signatures of overall purifying selection: Model 0 assumes a single d_N/d_S value for all sites of the alignment. We estimated a phylogeny for each alignment as above and compared the log-likelihood of model 0 where d_N/d_S is freely estimated with the log-likelihood of model 0 where d_N/d_S is fixed at 1 (neutral evolution). We used a χ^2 test (1 degree of freedom) to determine whether twice the log-likelihoods difference between those two models is statistically significant (Yang 1997).

Data access

Sequences produced in this study have been submitted to GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers MF134875–MF134883.

Acknowledgments

We thank M. Hays, L. Kursel, T. Levin, C. Schroeder, and P. Talbert for comments on the manuscript, S. Ramachandran and C. Schroeder for help analyzing the nucleosome crystal structure, and M. Doud for suggesting the Jensen-Shannon distance metric. This work was supported by a postdoctoral fellowship awarded to A.M. by the Damon Runyon Cancer Research Foundation (DRG:2192-14) and grants from the NIH (National Institute of General Medical Sciences) R01 GM074108 and from the Howard Hughes Medical Institute to H.S.M. The funders played no role in study design, data collection and interpretation, or the decision to publish this study. H.S.M. is an Investigator of the Howard Hughes Medical Institute.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Angelov D, Verdel A, An W, Bondarenko V, Hans F, Doyen CM, Studitsky VM, Hamiche A, Roeder RG, Bouvet P, et al. 2004. SWI/SNF remodeling and p300-dependent transcription of histone variant H2ABbd nucleosomal arrays. *EMBO J* **23:** 3815–3824.

Arimura Y, Kimura H, Oda T, Sato K, Osakabe A, Tachiwana H, Sato Y, Kinugasa Y, Ikura T, Sugiyama M, et al. 2013. Structural basis of a nucleosome containing histone H2A.B/H2A.Bbd that transiently associates with reorganized chromatin. *Sci Rep* **3:** 3510.

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, et al. 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* **40:** W597–W603.

Aul RB, Oko RJ. 2001. The major subacrosomal occupant of bull spermatozoa is a novel histone H2B variant associated with the forming acrosome during spermiogenesis. *Dev Biol* **239:** 376–387.

Bachtrog D. 2014. Signs of genomic battles in mouse sex chromosomes. *Cell* **159:** 716–718.

Baker MA, Reeves G, Hetherington L, Muller J, Baur I, Aitken RJ. 2007. Identification of gene products present in Triton X-100 soluble and insoluble fractions of human spermatozoa lysates using LC-MS/MS analysis. *Proteomics Clin Appl* **1:** 524–532.

Baker MA, Hetherington L, Reeves GM, Aitken RJ. 2008. The mouse sperm proteome characterized via IPG strip prefractionation and LC-MS/MS identification. *Proteomics* **8:** 1720–1730.

Bao Y, Konesky K, Park YJ, Rosu S, Dyer PN, Rangasamy D, Tremethick DJ, Laybourn PJ, Luger K. 2004. Nucleosomes containing the histone variant H2A.Bbd organize only 118 base pairs of DNA. *EMBO J* **23:** 3314–3324.

Barral S, Morozumi Y, Tanaka H, Montellier E, Govin J, de Dieuleveult M, Charbonnier G, Coute Y, Puthier D, Buchou T, et al. 2017. Histone variant H2A.L.2 guides transition protein-dependent protamine assembly in male germ cells. *Mol Cell* **66:** 89–101.e108.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446:** 507–512.

Bonisch C, Hake SB. 2012. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res* **40:** 10719–10741.

Boussouar F, Rousseaux S, Khochbin S. 2008. A new insight into male genome reprogramming by histone variants and histone code. *Cell Cycle* **7:** 3499–3502.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.

Buschbeck M, Hake SB. 2017. Variants of core histones and their roles in cell fate decisions, development and cancer. *Nat Rev Mol Cell Biol* **18:** 299–314.

Chadwick BP, Willard HF. 2001. A novel chromatin protein, distantly related to histone H2A, is largely excluded from the inactive X chromosome. *J Cell Biol* **152:** 375–384.

Chakravarthy S, Bao Y, Roberts VA, Tremethick D, Luger K. 2004. Structural characterization of histone H2A variants. *Cold Spring Harb Symp Quant Biol* **69:** 227–234.

Churikov D, Siino J, Svetlova M, Zhang K, Gineitis A, Morton Bradbury E, Zalensky A. 2004. Novel human testis-specific histone H2B encoded by the interrupted gene on the X chromosome. *Genomics* **84:** 745–756.

Coen E, Strachan T, Dover G. 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. *J Mol Biol* **158:** 17–35.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.

Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26:** 2455–2457.

Doud MB, Ashenberg O, Bloom JD. 2015. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol* **32:** 2944–2960.

Doyen CM, Montel F, Gautier T, Menoni H, Claudet C, Delacour-Larose M, Angelov D, Hamiche A, Bednar J, Faivre-Moskalenko C, et al. 2006. Dissection of the unusual structural and functional properties of the variant H2A.Bbd nucleosome. *EMBO J* **25:** 4234–4244.

Draizen EJ, Shaytan AK, Marino-Ramirez L, Talbert PB, Landsman D, Panchenko AR. 2016. HistoneDB 2.0: a histone database with variants—an integrated resource to explore histones and their variants. *Database (Oxford)* **2016:** baw014.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* **287:** 560–561.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Ehrlich M, Zhang XY, Inamdar NM. 1990. Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. *Mutat Res* **238:** 277–286.

Eirin-Lopez JM, Ishibashi T, Ausio J. 2008. H2A.Bbd: a quickly evolving hypervariable mammalian histone that destabilizes nucleosomes in an acetylation-independent way. *FASEB J* **22:** 316–326.
Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29:** 569–574.
El Kennani S, Adrait A, Shaytan AK, Khochbin S, Bruley C, Panchenko AR, Landsman D, Pflieger D, Govin J. 2017. MS_HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones. *Epigenetics Chromatin* **10:** 2.
Ellegren H. 2011. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Nat Rev Genet* **12:** 157–166.
Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet* **8:** 689–698.
Erives AJ. 2017. Phylogenetic analysis of the core histone doublet and DNA topo II genes of Marseilleviridae: evidence of proto-eukaryotic provenance. *Epigenetics Chromatin* **10:** 55.
Ferguson L, Ellis PJ, Affara NA. 2009. Two novel mouse genes mapped to chromosome Yp are expressed specifically in spermatids. *Mamm Genome* **20:** 193–206.
Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32:** W273–W279.
Gautier T, Abbott DW, Molla A, Verdel A, Ausio J, Dimitrov S. 2004. Histone variant H2ABbd confers lower stability to the nucleosome. *EMBO Rep* **5:** 715–720.
Gonzalez-Romero R, Mendez J, Ausio J, Eirin-Lopez JM. 2008. Quickly evolving histones, nucleosome stability and chromatin folding: all about histone H2A.Bbd. *Gene* **413:** 1–7.
Gonzalez-Romero R, Rivera-Casas C, Ausio J, Mendez J, Eirin-Lopez JM. 2010. Birth-and-death long-term evolution promotes histone H2B variant diversification in the male germinal cell line. *Mol Biol Evol* **27:** 1802–1812.
Govin J, Escoffier E, Rousseaux S, Kuhn L, Ferro M, Thevenon J, Catena R, Davidson I, Garin J, Khochbin S, et al. 2007. Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis. *J Cell Biol* **176:** 283–294.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307–321.
Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. 2009. Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460:** 473–478.
Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32:** 835–845.
Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293:** 1098–1102.
Ishibashi T, Li A, Eirin-Lopez JM, Zhao M, Missiaen K, Abbott DW, Meistrich M, Hendzel MJ, Ausio J. 2010. H2A.Bbd: an X-chromosome-encoded histone involved in mammalian spermiogenesis. *Nucleic Acids Res* **38:** 1780–1789.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8:** 275–282.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059–3066.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.
Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* **184:** 868–871.
Kornberg RD, Thomas JO. 1974. Chromatin structure; oligomers of the histones. *Science* **184:** 865–868.
Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22:** 3096–3098.
Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286:** 964–967.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.
Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389:** 251–260.
Malik HS, Henikoff S. 2001. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics* **157:** 1293–1298.
Malik HS, Henikoff S. 2003. Phylogenomics of the nucleosome. *Nat Struct Biol* **10:** 882–891.
Malik HS, Henikoff S. 2009. Major evolutionary transitions in centromere complexity. *Cell* **138:** 1067–1082.
Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. 2014. Integrated transcriptome analysis of mouse spermatogenesis. *BMC Genomics* **15:** 39.
Marino-Ramirez L, Hsu B, Baxevanis AD, Landsman D. 2006. The Histone Database: a comprehensive resource for histones and histone fold-containing proteins. *Proteins* **62:** 838–842.
Martin-Coello J, Dopazo H, Arbiza L, Ausio J, Roldan ER, Gomendio M. 2009. Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness. *Proc Biol Sci* **276:** 2427–2436.
Marzluff WF, Gongidi P, Woods KR, Jin J, Maltais LJ. 2002. The human and mouse replication-dependent histone genes. *Genomics* **80:** 487–498.
Marzluff WF, Wagner EJ, Duronio RJ. 2008. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet* **9:** 843–854.
Moretti C, Serrentino ME, Ialy-Radio C, Delessard M, Soboleva TA, Tores F, Leduc M, Nitschke P, Drevet JR, Tremethick DJ, et al. 2017. SLY regulates genes involved in chromatin remodeling and interacts with TBL1XR1 during sperm differentiation. *Cell Death Differ* **24:** 1029–1044.
Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol Biol Evol* **30:** 1196–1205.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39:** 121–152.
Oliva R, Dixon GH. 1991. Vertebrate protamine genes and the histone-to-protamine replacement reaction. *Prog Nucleic Acid Res Mol Biol* **40:** 25–94.
Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet* **7:** e1001342.
Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* **25:** 1605–1612.
Piontkivska H, Rooney AP, Nei M. 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol Biol Evol* **19:** 689–697.
Poccia DL, Green GR. 1992. Packaging and unpackaging the sea urchin sperm genome. *Trends Biochem Sci* **17:** 223–227.
Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21:** 676–679.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.
R Core Team. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Retief JD, Dixon GH. 1993. Evolution of pro-protamine P2 genes in primates. *Eur J Biochem* **214:** 609–615.
Rice WR. 1984. Sex-chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735–742.
Rivera-Casas C, Gonzalez-Romero R, Cheema MS, Ausio J, Eirin-Lopez JM. 2016. The characterization of macroH2A beyond vertebrates supports an ancestral origin and conserved role for histone variants in chromatin. *Epigenetics* **11:** 415–425.
Rooney AP, Zhang J, Nei M. 2000. An unusual form of purifying selection in a sperm protein. *Mol Biol Evol* **17:** 278–283.
Rooney AP, Piontkivska H, Nei M. 2002. Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. *Mol Biol Evol* **19:** 68–75.
Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434:** 325–337.
Sandman K, Reeve JN. 2006. Archaeal histones and the origin of the histone fold. *Curr Opin Microbiol* **9:** 520–525.
Sandstedt SA, Tucker PK. 2004. Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. *Genome Res* **14:** 267–272.
Sansoni V, Casas-Delucchi CS, Rajan M, Schmidt A, Bonisch C, Thomae AW, Staege MS, Hake SB, Cardoso MC, Imhof A. 2014. The histone variant H2A.Bbd is enriched at sites of DNA synthesis. *Nucleic Acids Res* **42:** 6405–6420.
Schenk R, Jenke A, Zilbauer M, Wirth S, Postberg J. 2011. H3.5 is a novel hominid-specific histone H3 variant that is specifically expressed in the seminiferous tubules of human testes. *Chromosoma* **120:** 275–285.
Schueler MG, Swanson W, Thomas PJ, Program NCS, Green ED. 2010. Adaptive evolution of foundation kinetochore proteins in primates. *Mol Biol Evol* **27:** 1585–1597.

Shaytan AK, Landsman D, Panchenko AR. 2015. Nucleosome adaptability conferred by sequence and structural variations in histone H2A–H2B dimers. *Curr Opin Struct Biol* **32:** 48–57.
Soboleva TA, Nekrasov M, Pahwa A, Williams R, Huttley GA, Tremethick DJ. 2012. A unique H2A histone variant occupies the transcriptional start site of active genes. *Nat Struct Mol Biol* **19:** 25–30.
Soboleva TA, Parker BJ, Nekrasov M, Hart-Smith G, Tay YJ, Tng WQ, Wilkins M, Ryan D, Tremethick DJ. 2017. A new link between transcriptional initiation and pre-mRNA splicing: the RNA binding histone variant H2A.B. *PLoS Genet* **13:** e1006633.
Soh YQ, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159:** 800–813.
Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–GC10.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609–W612.
Suzuki M. 1989. SPKK, a new nucleic acid-binding unit of protein found in histone. *EMBO J* **8:** 797–804.
Syed SH, Boulard M, Shukla MS, Gautier T, Travers A, Bednar J, Faivre-Moskalenko C, Dimitrov S, Angelov D. 2009. The incorporation of the novel histone variant H2AL2 confers unusual structural and functional properties of the nucleosome. *Nucleic Acids Res* **37:** 4684–4695.
Talbert PB, Henikoff S. 2010. Histone variants—ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol* **11:** 264–275.
Talbert PB, Henikoff S. 2017. Histone variants on the move: substrates for chromatin dynamics. *Nat Rev Mol Cell Biol* **18:** 115–126.
Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. 2002. Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *Plant Cell* **14:** 1053–1066.
Talbert PB, Ahmad K, Almouzni G, Ausio J, Berger F, Bhalla PL, Bonner WM, Cande WZ, Chadwick BP, Chan SW, et al. 2012. A unified phylogeny-based nomenclature for histone variants. *Epigenetics Chromatin* **5:** 7.
Tolstorukov MY, Goldman JA, Gilbert C, Ogryzko V, Kingston RE, Park PJ. 2012. Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Mol Cell* **47:** 596–607.
Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol Evol* **19:** 1973–1980.
Turner JM. 2015. Meiotic silencing in mammals. *Annu Rev Genet* **49:** 395–412.
Weber CM, Henikoff S. 2014. Histone variants: dynamic punctuation in transcription. *Genes Dev* **28:** 672–682.
Witt O, Albig W, Doenecke D. 1996. Testis-specific expression of a novel human H3 histone gene. *Exp Cell Res* **229:** 301–306.
Wu F, Caron C, De Robertis C, Khochbin S, Rousseaux S. 2008. Testis-specific histone variants H2AL1/2 rapidly disappear from paternal heterochromatin after fertilization. *J Reprod Dev* **54:** 413–417.
Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403:** 304–309.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555–556.
Yang Y, Liang G, Niu G, Zhang Y, Zhou R, Wang Y, Mu Y, Tang Z, Li K. 2017. Comparative analysis of DNA methylome and transcriptome of skeletal muscle in lean-, obese-, and mini-type pigs. *Sci Rep* **7:** 39883.
Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2017. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761.

Received September 1, 2017; accepted in revised form February 13, 2018.