Similar evolutionary trajectories for retrotransposon accumulation in mammals

Reuben M Buckley[1], R Daniel Kortschak[2], Joy M Raison[1], David L Adelson[1,*]

1 Department of Genetics and Evolution, The University of Adelaide, North Tce, 5005, Adelaide, Australia

*** david.adelson@adelaide.edu.au**

Abstract

The factors guiding retrotransposon insertion site preference are not well understood. Different types of retrotransposons share common replication machinery and yet occupy distinct genomic domains. Autonomous long interspersed elements accumulate in gene-poor domains and their non-autonomous short interspersed elements accumulate in gene-rich domains. To determine genomic factors that contribute to this discrepancy we analysed the distribution of retrotransposons within the framework of chromosomal domains and regulatory elements. Using comparative genomics, we identified large-scale conserved patterns of retrotransposon accumulation across several mammalian genomes. Importantly, retrotransposons that were active after our sample-species diverged accumulated in orthologous regions. This suggested a conserved interaction between retrotransposon activity and conserved genome architecture. In addition, we found that retrotransposons accumulated at regulatory element boundaries in open chromatin, where accumulation of particular retrotransposon types depended on insertion size and local regulatory element density. From our results, we propose a model where density and distribution of genes and regulatory elements canalise the accumulation of retrotransposons. Through conservation of synteny, gene regulation and nuclear organisation, we have found that mammalian genomes follow similar evolutionary trajectories.

Introduction

An understanding of the dynamics of evolutionary changes in mammalian genomes is critical for understanding the diversity of mammalian biology. Most work on mammalian molecular evolution is on protein coding genes, based on the assumed centrality of their roles and because of the lack of appropriate methods to identify the evolutionary conservation of apparently non-conserved, non-coding sequences. Consequently, this approach addresses only a tiny fraction (less than 2%) of a species' genome, leaving significant gaps in our understanding of evolutionary processes [1, 2]. In this report we describe how large scale positional conservation of non-coding, repetitive DNA sheds light on the possible conservation of mechanisms of genome evolution, particularly with respect to the acquisition of new DNA sequences.

Mammalian genomes are hierarchically organised into compositionally distinct hetero- or euchromatic large structural domains [3]. These domains are largely composed of mobile self-replicating non-long terminal repeat (non-LTR) retrotransposons; with Long INterspersed Elements (LINEs) in heterochromatic regions and Short INterspersed Elements (SINEs) in euchromatic regions [4]. The predominant LINE in most mammals is the ∼6 kb long L1. This autonomously replicating element is responsible for the mobilisation of an associated non-autonomous SINE, usually ∼300 bp long. Together, LINEs and SINEs occupy approximately 30% of the human genome [2], replicate via a well characterised RNA-mediated copy-and-paste mechanism [5] and co-evolve with host genomes [6–8].

The accumulation of L1s and their associated SINEs into distinct genomic regions depends on at least one of two factors. 1) Each element's insertion preference for particular genomic regions and 2) the ability of particular genomic regions to tolerate insertions. According to the current retrotransposon accumulation model, both L1s and SINEs likely share the same insertion patterns constrained by local sequence composition. Therefore, their accumulation in distinct genomic regions is a result of region specific tolerance to insertions. Because L1s are believed to have a greater capacity than SINEs to disrupt gene regulatory structures, they are evolutionarily purged from gene-rich euchromatic domains at a higher rate than SINEs. Consequently, this selection asymmetry in euchromatic gene-rich regions causes L1s to become enriched in gene-poor heterochromatic domains [2, 9–11].

An important genomic feature, not explored in the accumulation model, is the chromatin structure that surrounds potential retrotransposon insertion sites. Retrotransposons preferentially insert into open chromatin [12, 13], which is usually found overlapping gene regulatory elements. As disruption of regulatory elements can often be harmful, this creates a fundamental evolutionary conflict for retrotransposons: their immediate replication may be costly to the overall fitness of the genome in which they reside. Therefore, rather than local sequence composition and/or tolerance to insertion alone, retrotransposon accumulation is more likely to be constrained by an interaction between retrotransposon expression, openness of chromatin, susceptibility of a particular site to alter gene regulation, and the capacity of an insertion to impact on fitness.

To investigate the relationship between retrotransposon activity and genome evolution, we began by characterising the distribution and accumulation of non-LTR retrotransposons within placental mammalian genomes. Next, we compared retrotransposon accumulation patterns in five separate evolutionary paths by humanising the repeat content (see methods) of the chimpanzee, rhesus macaque, mouse and dog genomes. Finally, we analysed human retrotransposon accumulation in large hetero- and euchromatic structural domains, focussing on regions surrounding genes, exons and regulatory elements. Our results suggested that accumulation of particular retrotransposon families follows from insertion into open chromatin found adjacent to regulatory elements and depends on local gene and regulatory element density. From this we propose a refined retrotransposon accumulation model in which random insertion of retrotransposons is primarily constrained by chromatin structure rather than local sequence composition.

Results

Species selection and retrotransposon classification

We selected human, chimpanzee, rhesus macaque, mouse and dog as representative placental species because of their similar non-LTR retrotransposon composition (Fig. S1-S2) and phylogenetic relationships. Retrotransposon coordinates were obtained from the UCSC repeat masker tables [14, 15] and non-LTR retrotransposon families were grouped according to repeat type and period of activity as determined by genome-wide defragmentation [16]. Retrotransposons were placed into the following groups; new L1s, old L1s, new SINEs and ancient elements (for families in each group see Fig. S2). New L1s and new SINEs are retrotransposon families with high clade specificity and activity, while old L1s and ancient elements (SINE MIRs and LINE L2s) are retrotransposon families shared across taxa. We measured sequence similarity within retrotransposon families as percentage mismatch from family consensus sequences [17] and confirmed that our classification of retrotransposon groups agreed with ancestral and clade-specific periods of retrotransposon activity (Fig. S3).

Genomic distributions of retrotransposons

To analyse the large scale distribution of retrotransposons, we segmented each species genome into adjacent 1 Mb regions, tallied retrotransposon distributions, performed principal component analysis (PCA) and pairwise correlation analysis (see methods). From the PCA, we found that new SINEs and ancient elements strongly associated with the two major principal components (PC1 and PC2). Depending on this association we identified PC1 and PC2 as “New SINE PC” and “Ancient PC” respectively, or the converse (Fig. 1a). This showed that retrotransposon families from the same group accumulated in the same genomic regions. For all species examined, new SINEs were enriched in regions with few new L1s, and in all species except mouse — where ancient elements and old L1s were co-located — ancient elements were enriched in regions with few old L1s (Fig. 1a, S4). This mouse discordance has probably resulted from the increased genome turnover seen in the rodent lineage [18] disrupting the distribution of ancestral retrotransposon families (Fig. S1-S2). As the relationship between mouse clade-specific new retrotransposons is maintained, this discordance does not impact on downstream analyses. These results show that most genomic context associations between retrotransposon families are conserved across our sample species.

Retrotransposon accumulation and chromatin environment

In human and mouse, LINEs and SINEs differentially associate with distinct chromatin environments [19]. To determine how our retrotransposon groups associate with chromatin accessibility, we obtained cell line Repli-Seq data [20] from the UCSC genome browser. Repli-Seq measures the timing of genome replication during S-phase, where accessible euchromatic domains replicate early and inaccessible heterochromatic domains replicate late. Across our segmented human genome, we found a high degree of covariation between mean replication timing in HUVEC cells and New SINE PC scores (Fig. 1c), new SINEs associated with early replication and new L1s associated with late replication. This result is probably not specific to HUVEC cells alone, since early and late replicating regions from various independent cell lines exhibit a high degree of overlap (Fig. S5). In addition, by splitting L1s into old and new groups, we observed a strong association between replication timing and retrotransposon age that was not reported in previous analyses [21]. To confirm these results, we analysed retrotransposon accumulation at the boundaries of previously identified replication domains (RDs) [22]. We focused primarily on early replicating domain (ERD) boundaries rather than late replicating domain (LRD) boundaries. ERD boundaries mark the transition from open chromatin states to closed chromatin states and overlap with topologically associated domain (TAD) boundaries [21]. Consistent with our earlier results, significant density fluctuations at ERD boundaries were only observed for new L1s and new SINEs (Fig. 1c). Because RD timing and genomic distributions of clade-specific retrotransposons are both largely conserved across human and mouse [23], these results suggest that the relationship between retrotransposon accumulation and RD timing may be conserved across mammals.

The genomic distribution of retrotransposons is conserved across species

Our results showed that the genomic distribution of retrotransposons was similar across species (Fig 1a). To determine whether our observations resulted from retrotransposon insertion into orthologous regions, we used coordinate mappings between species to humanise retrotransposon family distributions and PC scores (see methods). From this, we found that retrotransposon families in different species that identified as the same group, accumulated in regions with shared common ancestry (Fig. S6-S9). In addition, humanised genome segments from the 20% tails of the New SINE and Ancient PC score distributions showed high degrees of genomic overlap and associated with human RDs as described above (Fig. 1b). With regard to sequence conservation and retrotransposon accumulation, regions enriched for ancient elements shared the highest degree of pairwise similarity across our species (Fig. S10-S11). This demonstrates that regions enriched for ancient elements have likely been preserved throughout mammalian evolution [24, 25]. Our results are consistent with retrotransposon accumulation overlying a conserved ancient genome architecture.

Retrotransposon insertion in open chromatin surrounding regulatory elements

Retrotransposons preferentially insert into open chromatin, yet open chromatin usually overlaps gene regulatory elements. As stated above, this creates a fundamental evolutionary conflict for retrotransposons: their immediate replication may be costly to the overall fitness of the genome in which they reside. To investigate retrotransposon insertion/accumulation dynamics at open chromatin regions, we analysed DNase1 hypersensitive activity across 15 cell lines in both ERDs and LRDs. DNase1 hypersensitive sites obtained from the UCSC genome browser [1] were merged into DNase1 clusters and DNase1 clusters overlapping exons were excluded. As replication is sometimes cell type-specific we also constructed a set of constitutive ERDs and LRDs (cERDs and cLRDs) (see methods). Based on previous analyses, cERDs and cLRDs likely capture RD states present during developmental periods of heritable retrotransposition [26]. Our cERDs and cLRDs capture approximately 50% of the genome and contain regions representative of genome-wide intron and intergenic genome structure (Fig. S12). In both cERDs and cLRDs, we measured DNase1 cluster activity by counting the number of DNase1 peaks that overlapped each cluster. We found that DNase1 clusters in cERDs were much more active than DNase1 clusters in cLRDs (Fig. 2a). Next, we analysed retrotransposon accumulation both within and at the boundaries of DNase1 clusters. Consistent with disruption of gne regulation by retrotransposon insertion, non-ancient retrotransposon groups were depleted from DNase1 clusters (Fig. 2b). Intriguingly, ancient element density in DNase1 clusters remained relatively high, suggesting that some ancient elements may have been exapted. At DNase1 cluster boundaries after removing interval size bias (Fig. S13-S14) (see methods), retrotransposon density remained highly enriched in cERDs and close to expected levels in cLRDs (Fig. 2c). This suggests that chromatin is likely to be open at highly active cluster boundaries where insertion of retrotransposons is less likely to disrupt regulatory elements. These results are consistent with an interaction between retrotransposon insertion, open chromatin and regulatory activity, where insertions into open chromatin only persist if they do not interrupt regulatory elements.

Retrotransposon insertion size and regulatory element density

L1s and their associated SINEs differ in size by an order of magnitude, retrotranspose via the L1-encoded chromatin sensitive L1ORF2P and accumulate in compositionally distinct genomic domains [12, 13]. This suggests that retrotransposon insertion size determines observed accumulation patterns. L1 and *Alu* insertions occur via target-primed reverse transcription which is initiated at the 3′ end of each element. With L1 insertion, this process often results in 5′ truncation, causing extensive insertion size variation and an over representation of new L1 3′ ends, not seen with *Alu* elements (Fig. 3a). When we compared insertion size variation across cERDs and cLRDs we observed that smaller new L1s were enriched in cERDs and *Alu* elements showed no RD insertion size preference (Fig. 3b). The effect of insertion size on retrotransposon accumulation was estimated by comparing insertion rates of each retrotransposon group at DNase1 cluster boundaries in cERDs and cLRDs. We found that *Alu* insertion rates at DNase1 cluster boundaries were similarly above expected levels both in cERDs and cLRDs (Fig. 3c), whereas new L1 insertion rates at DNase1 cluster boundaries were further above expected levels in cERDs than cLRDs (Fig. 3d). By comparing the insertion rate of new L1s — retrotransposons that exhibited RD specific insertion size variation — we found a negative correlation between element insertion size and gene/regulatory element density. Thus smaller elements, such as *Alu* elements, accumulate more in cERDs than do larger elements, such as new L1s, suggesting that smaller elements are more tolerated.

Retrotransposon insertion within gene and exon structures

Regulatory element organisation is largely shaped by gene and exon/intron structure which likely impacts the retrotransposon component of genome architecture. Therefore, we analysed retrotransposons and DNase1 clusters (exon overlapping and not exon overlapping) at the boundaries of genes and exons. Human RefSeq gene models were obtained from the UCSC genome browser and both intergenic and intronic regions were extracted (Table S4). At gene (Fig. 4a) and exon (Fig. 4b) boundaries, we found a high density of exon overlapping DNase1 clusters and depletion of retrotransposons. This created a depleted retrotransposon boundary zone (DRBZ) specific for each retrotransposon group, a region extending from the gene or exon boundary to the point where retrotransposon levels begin to increase. The size of each DRBZ correlated with the average insertion size of each retrotransposon group, suggesting larger retrotransposons may have a greater capacity to disrupt important structural and regulatory genomic features. We also found that in cERDs the $5'$ gene boundary *Alu* DRBZ was larger than the $3'$ gene boundary *Alu* DRBZ. This difference was associated with increased exon overlapping DNase1 cluster density at $5'$ gene boundaries in cERDs (Fig. 4a), emphasising the importance of evolutionary constraints on promoter architecture. For ancient elements, their interval size corrected density approximately 1 kb from the $5'$ gene boundary was significantly higher than expected. This increase is consistent with exaptation of ancient elements into regulatory roles [27] (Fig. S15-S18). Moreover, the density peak corresponding to uncorrected ancient elements also overlapped with that of not exon overlapping DNase1 clusters (Fig. 4a). Collectively, these results demonstrate the evolutionary importance of maintaining gene structure and regulation and how this in turn has canalised similar patterns of accumulation and distribution of retrotransposon families in different species over time.

Discussion

In our study, we compared several mammalian genomes and analysed chromatin structure at both small and large scales to better characterise retrotransposon accumulation. Our genome-wide comparisons across species were consistent with previous analyses that reported high levels of positional conservation for L1s and their associated SINEs [28, 29]. Because new L1s and new SINEs underwent periods of activity after each of our sample species diverged form a common ancestor [16], our observations are likely the result of a conserved interaction between retrotransposon activity and genome architecture. Previous analyses have attempted to capture this interaction through various retrotransposon accumulation models [2]. Based on large-scale conservation of genome architecture and GC content [28, 29], the current model of retrotransposon accumulation suggests that random insertion of L1s and SINEs are similarly constrained by local sequence composition, where L1s are quickly purged from gene-rich regions via purifying selection at a higher rate than SINEs [9–11]. However, this model fails to account for the demonstrated impact of chromatin structure on insertion site preference [12, 13].

From analysing retrotransposon architecture and local chromatin structure we found that 1) following preferential insertion into open chromatin domains, retrotransposons were tolerated adjacent to regulatory elements where they were less likely to cause harm; 2) element insertion size was a key factor affecting retrotransposon accumulation, where large elements accumulated in gene poor regions where they were less likely to perturb gene regulation; and 3) insertion patterns surrounding regulatory elements were persistent at the gene level. Based on these results, we propose a significant change to the current retrotransposon accumulation model; rather than random insertion constrained by local sequence composition, we propose that insertion is instead primarily constrained by local chromatin structure. Following this, L1s and SINEs both preferentially insert into gene/regulatory element rich euchromatic domains, where L1s with their relatively high mutational burden are quickly eliminated via purifying selection at a much higher rate than SINEs. Over time this results in an enrichment of SINEs in euchromatic domains and an enrichment of L1s in heterochromatic domains.

Conclusions

In conjunction with large scale conservation of synteny [30], gene regulation [31] and the structure of RDs/TADs [23, 32], our findings suggest that large scale positional conservation of old and new non-LTR retrotransposons results from their association with the regulatory activity of large genomic domains. From this, we conclude that similar constraints on insertion and accumulation of retrotransposons in different species can define common trajectories for genome evolution.

Methods

Within species comparisons of retrotransposon genome distributions

Retrotransposon coordinates for each species were initially identified using RepeatMasker and obtained from UCSC genome browser (Table S1) [14, 15]. We grouped retrotransposon elements based on repeat IDs used in Giordano *et al* [16]. Retrotransposon coordinates were extracted from hg19, mm9, panTro4, rheMac3, and canFam3 assemblies. Each species genome was segmented into 1 Mb regions and the density of each retrotransposon family for each segment was calculated. From this, each species was organised into an n-by-p data matrix of n genomic segments and p retrotransposon families. Genome distributions of retrotransposons were then analysed using principle component analysis (PCA) and correlation analysis. For correlation analysis, for each retrotransposon family we calculated Pearson's correlation coefficient for each retrotransposon family across our genome segments.

Across species comparisons of retrotransposon genome distributions

To compare genome distributions across species, we humanised a query species genome using mapping coordinates extracted from net AXT alignment files located on the UCSC genome browser (Table S1). First, genomes were filtered by discarding segments below a minimum mapping fraction threshold, removing poorly represented regions (Fig S19a). Next, we used mapping coordinates to match fragments of query species segments to their corresponding human segments (Fig S19b). From this, the retrotransposon content and PC scores of the matched query segments were humanised following equation 1 (Fig S19c).

$$c_i^* = \frac{\sum_j c_{ij} l_j^Q / q_j}{\sum_j l_j^R / r}, \tag{1}$$

where c_{ij} is the density of retrotransposon family i in query segment j, l_j^Q is the total length of the matched fragments between query segment j and the reference segment, l_j^R is the total length of the reference segment fragments that match query segment j, q_j is the total length of the query segment j, and r is the total length of the reference segment. The result c_i^* is the humanised coverage fraction of retrotransposon family i that can now be compared to a specific reference segment. Once genomes were humanised, Pearson's correlation coefficient was used to determine the conservation between retrotransposon genomic distributions (Fig S19d). Using the Kolmogorov-Smirnov test, we measured the effect of humanising by comparing the humanised query retrotransposon density distribution to the query filtered retrotransposon density distribution (Fig S19e). The same was done to measure the effect of filtering by comparing the segmented human retrotransposon density distribution to the human filtered retrotransposon density distribution (Fig S19f). Spatial correlations and the P-values from measuring the effects of humanising and filtering were integrated into a heatmap (Fig S19g). The entire process was repeated several times at different minimum mapping fraction thresholds to optimally represent each retrotransposon families genomic distribution in a humanised genome (fig S20).

Replication timing boundaries and constitutive replication timing domains

ERDs, LRDs, and timing transition regions (TTRs) for each dataset were previously identified using a deep neural network hidden Markov model (Table S2) [22]. To determine RD boundary fluctuations of retrotransposon density, we defined ERD boundaries as the boundary of a TTR adjacent to an ERD. ERD boundaries from across each sample were pooled and retrotransposon density was calculated for 50 kb intervals from regions flanking each boundary 1 Mb upstream and downstream. Expected density and standard deviation for each retrotransposon group was derived from a background distribution generated by calculating the mean of 500 randomly sampled 50 kb genomic bins within 2000 kb of each ERD boundary, replicated 10000 times. We also obtained Repli-Seq replication timing profiles from the UCSC genome browser as a wavelet signal (Table S2) [33]. For each of our 50 kb intervals we calculated the mean replication timing from across each Repli-Seq sample. To identify cERDs and cLRDs, ERDs and LRDs classified by Liu *et al* [22] across each cell type were split into 1 kb intervals to find the intersection. If the classification of 12 out of 16 samples agreed at a certain region, we classified that region as belonging to a cERDs or a cLRDs, depending on that region's majority classification.

DNase1 cluster identification and activity

DNase1 sites across 15 cell lines were found using DNase-seq and DNase-chip as part of the open chromatin synthesis dataset for ENCODE (Table S3) [1]. Regions where P-values of contiguous base pairs were below 0.05 were identified as significant DNase1 hypersensitive sites [1]. From this we extracted significant DNase1 hypersensitive sites from each sample and pooled them. DNase1 hypersensitive sites were then merged into DNase1 clusters. Cluster activity was calculated as the number of total overlapping pooled DNase1 hypersensitive sites. We also extracted intervals between adjacent DNase1 clusters to look for enrichment of retrotransposons at DNase1 cluster boundaries.

Extraction of intergenic and intron intervals

hg19 RefSeq gene annotations obtained from UCSC genome browser were used to extract a set of introns and intergenic intervals (Table S4). RefSeq gene annotations were merged and intergenic regions were classified as regions between the start and end of merged gene models. We used the strandedness of gene model boundaries to classify adjacent intergenic region boundaries as upstream or downstream. We discarded intergenic intervals adjacent to gene models where gene boundaries were annotated as both $+$ and $-$ strand. Regions between adjacent RefSeq exons within a single gene model were classified as introns. Introns interrupted by exons in alternatively spliced transcripts and introns overlapped by other gene models were excluded. Upstream and downstream intron boundaries were then annotated depending on the strandedness of the gene they were extracted from.

Interval boundary density of retrotransposons

Intervals were split in half and positions were reckoned relative to the feature adjacent boundary, where the feature was either a gene, exon, or DNase1 cluster (Fig S21). To calculate the retrotransposon density at each position, we measured the fraction of bases at each position annotated as a retrotransposon. Next, we smoothed retrotransposon densities by calculating the mean and standard deviation of retrotransposon densities within an expanding window, where window size grew as a function of distance from the boundary as position depth decreased. This made it possible to accurately compare the retrotransposon density at positions where retrotransposon insertions were sparse and density levels at each position fluctuated drastically. At positions with a high base pair density a small window was used and at positions with a low base pair density a large window was used. Expected retrotransposon density p was calculated as the total proportion of bases covered by retrotransposons across all intervals. Standard deviation at each position was calculated as $\sqrt{npq}$, where n is the total number of bases at a given position and q is equal to $1 - p$.

Interval size bias correction of retrotransposon densities

Interval boundary density is sensitive to retrotransposon insertion preferences into intervals of a certain size (Fig S22). To determine interval size retrotransposon density bias, we grouped intervals according to size and measured the retrotransposon density of each interval size group. Retrotransposon density bias was calculated as the observed retrotransposon density of an interval size group divided by the expected retrotransposon density, where the expected retrotransposon density is the total retrotransposon density across all intervals. Next, using the intervals that contribute to the position depth at each position adjacent to feature boundaries, we calculated the mean interval size. From this we corrected retrotransposon density at each position by dividing the observed retrotransposon density by the retrotransposon density bias that corresponded with that position's mean interval size.

Software and data analysis

All statistical analyses were performed using R [34] with the packages GenomicRanges [35] and rtracklayer [36]. R scripts used to perform analyses can be found at:
https://github.com/AdelaideBioinfo/retrotransposonAccumulation .

Additional Files

Additional file 1 — Supplementary information

Figures S1–S22, Tables S1–S4.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

R.M.B., R.D.K., J.M.R., and D.L.A. designed research; R.M.B. performed research; and R.M.B., R.D.K., and D.L.A. wrote the paper.

Acknowledgements

For reviewing our manuscript and providing helpful advice we would like to thank the following: Simon Baxter, Atma Ivancevic and Lu Zeng from the University of Adelaide; Kirsty Kitto from Queensland University of Technology; and Udaya DeSilva from Oklahoma State University.

Availability of data and materials

All data was obtained from publicly available repositories, urls can be found in supporting material (Table S1–S3). R scripts used to perform analyses can be found at https://github.com/AdelaideBioinfo/retrotransposonAccumulation.

References

1. ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

2. Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

3. Johan H Gibcus and Job Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782, 2013.

4. Patrik Medstrand, Louie N Van De Lagemaat, and Dixie L Mager. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome research*, 12(10):1483–1495, 2002.

5. Gregory J Cost, Qinghua Feng, Alain Jacquier, and Jef D Boeke. Human l1 element target-primed reverse transcription in vitro. *The EMBO Journal*, 21(21):5899–5910, 2002.

6. DA Kramerov and NS Vassetzky. Origin and evolution of sines in eukaryotic genomes. *Heredity*, 107(6):487–495, 2011.

7. Domitille Chalopin, Magali Naville, Floriane Plard, Delphine Galiana, and Jean-Nicolas Volff. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome biology and evolution*, 7(2):567–580, 2015.

8. Anthony V Furano, David D Duvernell, and Stephane Boissinot. L1 (line-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends in Genetics*, 20(1):9–14, 2004.

9. Todd Graham and Stephane Boissinot. The genomic distribution of l1 elements: the role of insertion bias and natural selection. *BioMed Research International*, 2006, 2006.

10. Stephen L Gasior, Graeme Preston, Dale J Hedges, Nicolas Gilbert, John V Moran, and Prescott L Deininger. Characterization of pre-insertion loci of de novo l1 insertions. *Gene*, 390(1):190–198, 2007.

11. Erika M Kvikstad and Kateryna D Makova. The (r) evolution of sine versus line distributions in primate genomes: sex chromosomes are important. *Genome research*, 20(5):600–613, 2010.

12. Gregory J Cost, Amit Golding, Mark S Schlissel, and Jef D Boeke. Target dna chromatinization modulates nicking by l1 endonuclease. *Nucleic acids research*, 29(2):573–577, 2001.

13. J Kenneth Baillie, Mark W Barnett, Kyle R Upton, Daniel J Gerhardt, Todd A Richmond, Fioravante De Sapio, Paul M Brennan, Patrizia Rizzu, Sarah Smith, Mark Fell, et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, 479(7374):534–537, 2011.

14. Kate R Rosenbloom, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2015 update. *Nucleic acids research*, 43(D1):D670–D681, 2015.

15. Arian FA Smit, Robert Hubley, and Phil Green. Repeatmasker open-3.0, 1996.

16. Joti Giordano, Yongchao Ge, Yevgeniy Gelfand, György Abrusán, Gary Benson, and Peter E Warburton. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput Biol*, 3(7):e137, 2007.

17. Weidong Bao, Kenji K Kojima, and Oleksiy Kohany. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1):1, 2015.

18. William J Murphy, Denis M Larkin, Annelie Everts-van der Wind, Guillaume Bourque, Glenn Tesler, Loretta Auvil, Jonathan E Beever, Bhanu P Chowdhary, Francis Galibert, Lisa Gatzke, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734):613–617, 2005.

19. Hiroki Ashida, Kiyoshi Asai, and Michiaki Hamada. Shape-based alignment of genomic landscapes in multi-scale resolution. *Nucleic acids research*, 40(14):6435–6448, 2012.

20. R Scott Hansen, Sean Thomas, Richard Sandstrom, Theresa K Canfield, Robert E Thurman, Molly Weaver, Michael O Dorschner, Stanley M Gartler, and John A Stamatoyannopoulos. Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144, 2010.

21. Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014.

22. Feng Liu, Chao Ren, Hao Li, Pingkun Zhou, Xiaochen Bo, and Wenjie Shu. De novo identification of replication-timing domains in the human genome by deep learning. *Bioinformatics*, page btv643, 2015.

23. Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C Schulz, Allan J Robins, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20(6):761–770, 2010.

24. David L Adelson, Joy M Raison, and Robert C Edgar. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences*, 106(31):12855–12860, 2009.

25. DL Adelson, JM Raison, M Garber, and RC Edgar. Interspersed repeats in the horse (equus caballus); spatial correlations highlight conserved chromosomal domains. *Animal genetics*, 41(s2):91–99, 2010.

26. Juan Carlos Rivera-Mulia, Quinton Buckley, Takayo Sasaki, Jared Zimmerman, Ruth A Didier, Kristopher Nazor, Jeanne F Loring, Zheng Lian, Sherman Weissman, Allan J Robins, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome research*, 2015.

27. Craig B Lowe, Gill Bejerano, and David Haussler. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, 104(19):8005–8010, 2007.

28. Asif T Chinwalla, Lisa L Cook, Kimberly D Delehaunty, Ginger A Fewell, Lucinda A Fulton, Robert S Fulton, Tina A Graves, LaDeana W Hillier, Elaine R Mardis, John D McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

29. Richard A Gibbs, George M Weinstock, Michael L Metzker, Donna M Muzny, Erica J Sodergren, Steven Scherer, Graham Scott, David Steffen, Kim C Worley, Paula E Burch, et al. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, 2004.

30. Bhanu P Chowdhary, Terje Raudsepp, Lutz Frönicke, and Harry Scherthan. Emerging patterns of comparative genome organization in some mammalian species as revealed by zoo-fish. *Genome research*, 8(6):577–589, 1998.

31. Esther T Chan, Gerald T Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A Zirngibl, Jane Aubin, Michael JH Ratcliffe, Andrew Wilde, Michael Brudno, et al. Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3):1, 2009.

32. Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

33. R. Scott Hansen, Sean Thomas, Richard Sandstrom, Theresa K. Canfield, Robert E. Thurman, Molly Weaver, Michael O. Dorschner, Stanley M. Gartler, and John A. Stamatoyannopoulos. Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144, 2010.

34. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

35. Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

36. Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009.

Figures

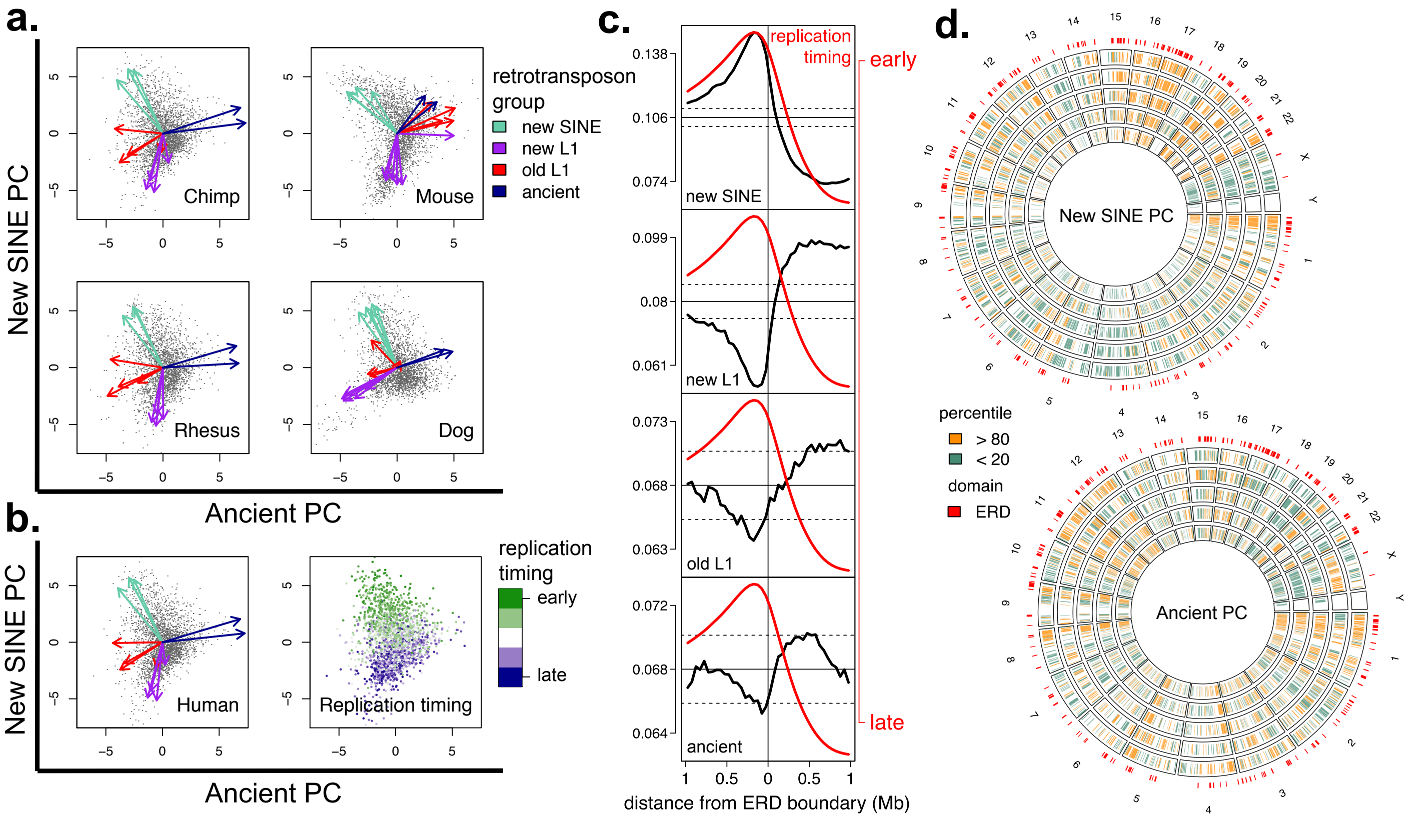

Figure 1. Large-scale genome distributions of retrotransposons are strongly associated with replication timing and conserved in distant mammalian species. **a**, PCA of non-human genome retrotransposon content, each vector loading has been coloured according to the retrotransposon group it represents. PC1 and PC2 have been renamed according to the retrotransposon group whose variance they principally account for. **b**, PCA of human retrotransposon content and mean genome replication timing in HUVEC cells. **c**, Retrotransposon density per non-overlapping 50 kb intervals from a pooled set of ERD boundaries across all 16 cell lines. Black dashed lines indicate 2 standard deviations from the mean (solid horizontal black line). Red line indicates mean replication timing across all samples. **d**, 20% tails of New SINE and Ancient PC scores of humanised genomes plotted against human, large ERDs (> 2 Mb) from HUVEC cells are marked in red. Species from centre are human, chimpanzee, rhesus macaque, mouse and dog.

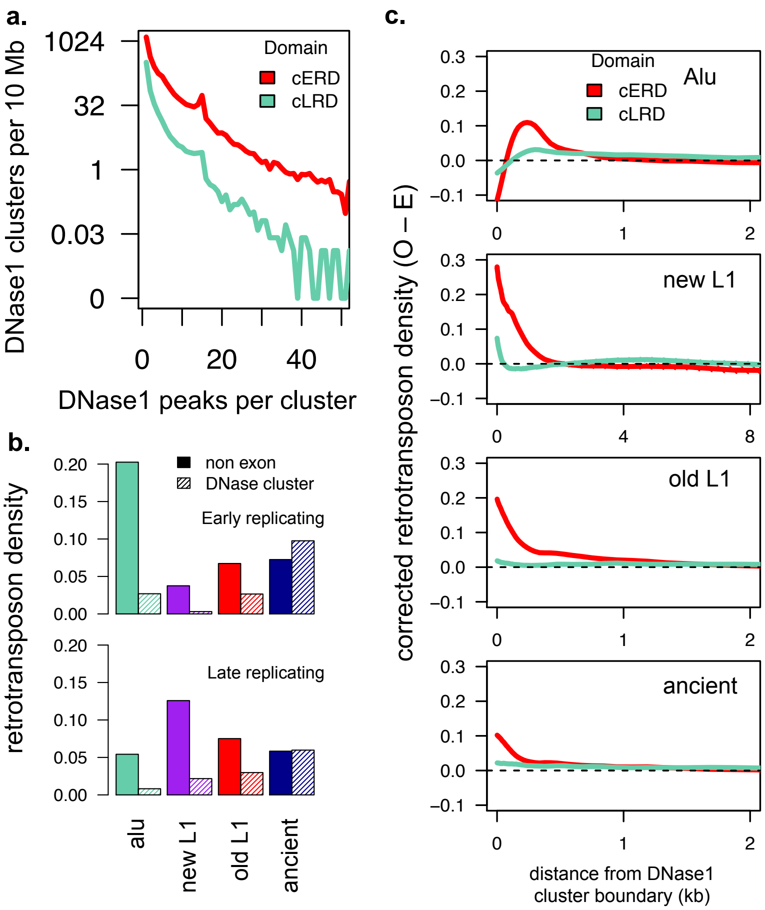

Figure 2. Retrotransposon accumulation occurs in open chromatin near regulatory regions. a, The activity of DNase1 clusters in cERDs and cLRDs. DNase1 clusters were identified by merging DNase1 hypersensitive sites across 15 tissues. Their activity levels were measured by the number of DNase1 hypersensitive sites overlapping each DNase1 cluster. **b**, Retrotransposon density of non-exonic regions and DNase1 clusters in cERDs and cLRDs. **c**, Observed minus expected retrotransposon density at the boundary of DNase1 clusters corrected for interval size bias (see methods). Expected retrotransposon density was calculated as each group's non-exonic total retrotransposon density across cERDs and cLRDs. A confidence interval of 3 standard deviations from expected retrotransposon density was also calculated, however the level of variation was negligible.

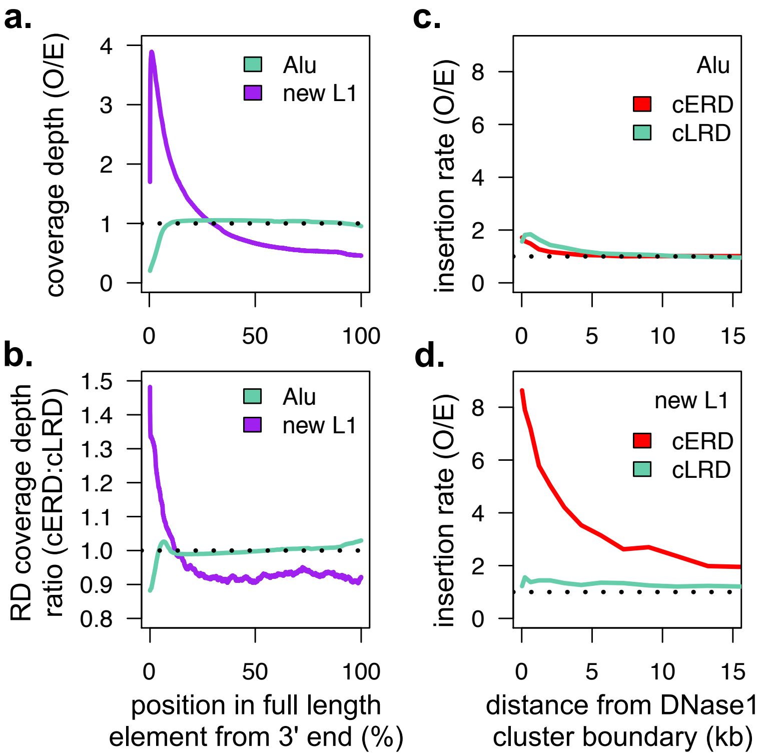

Figure 3. Retrotransposon insertion size is inversely proportional to local regulatory element density. **a**, Observed to expected ratio of retrotransposon position coverage depth measured from consensus 3′ end. Expected retrotransposon position coverage depth was calculated as total retrotransposon coverage over consensus element length. We used 6 kb as the consensus new L1 length and 300 bp as the consensus *Alu* length. **b**, New L1 and *Alu* position density ratio (cERDs:cLRDs). **c**, *Alu* and **d**, new L1 observed over expected retrotransposon insertion rates at DNase1 cluster boundaries in cERDs and cLRDs. Insertion rates were measured by prevalence of 3′ ends and expected levels were calculated as the per Mb insertion rate across cERDs and cLRDs.

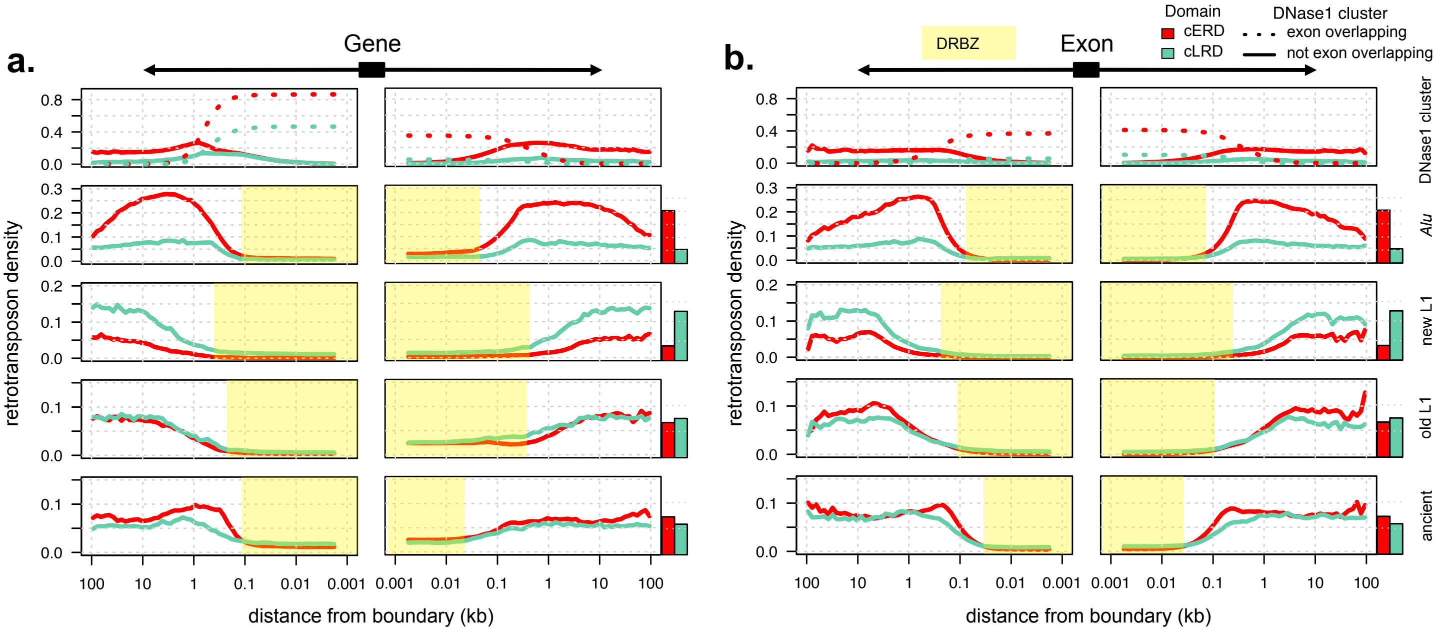

Figure 4. Retrotransposon accumulation within intergenic and intronic regions correlates with the distribution of DNase1 clusters. Density of DNase1 clusters and retrotransposons at each position upstream and downstream of genes and exons in **a**, intergenic and **b**, intronic regions. For DNase1 clusters, dotted lines represent exon overlapping clusters and solid lines represent not exon overlapping clusters. For retrotransposons, solid lines represent the uncorrected retrotransposon density at exon and gene boundaries. Bar plots show expected retrotransposon density across cERDs and cLRDs. Highlighted regions outline DRBZs, regions extending from the gene or exon boundary to the point where retrotransposon levels begin to increase.