PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse

Reuben M. Buckley, R. Daniel Kortschak, David L. Adelson*

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

* david.adelson@adelaide.edu.au

OPEN ACCESS

Citation: Buckley RM, Kortschak RD, Adelson DL (2018) Divergent genome evolution caused by regional variation in DNA gain and loss between human and mouse. PLoS Comput Biol 14(4): e1006091. https://doi.org/10.1371/journal.pcbi.1006091

Editor: Jian Ma, Carnegie Mellon University, UNITED STATES

Received: September 6, 2017

Accepted: March 15, 2018

Published: April 20, 2018

Data Availability Statement: Data used for analyses were obtained from the following URLs. Data for whole genome alignments were downloaded from: rsync://hgdownload.cse.ucsc.edu/goldenPath/. For a full listing of ingroup and outgroup genome alignments see S1 Text. Data for Conserved syntenic regions was downloaded from: http://bioinfo.konkuk.ac.kr/synteny_portal/. Data for DNaseI HS peaks human was downloaded from: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseMasterSites/. Data for DNaseI

Abstract

The forces driving the accumulation and removal of non-coding DNA and ultimately the evolution of genome size in complex organisms are intimately linked to genome structure and organisation. Our analysis provides a novel method for capturing the regional variation of lineage-specific DNA gain and loss events in their respective genomic contexts. To further understand this connection we used comparative genomics to identify genome-wide individual DNA gain and loss events in the human and mouse genomes. Focusing on the distribution of DNA gains and losses, relationships to important structural features and potential impact on biological processes, we found that in autosomes, DNA gains and losses both followed separate lineage-specific accumulation patterns. However, in both species chromosome X was particularly enriched for DNA gain, consistent with its high L1 retrotransposon content required for X inactivation. We found that DNA loss was associated with gene-rich open chromatin regions and DNA gain events with gene-poor closed chromatin regions. Additionally, we found that DNA loss events tended to be smaller than DNA gain events suggesting that they were able to accumulate in gene-rich open chromatin regions due to their reduced capacity to interrupt gene regulatory architecture. GO term enrichment showed that mouse loss hotspots were strongly enriched for terms related to developmental processes. However, these genes were also located in regions with a high density of conserved elements, suggesting that despite high levels of DNA loss, gene regulatory architecture remained conserved. This is consistent with a model in which DNA gain and loss results in turnover or "churning" in regulatory element dense regions of open chromatin, where interruption of regulatory elements is selected against.

Author summary

Approximately 2% of a mammalian genome is protein-coding DNA, the remainder is non-coding DNA. In mammals, this non-coding DNA fraction has undergone large amounts of turnover since placental mammals diverged from a common ancestor. For example, human and mouse, two species who diverged approximately 100 million years ago, share only approximately 40% of their DNA sequence. Given that genome size has

HS peaks mouse was downloaded from: https://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDgf. Data for RepeatMasker database was downloaded from: http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html. Data for Lamina associated domains human was downloaded from: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/laminB1Lads.txt.gz. Data for Lamina associated domains mouse was downloaded from: https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE17051&format=file&file=GSE17051%5FcLAD%5Fregions%2Ebed%2Egz. Data for Recombination rate human was downloaded from: ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/. Data for Recombination rate mouse was downloaded from: http://www.genetics.org/highwire/filestream/412790/field_highwire_adjunct_files/11/TableS1.csv. Data for Recombination hotspot human was downloaded from: ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2006-10_rel21_phaseI+II/hotspots/. Data for Recombination hotspot mouse was downloaded from: http://www.genetics.org/highwire/filestream/412790/field_highwire_adjunct_files/12/TableS2.csv. Data for Gene Regulatory blocks was downloaded from: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-017-00524-5/MediaObjects/41467_2017_524_MOESM2_ESM.txt. Data for Human and mouse gene expression comparison was downloaded from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4654840/bin/12862_2015_534_MOESM1_ESM.xls. Data for CpG islands human and mouse was downloaded from: http://hgdownload.soe.ucsc.edu/goldenPath/.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared no competing interests exist.

remained relatively constant since their divergence, this low level of ancestral DNA suggests there has been large amounts of DNA gain and loss in both lineages. To understand the cause and evolutionary impact of DNA gain and loss in mammalian genomes, we developed novel techniques that mapped individual DNA gain and loss events across distantly related species. By tallying the amount of DNA gained and lost across genomic regions we were able to measure its association with various genomic features. Our results showed that DNA loss in human and mouse mainly occurs in gene-rich open chromatin regions. In contrast DNA gain was mainly driven by transposition. In each lineage the proportion of total gain could be assigned to distinct transposon types. This meant that based on the differential activity of specific transposon types region-specific gain was following lineage-specific accumulation patterns, ultimately leading to divergent genome evolution. In addition, we measured how genes in DNA gain and loss hotspots associated with particular biological processes. Perhaps most strikingly, we found that mouse DNA loss hotspots overlapped highly conserved regions containing genes involved in development. This suggests that while the genomic environment in these regions is prone to DNA loss events, those that interrupt regulatory elements are strongly selected against.

Introduction

Evolution as a result of natural selection has led to many streamlined forms which follow directly from their function. However, in the case of genome evolution of complex organisms this connection is not quite so direct. One example is the evolution of genome size. In vertebrates, gene content has remained relatively constant, while the fraction of non-coding DNA varies drastically [1–3]. This observation is at the heart of the C-value enigma and raises many questions regarding the molecular drivers and evolutionary impacts of genome size variation. The major factor contributing to the total non-coding DNA genomic fraction is transposon load, due to mobile DNA elements that have actively replicated throughout evolution [2, 3]. In humans, since their divergence from the common placental ancestor, transposon activity has caused approximately 815 Mb of DNA gain, almost one third of their extant genome [4, 5]. However, this is not the only factor driving genome size evolution. DNA loss via deletion also plays a role, with approximately 650 Mb of the human genome being lost over the same time period [4]. Across mammals and birds these two forces operate in opposition to each other leading to the accordion model of genome evolution, where departures from this DNA gain and loss equilibrium cause genomes to either grow or shrink [4]. Importantly, our understanding of DNA gain and loss stems from genome-wide estimates rather than detection of individual events. Therefore, the role of genome structure on widespread DNA gain and loss and its subsequent impact on lineage-specific species evolution remains unknown.

The 'accordion' model of genome size evolution raises important questions regarding the roles of natural selection and genetic drift. Genome size, like any other heritable trait, is shaped by a combination of both of these factors [6]. However, the contribution of each mechanism in diverse taxa remains an open question in biology, with evidence to support the impact of each [7]. For genome evolution driven by selection there are observations of various phenotypic correlates consistent across both mammals and birds. One example is the evolution of powered flight in bats and birds which requires a high metabolic rate. Because metabolism is more efficient in smaller cells, it has been suggested that in flying species there is particularly strong selection pressure against genome growth [4, 8, 9]. Alternatively, observed genome size variation can result from neutral evolutionary processes. Many higher order vertebrates have low

effective population sizes resulting from reduced efficiency of selection [10], suggesting that neutral or mildly deleterious mutations such as some transposon insertions can easily reach fixation. Moreover, as transposons quickly accumulate the probability of deletions through non-allelic homologous recombination also increases, counteracting their initial impact on genome growth [11, 12]. Within this context, the accordion model is an emergent property based on transposon accumulation dynamics. Importantly, the signatures of selection for an optimal genome size are not always consistent; the Chinese tree shrew has a high metabolic rate but a relatively large genome of 2.86 GB [13]. This suggests that the role selection plays in driving genome size evolution is likely taxon-specific. Further, neither mechanism takes into account the underlying genome structure.

The genomic DNA of complex organisms is wrapped around nucleosomes and packaged into various conformations that regulate the access of different gene regulatory factors to their target sites. This hierarchical genome structure means that the impact and likelihood of particular mutations is highly context-specific, resulting in regional variation in both the susceptibility and tolerance to mutations. Here, susceptibility is the likelihood of a mutation occurring and tolerance is the degree to which the mutation does not adversely impact fitness. The observed accumulation patterns of DNA gain and loss events arise from the interaction of region-specific susceptibility and tolerance. For example, small ($\leq$ 30 bp) insertion or deletion (indel) events in the human genome are correlated with recombination rate and are enriched for topoisomerase cleavage sites [14, 15]. This suggests that the biological role of certain regions may cause them to be particularly susceptible to indel mutations. In the case of larger events such as transposon insertions, the prevailing model suggests that long interspersed elements (LINEs) accumulate in gene-poor regions where they are most tolerated [16]. The evolution of genome size via DNA gain and loss is not only shaped by higher order factors such as cell size and metabolic rate, but is intimately linked to the underlying genome structure.

To better characterise the molecular drivers and evolutionary impacts of DNA gain and loss, we calculated lineage-specific gain and loss rates across the human and mouse genomes. Human and mouse were chosen specifically for three reasons. Firstly, both species have well characterised genomes with highly accurate and well annotated assemblies [5, 17] and have both been used frequently in comparative genomic analyses resulting in many easily accessible pairwise alignment datasets available on the UCSC genome browser [18]. This makes it possible to compare them to a wide variety of outgroup species and detect genomic features that associate with DNA gain and loss. Secondly, the mouse genome is significantly smaller than the human genome, making it possible to detect a large number of lineage-specific deletion events [17, 19]. Finally, human and mouse genomes contain similar lineage-specific transposon families [17]. This means that both species share similar mechanisms for DNA gain, making it easier to compare differences between associations with other types of genomic features.

For our analysis, we detected DNA gain and loss events using two distinct, yet complementary, methods from which we characterised DNA gain and loss hotspots. From this we compared the genomic distributions of our hotspots to the genomic distribution of various features associated with genome evolution and genes that participate in particular biological processes. Our results revealed that DNA gains and losses occur in different regions across autosomes, while DNA gains from both species are particularly enriched on the X chromosome where they overlap. DNA gain events generally associate with L1 accumulation and DNA loss occurs in regions associated with biological activity such as transcription and regulation. Although DNA gain and loss in human occurred mostly in different regions, they both tended to impact on the same biological processes, while in mouse DNA loss was enriched for developmental genes and DNA gain did not associate with any particular biological process.

Materials and methods

Net data structure and feature extraction

For feature extraction, nets were obtained from the UCSC genome browser [20, 21]. Nets are a common format for representing pairwise genome alignments. Each net contains chained blocks of aligning sequence shared between a reference and a query genome. In order for alignment blocks to be chained together their ordering must be consistent between both genomes. Often gaps between chained blocks can contain smaller chains. It is this hierarchical structuring of the highest scoring chains at the top level with lower scoring chains filling in alignment gaps that makes nets. Importantly, in the reference genome nets provide only a single layer of coverage. However, two separate nets may occasionally overlap in the query; this is usually caused by segmental duplication in the reference. These conflicts were resolved by discarding all reference nets that did not overlap nets generated from a query reference alignment. Following this filtering process, only reciprocal best hit (RBH) nets remained. In our analysis we referred to alignment blocks within a chain as 'chain-blocks' and the spaces between chain-blocks also within a chain as 'chain-gaps'. The start and end coordinates in both the reference and query genome were recorded for each chain-block and chain-gap. The programs get_gaps_net.go and get_fills_net.go were used to extract all chain-gaps and chain-blocks respectively. Regions of chain-gaps that were overlapped by chain-blocks in lower ranked chains were discarded. Additionally, regions that were discarded as non-RBHs or fell outside of nets were plotted against synteny blocks to determine the loci hidden from our analysis in both species. Synteny data was obtained from the synteny portal (http://bioinfo.konkuk.ac.kr/synteny_portal/) [22].

Identifying ancestral elements

Chain-blocks were extracted from all genomes identified as outgroups to human and mouse. They were combined into a single file and merged using the bedtools genomecov function with the '-bg' option. This process returned a set of potential 'ancestral elements' along with their corresponding coverage depth. To identify false-positives and estimate the type 1 error rate, we used the genomic positions of a set of known lineage-specific repeat families in human and mouse, since lineage-specific repeat insertions should not overlap ancestral elements. The percentage overlap of our lineage-specific repeats set with ancestral elements was measured at each minimum coverage level. A similar approach was used to estimate the type 2 error rate; the type 2 error rate was estimated as the percentage of chain-blocks that did not overlap ancestral elements. To minimise our type 1 errors we selected a minimum coverage depth threshold independently for both hg19 and mm10, where nucleotide positions with coverage depth below the threshold were not considered as ancestral elements. The basis for this approach was that nucleotide positions in our reference genomes that aligned to a large number of outgroup species were highly likely to share ancestry with those species. In contrast, nucleotide positions in our reference genomes that aligned to very few outgroup species were likely errors caused by spurious alignments between complex regions that are difficult to map. Importantly, reductions in our type 1 error caused an increase in our type 2 error. Therefore, we chose the highest possible minimum coverage threshold, where the gain in the cumulative proportion of type 1 errors from lower threshold values was greater than the gain in proportional increase of type 2 errors.

Identifying recent transposon insertions

For both hg19 and mm10, genomic coordinates for transposons were obtained from the Repeat Masker database [23]. Based on their overlap with chain-blocks or ancestral elements, individual transposons were classified as either recent or ancestral. In addition to this, the

percent divergence from consensus family sequence and the proportion of total sequences of transposon family members that overlapped ancestral elements or chain-blocks were calculated. These data were then used in linear discriminant analysis to build a transposon family classifier. Our classifier was trained using the original individual transposon classifications. After training, entire families were classified as either recent or ancient using the family-wise means of the feature values. Finally, transposons from families classified as recent but overlapping gaps between reference and query were classed as lineage-specific insertions.

Gap annotation and placement

Chain-gaps extracted from nets were annotated as either DNA gain or DNA loss based on two distinct yet complementary annotation methods; the recent transposon-based method and the ancestral element-based method. The ancestral element-based method infers the ancestral state of a gap. For example, an mm10 gap overlapping an ancestral element would be annotated as an mm10 loss, whereas the same gap not overlapping an ancestral element would be annotated as an hg19 gain. The recent transposon-based method instead identifies DNA gains. In this case an mm10 gap overlapping a recent transposon would be annotated as an hg19 gain, while an mm10 gap not overlapping a recent transposon would be annotated as an mm10 loss.

After all chain-gaps between a reference and query were annotated in both genomes, the remaining non-aligning sequences were 'placed' in the genomes they were absent from. This process is referred to as 'gap placement' and is performed on the non-aligning sequence of chain-gaps that remain in the reference genome after a reference query alignment. These non-aligning reference sequences are absent from the query and are either the result of DNA gain in the reference or DNA loss in the query. Using the coordinate mappings of the 5′ and 3′ adjacent chain-blocks of each chain-gap, the non-aligning reference sequence of a chain-gap is inserted into the query genome at the corresponding position, where placed gaps are oriented relative to the genome they are placed in. Importantly, gap placement begins by placing chain-gaps at the bottom chain level of nets and ends by placing chain-gaps at the top chain level. This process ensures that non-aligning sequence in overlapping chain-gap annotations caused by hierarchical structure of nets are only placed once. Once the corresponding position of a gap has been identified, the downstream query coordinates are incremented by the size of the annotated chain-gap being placed. This creates a synthetic genome consisting of DNA gains and losses that occurred across both the reference and query lineages. The total length of our synthetic genomes is equal to the total length of the query genome and the total length of annotated chain-gaps from the reference. Finally, the synthetic genomes were segmented at a window size of 200kb into distinct genomic bins where the total size of each gap annotation was tallied. Genomic bins with less than 150 kb that did not belong to assembly gaps or non-RBH regions were discarded. Importantly, our decision to use a synthetic genome meant that placed chain-gaps larger than our window size would spread across window boundaries, ensuring that genomic bins would contain no more than 200 kb of sequence.

Hotspot identification

Hotspots for reference gain, reference loss, query gain and query loss in both hg19 and mm10 were identified using the Getis-Ord local statistic found in the R package 'spdep' [24, 25]. The Getis-Ord local statistic for genomic bin i is calculated as:

$$G_i^* = \frac{\sum w_{i,j}x_j - \bar{X}\sum w_{i,j}}{S\sqrt{\frac{n\sum w_{i,j}^2 - (\sum w_{i,j})^2}{n-1}}}, \quad (1)$$

where x_j is the number of bp belonging to a particular gap annotation within bin j, $w_{i,j}$ is the spatial weight between bin i and j, n is the number of bins for a particular genome, $\bar{X} = \frac{\sum x_j}{n}$ and $S = \sqrt{\frac{\sum x_j^2}{n} - \bar{X}^2}$ [26]. For the neighbourhood weight matrix W, $w_{i,j}$ was given a spatial weight of 1 if bin i and bin j were considered neighbours. For bin i and j to be considered neighbours bin j had to be within 600 kb of bin i. After calculating G_i^* for each bin and each gap annotation in both genomes, all G_i^* values were converted to P-values and adjusted for multiple testing using the false discovery rate (FDR). Bins were only considered hotspots if their G_i^* was > 0 and had a FDR < 0.05. Additionally, bins were considered coldspots if their G_i^* was < 0 and had a FDR < 0.05.

Obtaining genomic features

A set of genomic features was obtained from a range of sources to identify factors potentially driving DNA gain and loss. GC content was calculated as the proportion of chain-blocks per bin using the hg19 and mm10 Biostrings-based genome R packages [27–29]. CpG islands for both hg19 and mm10 were obtained from the UCSC genome browser [18]. DNaseI hypersensitivity (DNaseI HS) peaks for hg19 were obtained from UCSC as part of the DNaseI master track (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseMasterSites/). The master track was generated by combining DNaseI HS sites from across 125 cell lines produced by the University of Washington and Duke University ENCODE groups [30]. The Individual cell line data can be located using the accessions GSE29692 and GSE32970. DNaseI HS peaks for mm10 were obtained from UCSC as individual samples mapped to mm9 (https://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDgf). Individual peaks from each sample were merged into a single file, creating a single set of DNase1 HS peaks. The merged mm9 peaks were then converted to the mm10 assembly using the UCSC liftover tool [31]. Mouse DNaseI HS peaks were generated using DNaseI digital genomic foot-printing performed by the University of Washington ENCODE group [30]. This dataset can also be obtained using the accession GSE40869. Importantly, as part of the ENCODE pipeline, multi-mapping reads were discarded. To remove this bias from the analysis, genome-wide mappability tracks were used so that only uniquely mappable regions of the genome were considered. For hg19 the 36-mer mappability track was generated using the gem-mappability program with a mismatch score of 2, which was obtained form the UCSC genome browser [32]. For mm10 a 36-mer mappability track was instead generated locally using the same program and same parameters. Recombination rates for human were identified as part of the HapMap project (ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/) [33]. However, recombination hotspots were only available for earlier phases of the HapMap project (ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2006-10_rel21_phaseI+II/hotspots/). The hotspots were initially mapped to hg17 and then converted to hg19 coordinates using the UCSC liftover tool. Recombination hotspots were identified using the methods outlined in Winckler *et al* [34] and McVean *et al* [35]. Recombination rates and hotspots in mouse were calculated in mm9 based on two separate datasets [36–38]. They were converted to mm10 using the UCSC liftover tool. Importantly, recombination data was only available for mouse autosomes. During enrichment tests this was taken into account by removing the sex chromosomes from the sample space. Exons and introns for both hg19 and mm10 were extracted from UCSC genome annotations available from TXDB R packages [39–41]. Retrotransposon coordinates for hg19 and mm10 were obtained from the Repeat Masker database (http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html) [23]. The Repeat

Masker version used for hg19 and mm10 was open-4.0.5 with repeat library 20140131. Retrotransposons were sorted into the following categories: ancient elements, ancestral L1s, lineage-specific L1s and lineage-specific SINEs using prefixes for families of known lineage-specific and ancestral activity [42]. Ancient elements were identified by the class names 'SINE/MIR' and 'LINE/L2'. Ancestral L1s were identified using the family name prefixes 'L1ME', 'L1MD', 'L1MC', 'L1MB' and 'L1MA'. Human lineage-specific L1s were identified using the family name prefixes 'L1PB', 'L1PA' and 'L1HS'. Mouse lineage-specific L1s were identified using the family name prefixes 'Lx', 'L1Md', 'L1_Mus', 'L1_Mur' and 'L1_Mm'. Human lineage-specific SINEs were identified using the family name prefix 'Alu'. Mouse lineage-specific SINEs were identified using the family name prefixes 'PB', 'B1', 'B2', 'B3' and 'B4'. Lamina associated domains (LADs) for hg19 were obtained from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/laminB1Lads.txt.gz) [43]. LADs for mouse were constitutive across several samples and were obtained using the accession GSE17051, they were converted from mm9 assembly to mm10 assembly using the UCSC liftover tool [44]. For each feature, except recombination rate, the per 200 kb coverage level for each bin was calculated. For genomic features that are usually considered as binary features such as CpG islands and exons, we measured the total number of bases per bin that belong to each feature type. This in turn makes them comparable to continuous genomic features such as GC content in downstream analysis. For recombination rate the mean rate per bin was used.

Genomic feature enrichment

Feature enrichment was detected on the basis of a permutation test. For each feature and hotspot in both hg19 and mm10, a background distribution was generated by calculating the difference in means between a set of resampled hotspot and non-hotspot bins 10,000 times, resampling was performed without replacement. The background distribution was then used to convert the differences in means between observed hotspot and non-hotspot bins into a Z-score to allow standardisation between features and gap annotations and provide the direction of the association. Z-scores are only shown if they are outside the range of -3 to 3.

GO term enrichment analysis

Gene ontology (GO) term enrichment was calculated using the topGO package in R [45]. Genes within each hotspot region were independently tested against the genomic background. For enrichment, the Fisher test was used in combination with four separate algorithms: the classic algorithm treats each term independently whereas the elim, weight and parent-child algorithms factor in the GO inheritance structure [46–48]; the elim algorithm removes all genes annotated to a significantly enriched GO term from all of the terms ancestors; the weight algorithm behaves similarly, instead of removing genes from the ancestors of enriched GO terms, it creates a more subtle effect by reducing the weight of genes annotated to the ancestors of enriched GO terms [46]; for the parent-child algorithm, the enrichment score for a particular term takes into account the probability a random set of genes of the same size contains the same exact parents [47]. Because the non-classic algorithms adjust the enrichment probabilities they obviate the need to account for multiple testing [45]. For all non-classic algorithm a significance threshold of 0.05 was applied. Whenever a significance threshold was used with the classic algorithm, P-values were adjusted for multiple testing by calculating the FDR and a significance threshold of 0.05 was used.

Dating DNA gain and loss events

For hg19 and mm10, DNA gain and loss events were dated according to whether or not they were supported by an alignment gap from an ingroup species. In this case the ingroup species belonged to two main groups, human-related ingroup species and mouse-related ingroup species. The human-related ingroup species in order of relatedness to human were chimpanzee, baboon, tarsier and mouse lemur. The mouse-related ingroup species in order of relatedness to mouse were rat, kangaroo rat, and pika. These ingroup species were chosen as they each represented distinct lineages and divergence times between either human or mouse. The divergence times for each ingroup species were calculated using the "estimated divergence time" found on TimeTree [49]. Moreover, these species were also chosen as their net chain alignments contained only reciprocal best-hit alignments. For each ingroup species their whole genome alignments between both hg19 and mm10 were obtained from the UCSC genome browser. Alignment gaps between each ingroup species and human, and each ingroup species and mouse were extracted using the program get_gaps_net.go. In the case where human was used as a reference and mouse as a query, DNA gain events were dated by overlapping them with alignment gaps between human and human-related ingroup species. Comparisons between human and human-related ingroup species were made in order of most closely to least closely related to human, early DNA gain events were dated first and later DNA gain events were dated last. For example, hg19 DNA gain events overlapping gaps in the chimpanzee alignment were dated as occurring after human and chimpanzee divergence. From the remaining DNA gain events, those that overlapped gaps in the human and baboon alignment were then dated as occurring after human and baboon divergence and prior to human and chimpanzee divergence. This process of dating DNA gain events using human and ingroup species alignments occurred until all that remained were unsupported DNA gain events. These events were dated as occurring after human and mouse divergence and prior to human and mouse lemur divergence. Importantly, dating DNA loss events followed a slightly different procedure. This was because DNA from human DNA loss events is absent from hg19 and instead located in mm10. This meant that human DNA loss events were dated using alignments gaps between the human-related ingroup species and mm10. In contrast to dating DNA gain events, comparisons between mm10 and human-related ingroup species went in order of least related to human to most related to human. This was because DNA loss events that occurred early during human lineage specification are shared across all human-related ingroup species, while DNA loss events that occurred recently are only shared with recently diverged species.

Software and data analysis

All statistical analyses were performed using R including the packages GenomicRanges, RMySQL, dplyr and Bioconductor [41, 50–53]. Code used to perform analyses can be found at: https://github.com/AdelaideBioinfo/regionalGenomeTurnover.

Results

Detecting DNA gain and loss events

Across genomes and throughout evolution DNA is frequently gained and lost by the processes of insertion and deletion. To identify individual events and quantify DNA gain and loss at a regional level in hg19 and mm10, we obtained pairwise alignment data between both genomes in the form of nets from the UCSC genome browser (Methods) [18, 21]. By taking advantage of the data's hierarchical structure we were able to estimate DNA gain and loss in regions that

have undergone rearrangements. We processed our data in three distinct steps; 1) extract features (Fig 1a), 2) annotate gaps (Fig 1b and 1c) and 3) place gaps (Fig 1d).

For step 1, chain-gaps and chain-blocks were extracted from nets considering only chain-gaps of at least 10 bp in size (Fig 1a) (Methods). Our approach allowed us to keep track of each feature's position in both the reference and query genome. This is especially important since it is not possible to identify deletions when the corresponding coordinates between species are lost. After extracting features we found that approximately 111 Mb of hg19 and 174 Mb of mm10 were not contained within nets (Table 1). Alignment gaps that did not belong to any nets in human and mouse tended to overlap regions between two conserved synteny blocks (S1 and S2 Figs). With the remaining features extracted from hg19 and mm10, we used the corresponding coordinates between reference and query to identify features that were reciprocal best hits (RBHs). This removed features in the reference genome that mapped to similar locations in the query, which are likely the result of segmental duplication. After filtering out non-net and non-RBH regions, 1014.3 Mb of chain-blocks and 1465.8 Mb of chain-gaps remained in hg19, and 994.4 Mb of chain-blocks and 1191.5 Mb of chain-gaps remained in mm10 (Table 1). Since our processed nets for each genome are supposed to only contain RBH features, it is expected that the coverage of chain-blocks should be equal between hg19 and mm10. To determine the source of this discrepancy, we analysed the number of chain-gaps below our minimum size cut off and found that when these were taken into consideration the difference in chain-block size was reduced to approximately 1 Mb.

Next, for step 2 we annotated chain-gaps as either lineage-specific DNA gain or DNA loss. To annotate gaps we used two complementary methods, an ancestral elements-based method and a recent transposon-based method. The ancestral element-based method uses outgroup species to annotate gaps by inferring their ancestral state (Fig 1b). For example, if a particular sequence between a reference and outgroup is conserved but presents as a gap in the query it is likely that this sequence was lost from the query. Alternatively, if this particular sequence in the reference presents as a gap in both the query and the outgroup it is likely that this sequence was instead gained in the reference. An important consideration for identifying ancestral elements is the type 1 (false positive) and type 2 (false negative) error rates, where type 1 errors are lineage-specific regions annotated as ancestral elements and type 2 errors are ancestral regions annotated as lineage-specific. To reduce our type 2 error rate we obtained the genomes of a large range of human and mouse outgroup species from the UCSC genome browser (S2 Table). Across all of our outgroup species we extracted all the chain-blocks and merged overlapping intervals to create our ancestral elements. This strategy increased the chance of finding ancestral DNA in our reference that may have been lost in one or more of our outgroup species. For both hg19 and mm10 we found that total genome coverage of ancestral elements reached asymptotic levels at approximately 18 outgroup species (S3 Fig). However, this strategy also came with the trade-off of increasing our type 1 error rate. To control error rates we measured how type 1 and type 2 errors responded to changes in coverage depth of outgroup chain-blocks at each position in hg19 and mm10 (S4 Fig). Based on these results we annotated human ancestral elements at an outgroup coverage depth ≥ 6 and mouse ancestral elements at an outgroup coverage depth ≥ 4 (S4 Fig). This strategy removed $> 85\%$ ancestral elements overlapping known lineage-specific repeats in mouse and $> 95\%$ of ancestral elements overlapping known lineage-specific repeats in human. For remaining chain-blocks, we found that 94.2% in human and 85.2% in mouse were supported by our annotated ancestral elements (Table 1). Our very low error rate in human indicates that we were able to accurately determine the amount of mm10 DNA loss and hg19 DNA gain. However, our error rates in mm10 suggest that ancestral regions alone are insufficient to accurately estimate hg19 DNA loss and mm10 DNA gain.

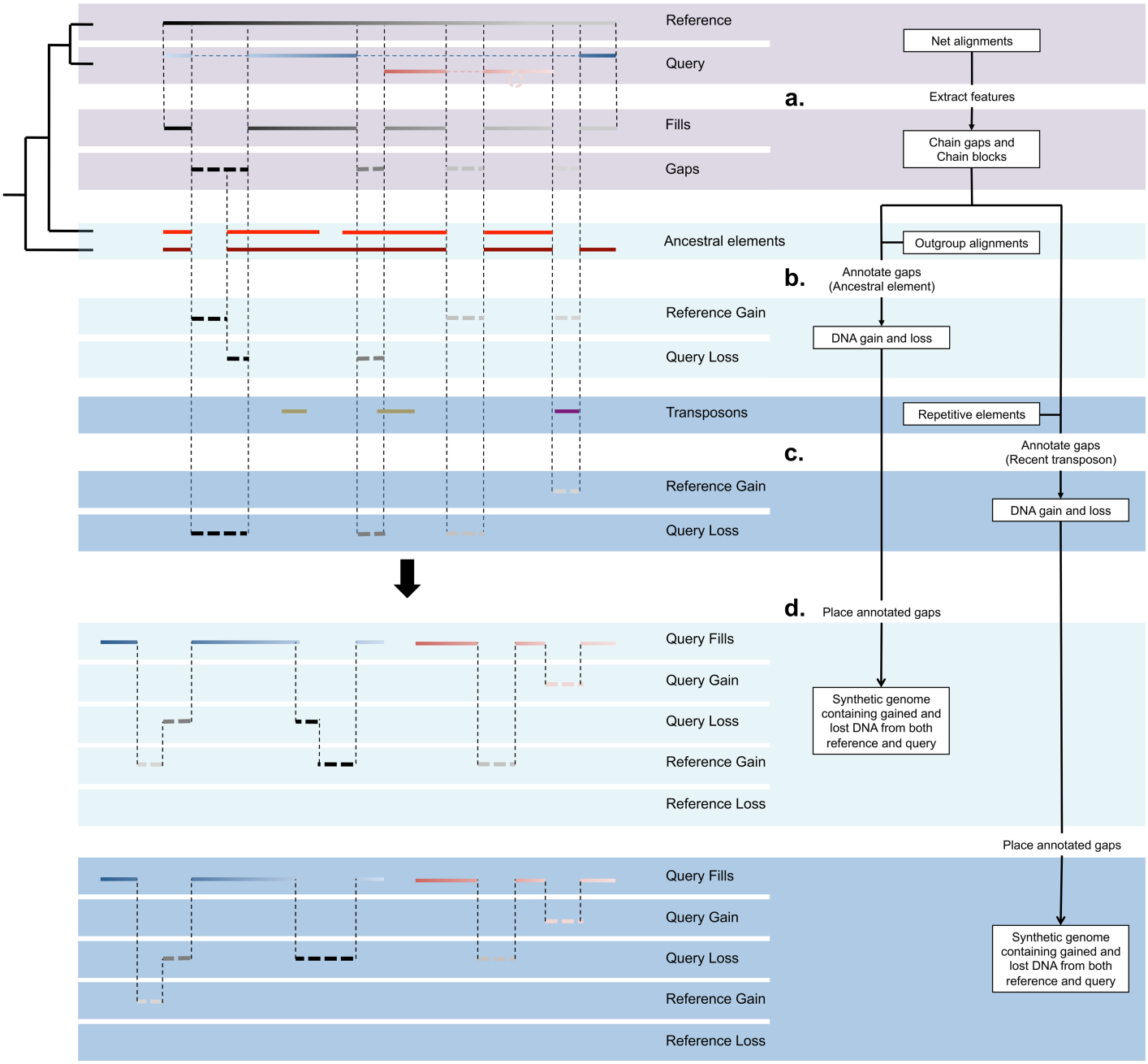


Fig 1. Detecting DNA gain and loss events between two species. Chain-gaps and chain-blocks are extracted from nets between reference and query (a). The resulting chain-gaps are essentially sequences from the reference genome that do not align to anything in the query genome. Chain-blocks are extracted from nets between reference and outgroup species as ancestral elements. Ancestral elements are then used to annotate chain-gaps as either gain or loss (b). Chain-gaps are annotated as query loss if they overlap ancestral elements or as reference gain if they do not. This is the ancestral element method for annotating gaps. The recent transposon method instead uses transposons classified as recent or ancestral to annotate gaps (c). Transposons are extracted from Repeat Masker files containing various classes of repetitive elements. Chain-gaps are annotated as reference gain if they overlap recent transposons or as query loss if they do not. After gaps are annotated they are placed within each genomic background creating a synthetic genome (d). Annotated chain-gaps are placed according to the edge coordinates of their adjacent chain-blocks within the same chain. Shown in the final two panels are chain-gaps extracted from the reference placed within the query genome. The different colours of the query chain-blocks show that gap annotations in the reference are placed on different chromosomes in the query. Differences in annotations are the results of conflicting information either resulting from incorrect identification of ancestral elements or recent transposons. Shading is used throughout the figure to help differentiate the ancestral element method from the recent transposon method.

https://doi.org/10.1371/journal.pcbi.1006091.g001

Table 1. Processing of net files.

Genomic regions (Mb)	hg19	mm10
Sequenced genome (Mb)	2897.0	2653.0
Gaps outside of nets (Mb)	111.1	174.0
Non-RBH chains (Mb)	306.1	293
Ancestral elements (Mb)	1726.0	1021.0
Remaining chain-blocks (Mb)	1014.3	994.4
Remaining chain-blocks ∩ ancestral elements (%)	94.2	85.2
Remaining chain-gaps (Mb)	1465.8	1191.5

https://doi.org/10.1371/journal.pcbi.1006091.t001

To complement and overcome potential shortcomings of the ancestral element-based method of estimating DNA gain and loss, we adopted a recent transposon-based method. We identified transposon families with lineage-specific activity and used them to annotate gaps as lineage-specific DNA gain or loss (Fig 1c). For example, recent transposon sequences in hg19 that overlap gaps in mm10 are annotated as hg19 gains, where ancestral transposon sequences in hg19 that overlap gaps in mm10 are annotated as mm10 losses. This approach has been used previously to identify DNA loss in the mouse and human lineages [17, 54].

In order to annotate gaps using the recent transposon method, we first had to identify transposon insertions that occurred after mouse and human diverged from their common ancestor. Because transposon families have undergone distinct bursts of activity at particular points in time, we decided to classify transposon families as either 'recent transposons' or 'ancestral transposons', and use members of those respective classifications to annotate our chain-gaps. The main challenge in this approach is identifying lineage-specific activity of transposons. Generally, transposon families are considered to be ancestral transposon families when they are shared between two species. However, there is a possibility some ancestral transposon families may have been active during the period of human and mouse divergence and continued replicating in each lineage independently. This means families that would have been otherwise classified as ancestral transposons may have actually undergone varying amounts of lineage-specific transposition.

To overcome the problem of misclassifying the activity of otherwise ancestral transposon families, we used linear discriminant analysis to build a transposon family classifier for both human and mouse. We initially obtained transposon coordinates from the Repeat Masker database and classified individual transposons as 'ancestral transposons' if they overlapped ancestral elements or chain-blocks and as 'recent transposons' if they did not. Next, we trained our classifier using two separate variables. The first variable was each transposon's percent divergence from their family consensus sequence, often used as an indicator of transposon age [55, 56]. The second variable was the proportional overlap between each transposon family and ancestral elements or chain-blocks as measured by bp coverage. After training we used our classifier to group each family based on the family-wise means for the variables above (S5 Fig). We identified 656 recent human transposon families and 689 recent mouse transposon families. Our results suggest that at least 176 families were active during human and mouse divergence leading to a mixture of both ancestral and lineage-specific insertions (S1 Table). Moreover, the percent divergence of these families is consistent with transposon activity occurring after the evolution of ancestral transposons and prior to the evolution of lineage-specific transposons (S6 Fig). Surprisingly, we also identified some transposon families that were not shared between human and mouse, and yet were annotated as ancestral. However, these families were usually small and together they covered less than 1 Mb of their respective

genomes (S1 Table). In addition, our results for mm10 indicate potential drawbacks in using the ancestral element-based method for annotating gaps; percent divergence from consensus for some recent transposon families is similar to ancestral transposon families. While this is consistent with an elevated rate of substitution in the rodent lineage, it suggests that a large number of regions in mm10 that share ancestry with our outgroup species may have diverged beyond the alignment threshold (S5 Fig). Collectively, these results demonstrate the difficulty of identifying recent transposon insertions based on family name alone. For this reason we decided to annotate chain-gaps using our newly classified recent transposon families, which were classified using a combination of family-wide and transposon-specific factors in conjunction with comparative genomic approaches.

DNA gain and loss annotation accuracy

Using both the ancestral element and recent transposon based methods, we annotated a large number of chain-gaps with varying levels of consistency. In hg19, both methods were largely consistent in identifying human-specific DNA gains and mouse-specific DNA loss. However, in mm10 there was less agreement between the methods; while the majority of mouse lineage-specific DNA gains identified by both methods tended to overlap, the majority of human lineage-specific DNA loss did not (Table 2). This is most likely due to limitations for detecting ancestral elements in mm10. We found that only 85% of mm10 chain-blocks were supported by ancestral elements as opposed to 95% in hg19 (Table 1), suggesting that many ancestral elements were not identified using our outgroup species. This is a key weakness in our approach; if there is an underlying error for detecting human DNA loss in mm10, it means that we would also be overestimating the amount of DNA gain in mm10. However, by using two distinct yet complementary methods, we are able to identify potential sources of error and estimate their impact. One explanation for missing ancestral elements may be that DNA gain and loss events that occurred in either the mouse or human clade overlap DNA gain and loss events that occurred across a large number of our outgroup species. However, as stated above, nucleotide divergence rates may also play a role. Some regions in mm10 may have diverged so much that it is impossible to perform a pairwise alignment with our outgroup species. Despite the above mentioned inconsistencies between the methods in mm10, it is clear that the amount of DNA loss in human is much smaller than the amount of DNA loss in mouse and the amount of DNA gain for both. The difference in loss rates for human and mouse is mostly

Table 2. hg19 and mm10 gap annotation. Chain-gaps were annotated using both the ancestral element and recent transposon method. Each number represents gap annotations in Mb.

hg19 chain-gaps			
Recent transposon	**Ancestral element**		
	hg19 gain	mm10 loss	**Total**
hg19 gain	685.0	37.8	722.8
mm10 loss	168.0	575.0	743.0
Total	853.0	612.8	1465.8
mm10 chain-gaps			
Recent transposon	**Ancestral element**		
	mm10 gain	hg19 loss	**Total**
mm10 gain	720.6	11.5	732.1
hg19 loss	356.1	103.4	459.5
Total	1076.7	114.9	1191.6

https://doi.org/10.1371/journal.pcbi.1006091.t002

consistent with a high deletion rate in the mouse genome that has caused it to shrink in size since divergence with human [17, 19].

To further characterise the results from each method we compared the length distributions of their gap annotations. For DNA gain events in hg19 and mm10, the ancestral element method displayed a much higher frequency of small elements than the recent transposon method. This may be caused by spurious alignments between similarly structured recent transposons found in reference and outgroup species, effectively separating the annotation gain events into smaller pieces. Moreover, the recent transposon method identified much higher frequencies of DNA gain events that correspond to full length consensus sequences of known transposon families (Fig 2a and 2b). Conversely, the length distributions for DNA loss events identified by each method were much more similar, especially in mm10. In hg19 the frequency of events detected by the ancestral element method were much lower than those detected by the recent transposon method (Fig 2c and 2d). This is consistent with the low number of ancestral elements in the mouse genome. However, the high level of consistency for both methods in identifying hg19 DNA gain and mm10 DNA loss where there is good support for outgroup species is highly encouraging. It indicates that the recent transposon method is a reasonably effective method in identifying DNA gain and loss in species where it is difficult to detect ancestral elements. Consistent between both methods is size distribution difference between DNA gain and loss. DNA gain events are mostly over 100 bp in length while DNA loss events are mostly under 100 bp.

In both hg19 and mm10 we annotated a large number of gain and loss events using two distinct methods. However, to measure the total amount of DNA turnover at particular loci, gaps annotated in a query genome needed to be mapped to a reference genome. Hence, gap annotations were placed using the reference and query coordinates we extracted from our nets in step 1 (Methods) (Fig 1d). To account for the placement of gaps from one genome into another, we adjusted the genomic coordinates at the target loci, resulting in a synthetic genome for both species (Methods). Each synthetic genome contains both hg19 and mm10 annotated gaps in either an hg19 or mm10 genomic background. Finally, our resulting dataset consists of 4 synthetic genomes; mm10 with gap annotations based on the ancestral element method, mm10 with gap annotations based on the recent transposon method, hg19 with gap annotations based on the ancestral element method and hg19 with gap annotations based on the recent transposon method. Collectively, these results demonstrate that it is possible to identify locations for the majority of DNA gain and loss events since human and mouse divergence. Using our identified DNA gain and loss events it is possible to characterise genome-wide patterns of DNA gain and loss and to begin to determine how DNA turnover may impact on mammalian genome evolution.

Genome-wide characteristics of DNA gain and loss

Genome size evolution in mammals follows an accordion model, where DNA gain is counteracted by DNA loss to maintain a relatively constant genome size [4]. To characterise how DNA gain and loss interacts with genome structure, we used our synthetic genomes to analyse the genomic distribution of DNA gain and loss events in hg19 and mm10. We began by segmenting synthetic genomes into 200 kb non-overlapping bins and tallying the total bp coverage of each type of gap annotation. Several bin sizes were tested, however we found that at 200 kb the total sum of gap annotations per bin averaged approximately 150 kb and all bins were less than 200 kb (S10 Fig). This meant that 200 kb could provide good genomic resolution and no single type of gap annotation would span the entire width of a single bin. Bins with less than 150 kb of DNA not belonging to RBH nets were removed and our tallies were normalised to reflect DNA gain and loss amounts per 200 kb. Additionally, because gap annotations from

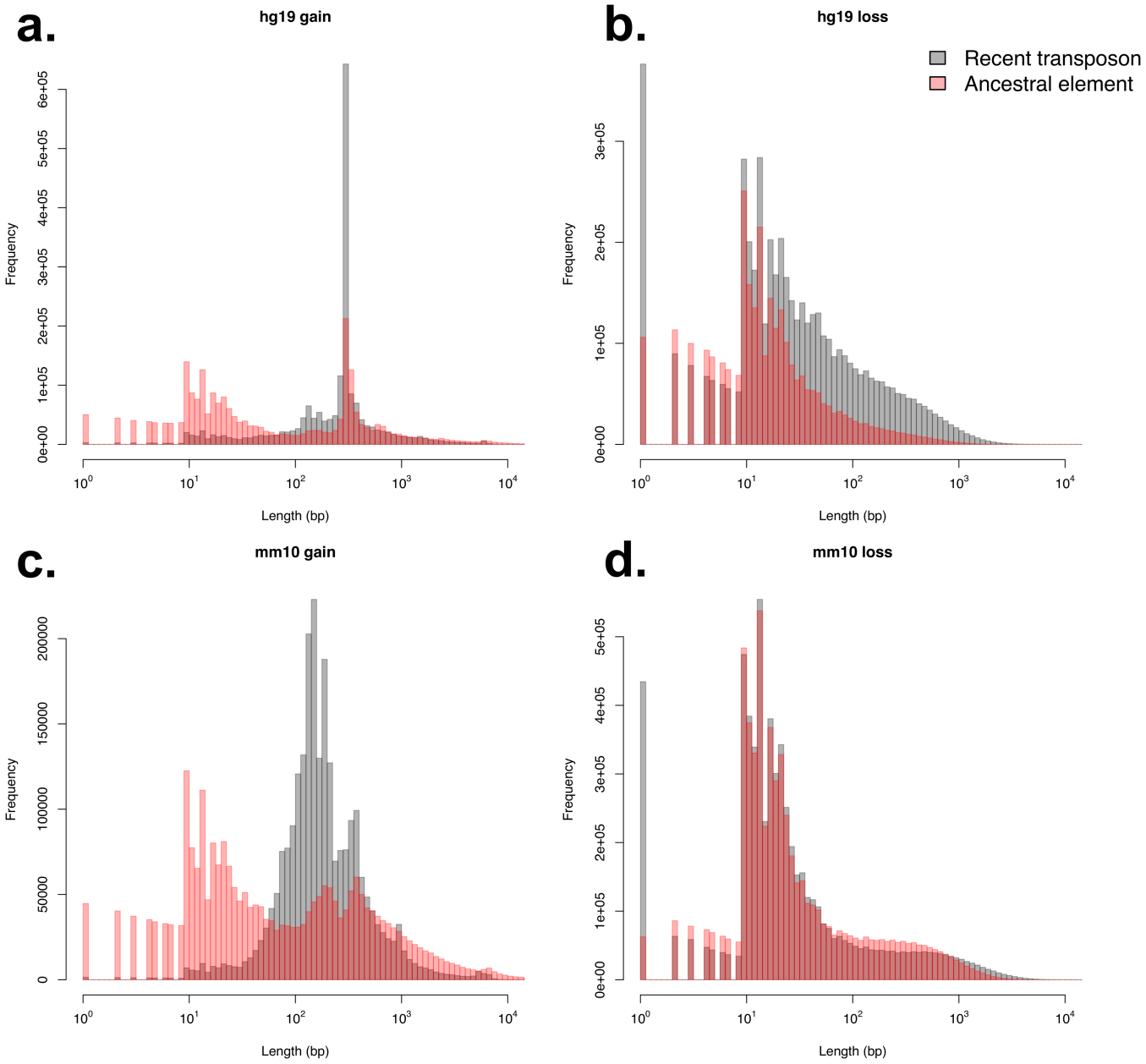


Fig 2. Length distributions of identified DNA gain and loss events. hg19 gain (a), hg19 loss (b), mm10 gain (c) and mm10 loss (d) events were identified using both the recent transposon and ancestral element method. Peaks for hg19 and mm10 gain, especially those detected by the recent transposon method, correspond to know lengths of transposon families.

https://doi.org/10.1371/journal.pcbi.1006091.g002

both species can be placed within a single genome, we are able to directly compare their genomic distributions.

Using our binned synthetic genomes we compared the variation and average amount of regional DNA gain and loss identified using each method. Our results showed that variation in regional DNA gain or loss was reasonably consistent across both methods (Fig 3). For DNA gain this was also quite large, in 200 kb genomic bins the amount of DNA gain in human and mouse spanned a range greater than 70 kb, indicating that some regions underwent much greater levels of DNA gain than others. While bin-wise variation in gain and loss rates was

consistent across methods, the average amount of DNA turnover was not. This makes it difficult to reliably calculate the regional amount of DNA turnover or genome growth. However, despite these inconsistencies, bin-wise levels of DNA gain and loss were highly correlated across all cases, with the exception of hg19 DNA loss (Fig 3a) (S7 and S8 Figs). Surprisingly, given that mm10 DNA gain is essentially the inverse of hg19 DNA loss, mm10 gain calculations are fairly consistent with respect to each method. This is because there has been a much higher level of mm10 DNA gain than hg19 DNA loss, causing calculations for the total amount of hg19 DNA loss to be much more sensitive to incorrect annotation (Table 2). Following this, we investigated regional DNA gain and loss dynamics by identifying DNA gain and loss genomic hotspots. Hotspots were identified by calculating G_i^* for each bin (Methods). For our hotspot identification, we used a neighbourhood size of 600 kb (3 neighbouring bins) both upstream and downstream of the bin in question. Before deciding to use 600 kb in our analysis we tested several other neighbour distances. Our results showed that at a neighbour distance of 3 bins, G_i^* scores show a relatively strong correlation with raw signal and also display a reasonably smooth signal (S11 Fig). More importantly, by plotting the locations of hotspots at different neighbour distances, we observed a strong tendency for hotspots to grow in size as neighbourhood distance increased (S12 Fig). We converted our G_i^* values to P-values and calculated the false discovery rate (FDR). Bins whose G_i^* was positive with FDR < 0.05 were considered hotspots. Hotspots were identified for each type of gap annotation found using both gap annotation methods in both synthetic genomes. We found that the size of the hotspot overlap between each gap annotation method for hg19 gain, mm10 gain and mm10 loss was larger than the sum of non-overlapping hotspots (Fig 3b). Using the hotspot intersect between gap annotation methods, we further characterised regional variation of DNA gain and loss across hg19 and mm10. For the remainder of the analysis the terms 'DNA gain hotspots' and 'DNA loss hotspots' refer to the hotspot intersect between each gap annotation method, except for hg19 DNA loss hotspots which instead refer to hg19 DNA loss hotspots identified through the recent transposon method. For mm10 DNA loss, mm10 DNA gain and hg19 DNA gain, the intersect was used as it provided a sample of genomic regions where regional DNA gain and loss dynamics were highly supported by both methods. For hg19 DNA loss we used hotspots that were identified using the recent transposon method because the ancestral based method was shown to largely underestimate the total amount of ancestral DNA.

Regional patterns of DNA gain and loss indicate lineage-specific divergence

The accordion model of genome evolution suggests DNA gain and loss is largely balanced across the entire genome. Whether the individual events are balanced at the local scale remains unknown. We analysed the genomic distribution of hg19 and mm10 gain and loss hotspots by focusing on the within-species overlap and the across species overlap. The within species overlap was designed to investigate whether DNA gain and loss is balanced on a regional level, indicating that despite large amounts of DNA turnover, local genome structures stay intact. The across species overlap was designed to investigate whether DNA gain and loss associated with lineage specific divergence in genome architecture. We found that almost 4% of human loss hotspots overlapped human gain hotspots and approximately 6% human gain hotspots overlapped human loss hotspots (Fig 4) (S13 Fig). These results showed that DNA gains and losses in human at a regional scale have occurred independently. Conversely, less than 1% of gain and loss hotspots in mouse overlapped each other, with a significant negative association. These results suggest that regional DNA gain and loss in both species is largely unbalanced. For the across species comparison, we found significant levels of overlap between DNA-loss hotspots and negative associations between all other hotspot types at varying levels of statistical

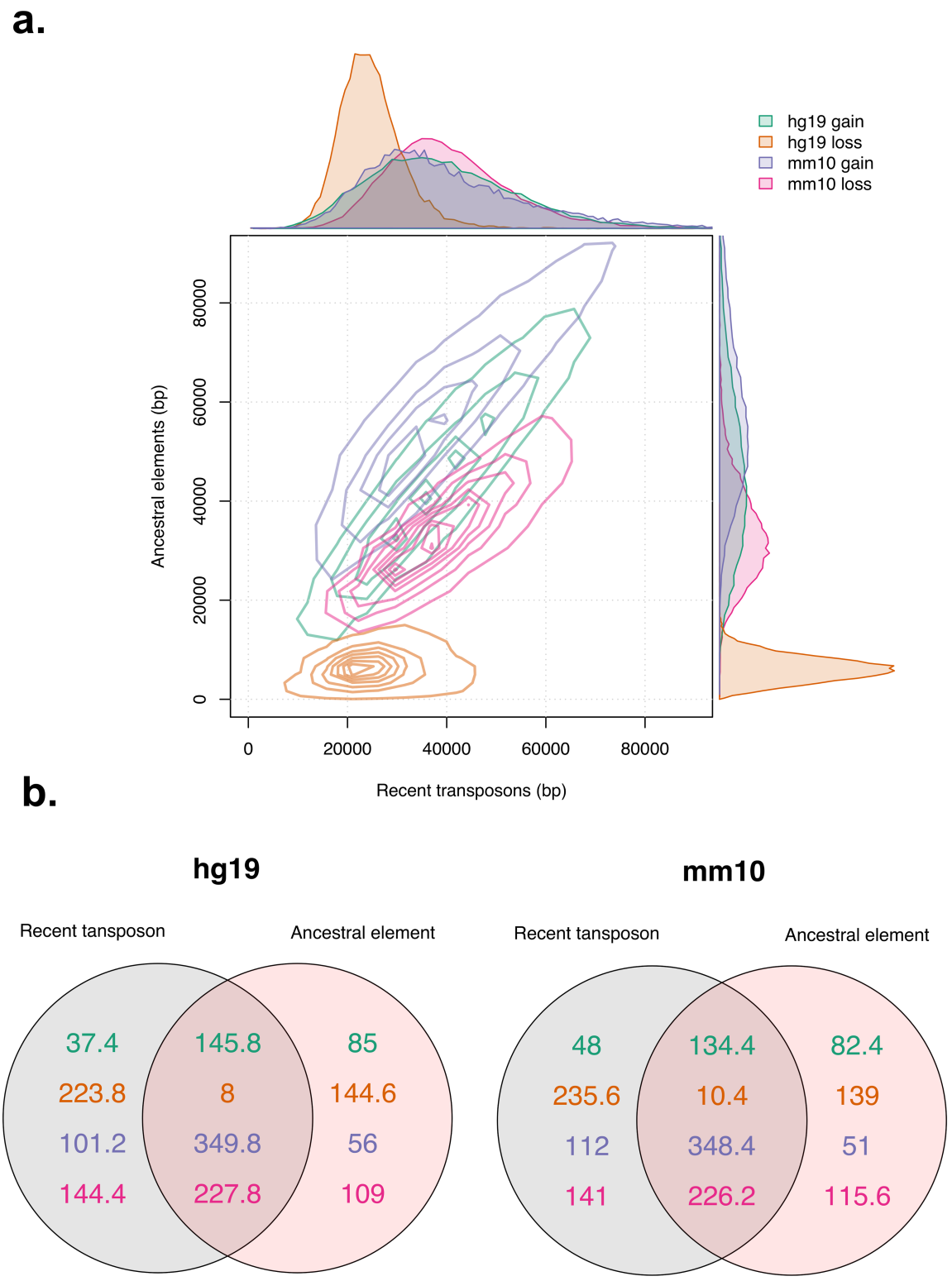


Fig 3. Comparison of gap annotation methods in binned synthetic genomes. Amount of DNA gain and loss per 200 kb in each bin for both hg19 and mm10 (a). For each gap annotation, contour lines begin at a 2D kernel density estimate of 2^{-10} and increase at regular intervals of 4^{-10}, except for hg19 which increase at regular intervals of 1.6^{-9}. Sizes of regions in Mb identified as hotspots for DNA gain or loss using the G_i^* statistic in each genome (b).

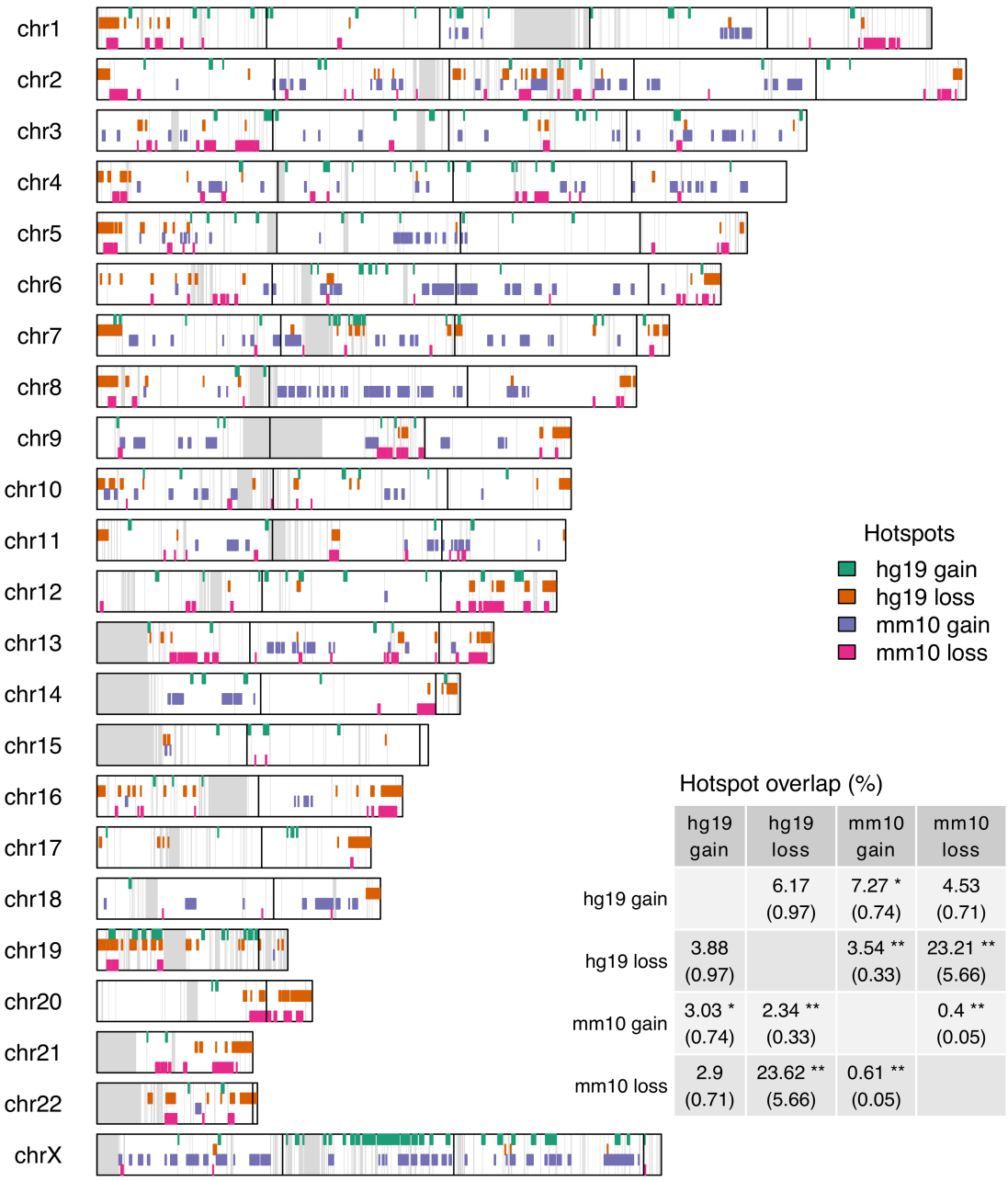
https://doi.org/10.1371/journal.pcbi.1006091.g003

significance depending on genomic background. This demonstrates that DNA loss dynamics in both hg19 and mm10 share some degree of conservation while DNA gain dynamics are mostly lineage-specific, suggesting that the acquisition of new DNA may be driving lineage-specific divergence of genome structure.

To further characterise the distribution of hg19 and mm10 gain and loss hotspots, we plotted them against both genomic backgrounds. hg19 and mm10 gain hotspots were most enriched on chromosome X (Fig 4) (S13 Fig). This is consistent with chromosome X as a hotspot for L1 insertion, a particularly large transposon with high levels of lineage specific activity that contributes to X inactivation [57]. For gain and loss hotspots themselves, hg19 gain hotspot regions were much more dispersed than other types of hotspot regions (Fig 4) (S13 Fig). Since DNA loss across both species overlaps significantly, this adds to the lineage-specific behaviour of DNA gain dynamics, where regional DNA gain in mouse is more concentrated than in human. Interestingly, DNA loss hotspots in the hg19 genomic background appear more concentrated towards telomeres, suggesting that chromosomal location may play a role in DNA loss dynamics (Fig 4). However, it is worth noting that this observation did not occur in the mm10 genomic background (S13 Fig). One explanation is that telomeres in mouse are quite recent as mouse chromosomes have undergone a high frequency of breakage and fusion events since divergence from a common ancestor [58]. In addition to analysing DNA gain and loss hotspot genomic distributions, we repeated the analyses but instead focused on the genomic distribution of DNA gain and loss coldspots (S14, S15 and S16 Figs). The most significant result was again on chromosome X, which was strongly enriched for DNA loss coldspots in human and mouse. This is consistent with low levels of homologous recombination observed on X chromosomes across mammals [59–61], as recombination is the primary mechanism that causes DNA loss [62]. Due to their evolutionary significance, we also analysed levels of DNA gain and loss surrounding chromosomal rearrangement breakpoints that were previously identified by Lemaitre *et al* [63]. We found that DNA gain and loss rates surrounding human/mouse chromosomal rearrangement breakpoints were similar to genome-wide levels (S9 Fig). Together, our results demonstrate that regional lineage-specific DNA gain and loss dynamics are relatively context-specific.

DNA gains and losses associate with distinct genomic environments

Various genomic structures and epigenetic states are known to shape and modify mutational landscapes across genomes [64]. Therefore, we examined whether gain and loss hotspots were correlated with a range of genomic features. The genomic features we analysed are non-randomly distributed and known to play various roles in genome biology. By investigating their association, we may begin to develop insight into the molecular drivers of DNA turnover. To measure the correlation between genomic features and particular gap annotations we performed feature enrichment analysis with 10,000 permutations (Methods). The analysis was performed for both mm10 gain and loss and hg19 gain and loss in both the genomic backgrounds. Using both genomic backgrounds we were able to analyse the genomic features from regions in a query genome that have been deleted from a reference. We specifically chose genomic features that could be found in both genomes as indicators for distinct aspects of genome biology. Intron density, exon density, DNaseI hypersensitivity (DNaseI HS) peaks, CpG islands, GC content and lamina-associated domains (LADs) are all indicators of genome activity [18, 30, 43, 44]. Most of these features, excluding LADs, are associated with gene dense areas and are linked to their expression or regulation [65]. LADs themselves are instead associated with gene-poor regions and gene silencing [43, 44]. We also investigated various groups of transposons whose genomic distributions have been previously characterised and used to



Hotspot overlap (%)

	hg19 gain	hg19 loss	mm10 gain	mm10 loss
hg19 gain		6.17 (0.97)	7.27 * (0.74)	4.53 (0.71)
hg19 loss	3.88 (0.97)		3.54 ** (0.33)	23.21 ** (5.66)
mm10 gain	3.03 * (0.74)	2.34 ** (0.33)		0.4 ** (0.05)
mm10 loss	2.9 (0.71)	23.62 ** (5.66)	0.61 ** (0.05)	

Fig 4. Genomic distribution of gain and loss *hotspots* for hg19 and mm10 plotted against *hg19 synthetic genome*. Grey regions indicate bins with < 150 kb of RBH nets and black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss hotspots. The percentages were calculated using the hotspots labelled in each row as the denominator. '*' and '**' represent p-values below 0.05 and 0.01 respectively based on the Fisher statistic. The odds ratio for each fisher test is reported within the brackets. An odds ratio > 1 represents a positive association and an odds ratio < 1 represents a negative association. DNA gain and loss hotspots, except for hg19 DNA loss, were identified by using both the recent transposon and ancestral element method and taking the intersect. For hg19 DNA loss, only the recent transposon method was used.

https://doi.org/10.1371/journal.pcbi.1006091.g004

investigate genome-wide DNA gain and loss rates. Lineage-specific L1s and SINEs are both major sources of DNA gain via retrotransposition, they both also have distinct accumulation profiles that are similar across both species [17]. Lineage-specific L1s tend to accumulate in gene-poor regions while lineage-specific SINEs accumulate in gene rich regions. Ancestral

L1s, and ancient elements (MIRs and L2s) have been used previously to indicate levels of DNA loss. Since these elements inserted prior to species divergence, they both provide signatures of ancestral DNA. Differences in the numbers of these elements in similar regions across species can indicate DNA loss [17, 19]. Finally, we investigated the genomic distribution of recombination hotspots and genome-wide profiles of recombination rates [33, 36]. We considered recombination as an indicator of genome instability, as meiotic recombination increases the potential for heritable genomic rearrangements [66]. Importantly, it is worth noting that recombination hotspots and recombination rates in mm10 are autosomal only. This was due to limited data availability for mouse.

Among our features we observed distinct profiles for DNA gain and loss that were largely consistent across both genomes. For DNA loss from both genomes and in both genomic backgrounds we found a strong positive associations with indicators of gene-rich/active genomic regions (Fig 5). This is surprising as biologically active genomic regions are likely to contain many important functional elements. However, it has recently been shown that these regions are particularly prone to genomic instability leading to evolutionary genomic rearrangements [67]. This also suggests DNA loss is linked to an open chromatin state as it is strongly negatively associated with LADs. In the hg19 genomic background we also found that ancient elements were positively associated with mm10 DNA loss (Fig 5). While ancient elements have been used as indicators of DNA loss, we did not expect they would be quite so strongly associated with it. Moreover, in hg19 ancient elements are negatively associated with DNA loss and have been predicted to play important roles in gene regulation [68]. In addition, the high DNA loss rate in these regions may lead to overestimates of the genome-wide DNA loss rate in mouse, as these elements have previously been used as markers for calculating deletion rates [5, 17]. Our results also showed that DNA loss in hg19 and mm10 in the hg19 genomic background was positively associated with genomic recombination (Fig 5). This is consistent with previous analyses that have identified an association between DNA loss and recombination [69]. Interestingly, we did not observe any association with recombination in the mm10 genomic background. This may be due to the decreased resolution used to calculate recombination rates and identify recombination hotspots in mouse compared to human [33, 36]. For DNA gain hotspots we found that their associations with genomic features was less consistent across both species than DNA loss hotspots (Fig 5). For sources of DNA gain, mm10 and hg19 DNA gains were both positively associated with lineage-specific L1s. However, while lineage-specific SINEs were associated with hg19 DNA gain, in mm10 they were associated with DNA loss (Fig 5). This paradoxical finding is likely caused by two separate contributing factors. The first is that lineage-specific SINEs in mouse are not a major contributor to DNA gain compared to human, as their overall coverage levels are much lower [17]. The second is that lineage-specific SINEs accumulate in gene-rich open chromatin areas which also happen to strongly associate with DNA loss [70]. These differences in sources of DNA gain may explain divergence patterns in both species DNA gain dynamics; lineage-specific SINEs are associated with gene-rich/active genomic regions and lineage-specific L1s are associated with gene-poor silent regions such as LADs. Ultimately, this suggests that DNA is accumulating/turned over in different regions at different rates by otherwise conserved mechanisms of DNA gain. Collectively, our results show that DNA gain and loss is associated with specific genomic contexts, leading to differences in genome structure.

Potential evolutionary impacts from DNA gain and loss

DNA gain and loss is non-random and may be a function of mammalian genome structure. However the evolutionary impact of DNA gain and loss is mainly determined by whether or

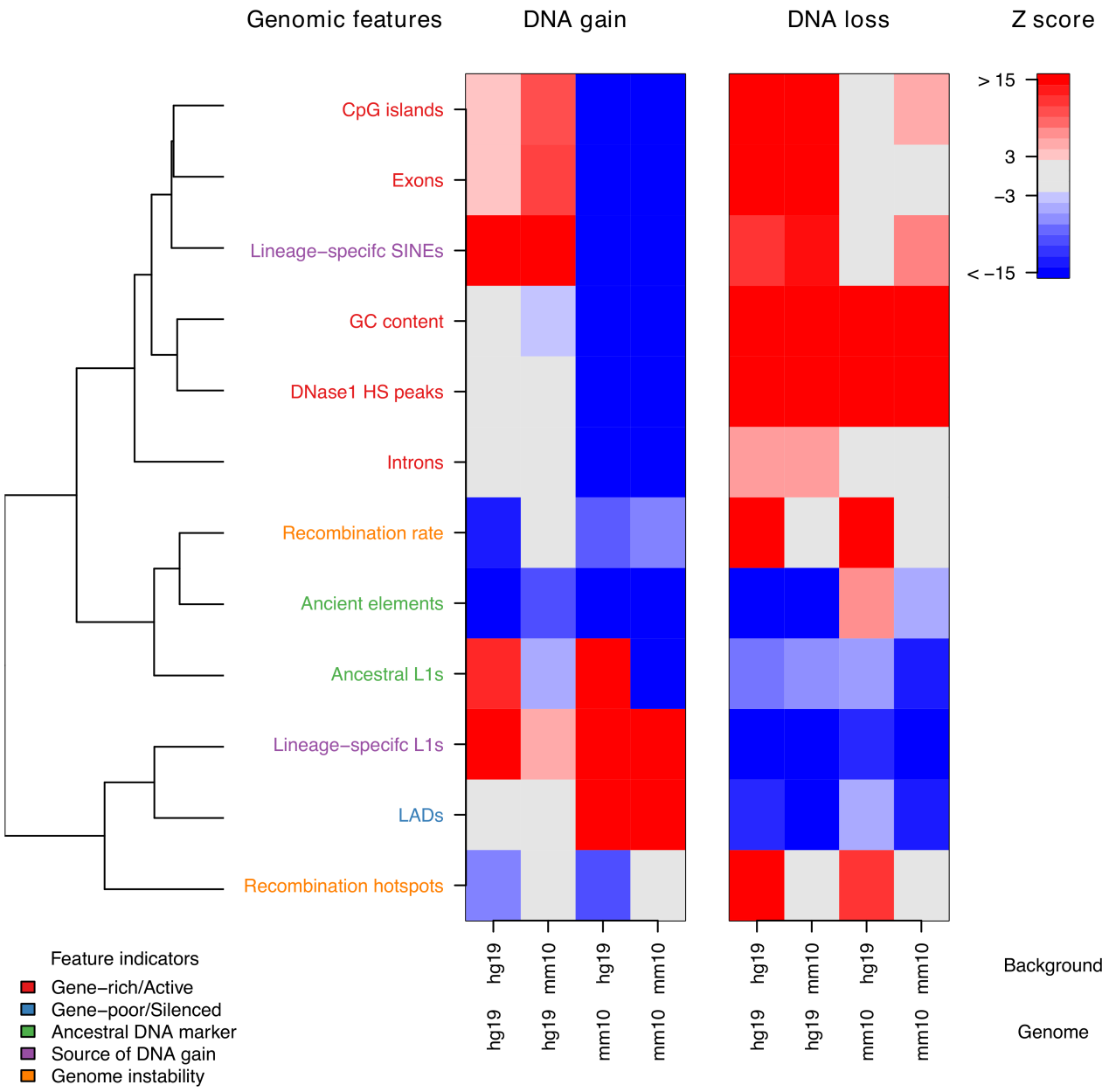


Fig 5. Association between genomic features and DNA gain or loss. Z scores are calculated using background distribution generated from 10000 permutations (Methods). A positive association indicates that a particular gap annotation and genomic feature co-locate. Alternatively, a negative association indicates that the gap annotation and genomic feature occupy distinct genomic regions. DNaseI HS peaks [30], recombination hotspots [33, 36], LADs [43, 44], CpG islands [18], gene annotations [39, 40] and Retrotransposons [23] were measured in each bin as coverage per 200 kb. Recombination rates were measured as the mean bin-wise recombination rate [33, 36]. GC content was measured as the proportion of G or C nucleotide residues in chain-blocks per bin [27, 28]. Genomic features are classified intro groups of feature indicators based on distinct aspects of genome biology they are known to associate with. The dendrogram represents spatial clustering of genomic features across both genomes,where two tightly clustered genomic features in the dendrogram are genomic features that tend to be co-located. The dendrogram was generated from a correlation matrix that consisted of pair-wise correlations between each feature across both binned genomes.

https://doi.org/10.1371/journal.pcbi.1006091.g005

not it affects particular phenotypes. To identify potentially impacted phenotypes we performed gene ontology (GO) enrichment analysis on genes in DNA gain and loss hotspots for biological process GO terms [48]. Because we are interested in identifying whether DNA gain and loss may have driven lineage-specific divergence we compared the significance levels of GO term

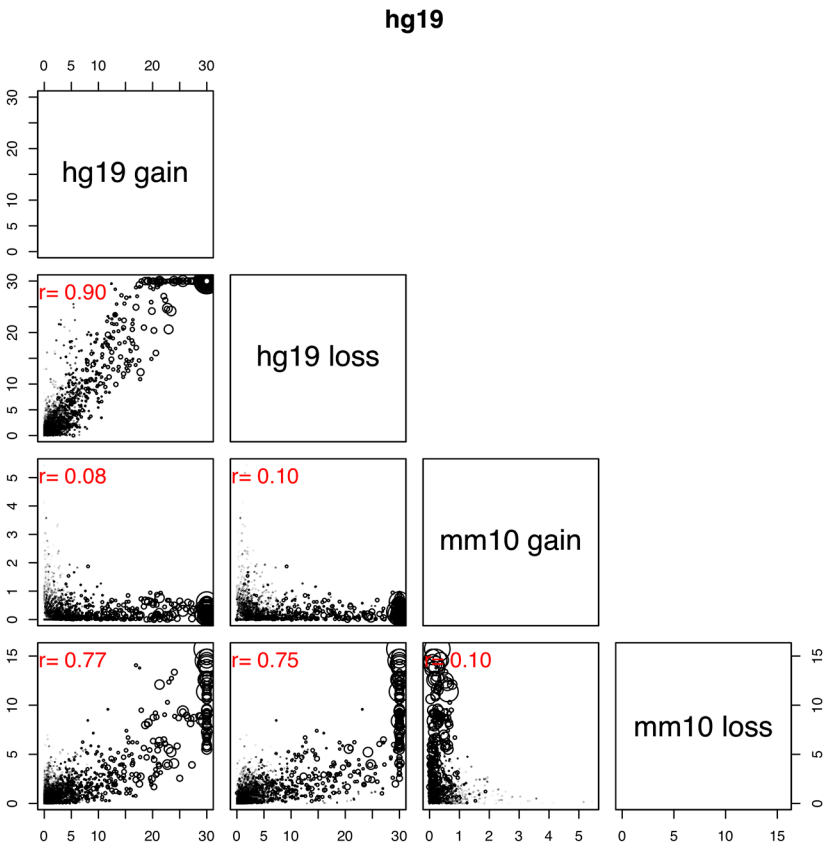


Fig 6. Over representation of biological process GO terms in gain and loss hotspots in hg19. The axes are marked according to -log10 P-values. The size of points represents the total number of annotations for each GO term. In red is the Pearson correlation coefficient.

https://doi.org/10.1371/journal.pcbi.1006091.g006

enrichment between our hotspot types. To do this we performed correlation analysis using the -log10 P-values for GO term enrichment as determined using a Fisher test combined with the 'classic' GO term enrichment algorithm (Methods) [45]. Surprisingly our results showed the highest level of similarity between hg19 DNA gain and hg19 DNA loss (Fig 6) (S17 Fig). This is interesting because the overlap between hg19 gain and loss was not statistically significant (Fig 4) (S13 Fig). Moreover, when we compare hg19 DNA loss with mm10 DNA loss; gap annotations with a significant degree of overlap (Fig 4) (S13 Fig), we found that GO terms were not as similar, particularly in the mm10 genomic background (S17 Fig). Alternatively, enriched GO terms found in mm10 DNA gain hotspots appeared distinct from GO terms enriched in other DNA gain and loss hotspots. These results echo our above findings from comparing hotspot overlap, where mm10 gains were least likely to significantly overlap other hotspot types (Fig 4) (S13 Fig).

To confirm our findings and examine the GO terms themselves, we calculated the proportion of significant terms that were descendants (child terms) of a high-order parent term. Child terms were identified as statistically significant at a FDR < 0.05 based on a Fisher test using the classic algorithm. Additionally, we extracted the 10 highest ranked terms discovered using the Fisher test combined with 3 other algorithms designed to reduce false positives generated by the inheritance problem (described in Methods) (S3, S4, S5 and S6 Tables) [46, 47]. Statistically significant terms for hg19 gain and loss mostly belonged to cellular processes, metabolic processes, single organism processes and biological regulation (Fig 7). For mm10, DNA

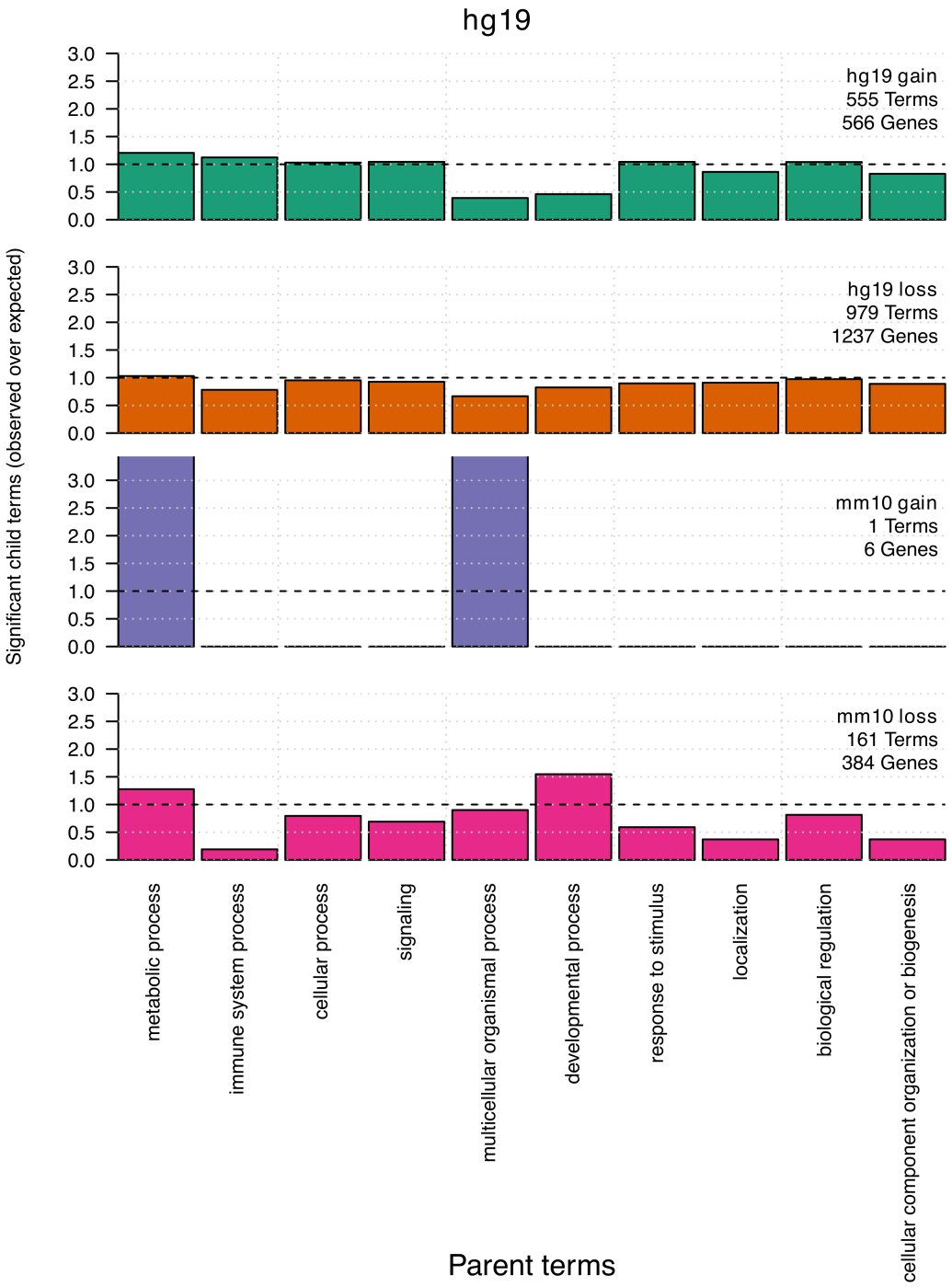


Fig 7. Significant biological process GO terms in hg19 background. Parent terms were the top level biological process GO terms while child terms were those beneath each parent term. Only Parent terms whose children make up > 5% of all terms in the genome are shown. Child terms were identified as significant at a FDR < 0.05 based on a Fisher test using the 'classic' algorithm. The Y axis represents the proportion of significant child terms belonging to a particular parent (observed), divided by the proportion of all child terms in the genome that belong to that same parent term (expected). Also shown is the number of non-redundant GO terms and genes annotated with significant GO terms for each gap annotation.

https://doi.org/10.1371/journal.pcbi.1006091.g007

loss hotspots were enriched for similar terms, including developmental processes, which were particularly enriched in the mm10 genomic background (S18 Fig). However, mm10 gain in the hg19 background was only enriched for a single term and in the mm10 background mm10 gain was not enriched for any terms. The difference in these results is consistent with how DNA gain and loss events in human and mouse associate with regions of varying gene density and biological activity (Fig 5).

Interestingly, while the genomic distributions of each hotspot type differed, their associated significant GO terms were highly similar. This may be caused by genes that contribute to similar biological processes being tightly clustered and located within regions that consist of overlapping hotspot types. To determine if this was the case we compared non-redundant statistically significant child terms and gene annotations across each hotspot type (S19 Fig). We found that the vast majority of genes annotated with significant GO terms were unique to a particular hotspot type. In contrast to this, the GO terms themselves were usually shared across hotspot types. This suggests that DNA gain and loss tend to associate with different genes that contribute to the same biological processes. Together our results show that particular biological processes are either prone to DNA gain or loss or are instead highly robust and able to withstand high levels of genomic turnover.

To determine whether or not increased DNA gain or loss likely had an evolutionary impact we compared human and mouse gene expression divergence. Gene expression divergence levels were obtained from [71] and were measured in terms of the number of commonly co-expressed genes between human and mouse one to one orthologs. We also considered orthologs which were outliers based on their levels of differential connectivity [71]. The number of genes within each group are shown in S7 Table. We found that for genes in human and mouse DNA gain and loss hotspots and developmental process genes in mouse DNA loss hotspots (identified using GO terms) there was no significant association with conserved or divergent expression patterns (S8 Table). In addition, we also measured how genes in DNA gain and loss hotspots associate with gene regulatory blocks (GRBs), genomic regions preserved between mammals and birds that are enriched for highly conserved elements [72]. Interestingly, we found that developmental genes in mm10 DNA loss hotspots were strongly enriched in GRBs (FDR < 0.001) (S8 Table), indicating that despite high levels of DNA turnover in these regions the regulatory architecture of developmental genes remains largely intact. Collectively, these results suggest that increased rates of DNA turnover have had little impact on altering gene expression patterns, since the majority of DNA turnover in these regions surrounding developmental genes has likely not interrupted regulatory element architecture.

Spatio-temporal dynamics of DNA gain and loss

Mouse and human diverged approximately 90 MYA. Over this period of time approximately 60% of their genomes have been turned over. However, changes in the rate of DNA turnover across this time-frame so far remains unknown. To better understand the spatio-temporal dynamics of DNA gain and loss, we dated individual DNA gain or loss events using a series of ingroup species that each mark specific divergence events between either human or mouse (Methods). Specifically, we dated gain or loss events that were annotated using the recent transposon method. Our results showed that both hg19 and mm10 underwent similar temporal patterns of DNA gain and loss. After the initial divergence event between human and mouse, both genomes underwent their highest rates of DNA loss which continued to slow down throughout their evolution (Fig 8a and 8b). Before humans diverged from their common ancestor with the mouse lemur, the human genome had lost approximately 275 kb (S20 Fig), and before mice had diverged from their common ancestor with the pika, the mouse genome had lost

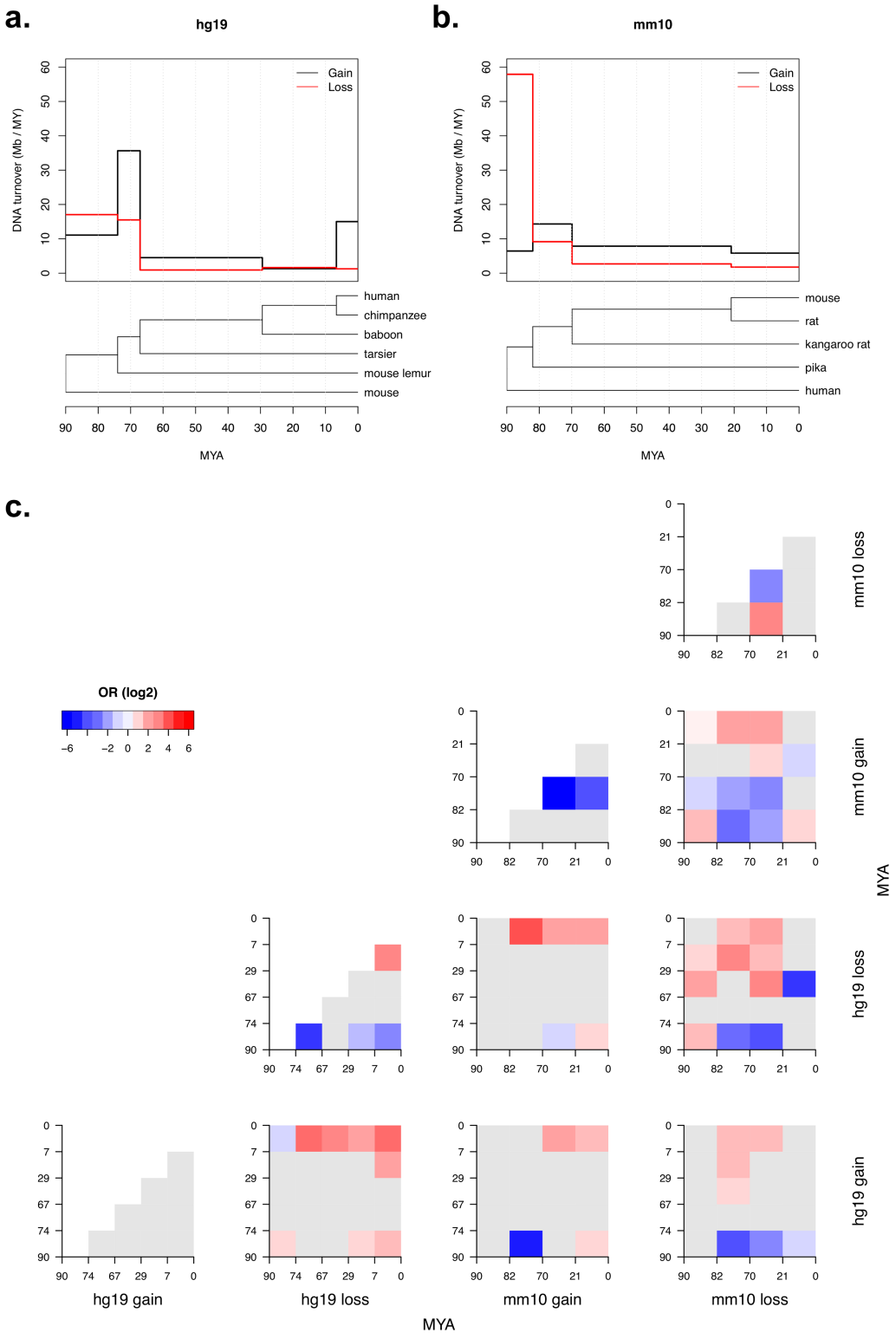


Fig 8. Rate of DNA gain and loss that occurred in Mb per million years (MY) since the human (a) and mouse (b) divergence event. Levels of DNA turnover reflect the rate of DNA gain or loss that can be attributed to each branch pictured below in the phylogenetic trees. Divergence times for the phylogenetic trees were calculated using time tree's "estimated divergence time" [49]. Overlap between DNA gain and loss hotspots in the mm10 genomic background specific to each divergence event (**c**). Significant positive and negative associations based on Fisher's exact test

(FDR < 0.05) are coloured according to their log2 odds ratio (OR). Numbers on each plot's x and y axis represent time periods which individual DNA gains and losses were assigned to.

https://doi.org/10.1371/journal.pcbi.1006091.g008

approximately 450 kb (S20 Fig). In both species this initial period of DNA loss constituted more than half of the total DNA loss they each experienced since divergence from each other (Table 2). This helps support our earlier observation about DNA loss being concentrated at telomeres in the hg19 genomic background shown in Fig 4. Since the majority of DNA loss occurred quite early after human and mouse diverged, their karyotypes were likely similar to the current human karyotype. This is likely true for two reasons; 1) there would not have been much time for a large number of chromosomal rearrangements to occur between these early ancestral human and mouse genomes, 2) and that since divergence with the boreoeutherian ancestor the human genome has undergone only a small number of chromosomal rearrangements meaning that many human telomeric regions are ancestral [58, 73]. Additionally, since over this time period DNA loss was greater than DNA gain, the results suggest that the human and mouse genomes both shrank in size before they began to grow due to transposon accumulation. Interestingly, humans underwent a recent burst of DNA gain after their divergence with chimpanzee which is consistent with rates of human-specific *Alu* and L1 activity [74] (Fig 8a).

To understand the relationship between both the spatial and temporal dynamics of DNA gain and loss, we analysed the genomic distribution of DNA gain and loss events that occurred between each divergence event. First, we identified DNA gain and loss hotspots using the hotspot identification procedure described in the methods section. Next, the genomic distribution for each set of time-specific DNA gain and loss hotspots were then compared by performing Fisher's exact test based on their overlap, hotspot overlaps were considered significant if their FDR was < 0.05. This analysis showed consistent results across both the hg19 (Fig 8c) and mm10 genomic backgrounds (S20 Fig). Overall, within species, we found that most successive time-periods of DNA gain or loss showed no statistical significant association with DNA turnover from the previous time-period. However, across species and between distant evolutionary time-periods there are particularly strong spatial associations. For example, it appears that only recent hg19 DNA gains tend to associate with DNA losses across multiple time-periods, which is consistent with recent SINE activity in human evolution [74] following insertion into gene-rich regions that are prone to DNA loss. Similarly to hg19, recent mm10 gains also strongly associated with mm10 losses from a range of time-periods. However in contrast to hg19, older mm10 DNA gain and loss events show strong negative associations with each other (Fig 8). Interestingly, hg19 loss hotspots and mm10 loss hotspots across different time-periods occasionally show negative associations. This is at odds with our earlier findings in Fig 4 that show a positive association between hg19 loss and mm10 loss. These results again indicate that genomic distributions of DNA loss have been dynamic throughout evolution. However, it is important to realise that the majority of DNA losses occurred early after human and mouse divergence, and at this early time-point hg19 and mm10 DNA loss hotspots show a positive genomic association (Fig 8). Collectively, our results show that the regional distribution of DNA gains and losses over time have been highly dynamic and most likely the result of complex interactions between genome organisation, genome biology and transposon activity.

Discussion

Genome-wide DNA gain and loss dynamics

Estimating the total amount of DNA turnover across two separate lineages over a time span of approximately 90 million years is a challenging task [49]. After this divergence period as little

as 40% of the extant human genome shares ancestry with mouse, suggesting that at least 60% has been turned over in either lineage. In order to understand gain and loss dynamics we must be able to correctly assign this non-aligning portion of the human genome as either human gain or mouse loss. Chinwalla *et al* [17] and Hardison *et al* [54] used an approach similar to our recent transposon based method. They used a set of lineage-specific transposons in human and mouse to identify regions of DNA gain. From this, the remaining non-aligning portion of one genome was assumed to be lost from the other. To confirm this approach, Chinwalla *et al* [17] checked to see if their inferred genome-wide rates of DNA loss were consistent with local estimates. They used the following equation;

$$G_E = G_A + G_G - G_L, \tag{2}$$

where G_E is the size of the extant genome, G_A is the size of the ancestral genome, G_G is the amount of lineage-specific genome gain and G_L is the amount of lineage-specific genome loss. For human and mouse they solved the equation for G_L where they estimated ancestral genome size within a range similar to the extant human genome size. This was chosen because it was similar to the average genome size for mammalian outgroup species. Estimates showed that DNA loss in mouse was almost double that of human, and consistent with the difference in the number of non-aligning non-recent transposon bases in each genome. While these estimates were consistent with expectations based on the assumption that non-aligning non-recent transposon regions were ancestral, their ancestral state remained unverified. Conversely, our ancestral based approach aimed to directly verify the ancestry status of non-aligning regions between human and mouse. This was achieved by using a wide variety of outgroup species alignments not available to Chinwalla *et al* [17] and Hardison *et al* [54] at the time of their analysis. In human, our results revealed that indeed many of the non-aligning non-recent transposon bases overlapped ancestral elements. However, approximately 168 Mb remained ambiguous (Table 2) which was more than double the 5.8% of the total non-aligning human genome, the fraction of known ancestral bases not supported by ancestral elements (Table 1). As stated in the results, this discrepancy was most likely caused by incorrect identification of DNA gain events or misidentification of ancestral elements. It is important to realise that the ancestral element-based approach has its limits, as orthologous sequences between species have the potential to diverge beyond recognition. This was the most likely reason that ancestral element detection in mouse was so much lower than in human, as the genome-wide substitution rate in mouse is approximately twice that of human.

An alternative way to verify the recent transposon based method was to use our estimated DNA loss rates to solve for G_A and to compare this to other estimates of ancestral genome sizes. After the mouse genome was completed many other mammalian genome projects also reached completion, allowing for the development of ancestral genome reconstruction techniques. While ancestral genome reconstruction is based on alignment it is much less susceptible to errors than our detection of ancestral elements. Instead of performing alignments directly between human or mouse and each individual outgroup species, it uses alignments between groups of more closely related species to build a phylogeny of ancestral states [73, 75]. Recently, Kim *et al* [76] estimated an ancestral euarchontoglires genome of 2.67 Gb in an analysis involving 19 placental mammals. Using Eq 2 and solving for G_A with extant genome sizes from Table 1 and gain and loss rates calculated by the recent transposon method (Table 2), we get estimated ancestral genome sizes of 2.64 Gb and 2.66 Gb for human and mouse respectively. Together our findings in the context of various other methods support the use of recent transposons to analyse DNA gain and loss dynamics.

While the recent transposon method provides an accurate estimate of DNA gain and loss dynamics it is important to realise these estimates are only a lower bound on the the total amount of DNA turnover since divergence. This is because both our analysis and previous analyses relied heavily on the assumption of parsimonious genome evolution, where lineage-specific gain and loss patterns are based on the fewest possible evolutionary changes. Unfortunately, in our case the assumption of parsimonious genome evolution is likely to cause various events to be hidden. For example, if a particular region underwent lineage-specific DNA gain that was subsequently lost, both the gain and loss events will not be detected. Additionally, DNA loss occurring in both lineages at the same loci would also go undetected. Depending on the frequency and magnitude of the above events we have likely underestimated the total amount of DNA gain and loss. A possible way to overcome this problem is to adopt model based approaches similar to those used in phylogenetic analyses. These approaches use a substitution model along with maximum likelihoods or Bayesian inference to allow for varying rates of evolution across lineages and sites [77]. However, given our current lack of understanding of the non-coding portion of the genome such an approach for estimating DNA turnover is likely to yield highly questionable results.

Evolutionary impact of large scale DNA gain and loss

During genome evolution the spectrum of possible mutations is extremely broad, ranging from single nucleotide substitutions all the way up to Mb-sized rearrangements and translocations. Importantly, the genomic distribution of events at each level of the mutation spectrum is non-random and highly context-dependent. Moreover, the regional susceptibility and tolerance to a particular mutation type is a mixture of various genomic and epigenomic features and selective pressures [64]. To understand the evolutionary impacts and trajectories of DNA gain and loss dynamics we analysed their genomic distributions in the context of various genomic features and biological processes.

In mammals synteny is highly conserved due to the frequent reuse of chromosome rearrangement breakpoints throughout their evolution [58]. Since chromosome rearrangement breakpoints were located outside of nets, many DNA gain and loss events went undetected (S1 and S2 Figs). Instead, we most likely identified regions where gain and loss dynamics impacted on local architecture, such as the genomic distances between neighbouring genes or intron size. However, due to the difficulty in mapping DNA gain and loss events across large evolutionary time scales, the impact of DNA gain and loss at this scale remains largely unknown. Our strategy has therefore allowed us for the first time to measure regional variation in DNA gain and loss across genome structures that have been resistant to large structural rearrangements. Our results revealed that DNA gains and losses in human and mouse were associated with the same kinds of features; DNA gains were most associated with L1 accumulation in gene poor regions with low biological activity while DNA losses occurred mostly in highly active gene-rich regions. Previous analyses have shown that genome organisation between human and mouse is largely conserved, where lineage-specific L1s and SINEs tend to accumulate in similar regions in different species [70]. Our results suggest that rather than certain types of events driving genome divergence, it is instead the rate at which each particular event type occurs that drives divergence. For example, mouse has a much higher deletion rate than human and a larger number of active L1s. This would suggest that particular regions in the mouse are growing or shrinking much more than in the human genome while their sequence composition remains similar. Alternatively, DNA gain rates were especially enriched on the X chromosome in both species with some degree of regional overlap (Fig 4) (S13 Fig). This is consistent with the high concentration of L1s that play a role in X inactivation [57].

Despite the amount of structural divergence between human and mouse, it is difficult to identify how much impact this might have on evolution at the level of phenotype. Interestingly, Human DNA gains and losses and mouse DNA losses all occurred near genes involved in fundamental cellular/metabolic processes. Because cellular/metabolic process genes likely evolved earlier in animals and probably have house keeping functions, their regulation is also likely highly conserved [78]. This suggests that for the most part the accumulation of DNA gains and losses have had little impact on phenotypic change. However, for some mouse DNA losses the case may be different, as in the mm10 genomic background they mostly occurred near genes involved in developmental processes. Developmental processes may be linked to traits that could have potentially undergone divergence, such as mouse-specific morphological characteristics. While this is an attractive idea, an analysis of regulatory element evolution shows that lineage-specific regulatory innovation for development occurred prior to human and mouse divergence [78]. Moreover, we observed that developmental genes associated with mm10 DNA loss hotspots were in genomic regions enriched for conserved elements that likely contribute to conservation of gene regulation [72]. Therefore, throughout mammalian evolution regulatory elements for development and cellular processes have likely remained intact while nearby DNA has been frequently turned over. This has important implications for calculating the "functional" proportion of mammalian genomes, depending on the methods used and how the term itself is applied, this value ranges widely. Using transcription and DNA binding to identify functional DNA, the ENCODE consortium estimated that as much as 80% of the human genome might be functional [30]. Alternatively, evolutionary approaches have been used to identify functional regions as those that are likely to have a measurable biological impact on cell function if perturbed. These kinds of approaches suggest that no more than 25% of the human genome is functional [79, 80]. Ultimately, given that we are able to detect little phenotypic impact where there are vast amounts of DNA turnover, our findings support lower estimates for the functional proportion of the human genome.

Conclusion

There are four key points from our results. First, hot spots for DNA gains and losses occur in different compartments; loss hotspots in open chromatin/regulatory regions and gain hotspots in heterochromatin. Because DNA loss is caused by repair of DNA Double Stranded Breaks (DSB) [81], this means that L1 ORF2p activity can both cause DNA gains and losses as a cause of DSB. However, this does not mean that gains and losses do not occur in the same regions. Second, mouse SINEs are strongly associated with DNA loss, indicating that losses in regulatory regions are accompanied by SINE insertions suggesting that there is extensive "churning" or turnover of sequences in these regions. The observed differences in associations between lineage-specific SINEs and gain and loss in mouse and human are likely due to differential expansion of LINEs vs SINEs in the two lineages. Thus, regional/species specific variation in DNA gain and loss are primarily driven by clade specific/recent transposons interacting with open chromatin either in the male germ line, female germ line or early embryo. Third, the X chromosome is largely devoid of loss hotspots, but has many gain hotspots, consistent with a continuing selection for insertion of L1 elements required for X inactivation. Fourth, the observed autosomal divergence of gain and loss hotspot patterns in proximity to genes supports a model in which developmental/regulatory mechanisms (based on GO term results) are robust to large amounts of transposon driven DNA gain and loss. This has implications for our views regarding the "functional" proportion of the genome that is under selection and contributing to phenotypic divergence.

Supporting information

S1 Fig. Genomic regions filtered from hg19. Gaps outside of nets $\geq$ 10 kb are shown in black above each chromosome. non-RBH regions $\geq$ 10 kb are shown in red below each chromosome. Assembly gaps are plotted in black within chromosomes. Syntenic blocks are coloured according to which chromosome they belong to in mm10. The trace running through each syntenic block represents its mm10 chromosomal position and orientation, running top to bottom (5′ to 3′).
(TIFF)

S2 Fig. Genomic regions filtered from mm10. Gaps outside of nets $\geq$ 10 kb are shown in black above each chromosome. non-RBH regions $\geq$ 10 kb are shown in red below each chromosome. Assembly gaps are plotted in black within chromosomes. Syntenic blocks are coloured according to which chromosome they belong to in hg19. The trace running through each syntenic block represents its hg19 chromosomal position and orientation, running top to bottom (5′ to 3′).
(TIFF)

S3 Fig. Coverage depth of chain-blocks extracted from outgroup species. Coverage depth is measured by number of overlapping outgroup species. Ancestral DNA % is the proportion of total bp in hg19 and mm10 that overlap at least one chain-block extracted from an outgroup species.
(TIFF)

S4 Fig. Error profile and coverage depth for identifying ancestral elements. Minimum coverage depth threshold for identifying ancestral elements is plotted against total proportion of identified type 1 errors and the proportional increase in type 2 error rate. Type 1 errors are identified as known recent transposons that overlap chain-blocks extracted from outgroup species. Type 2 errors are identified as chain-blocks between hg19 and mm10 that do not overlap chain-blocks extracted from outgroup species. Type 2 error increase is the reduction in the overlap between outgroup and ingroup (hg19 and mm10) chain-blocks as minimum coverage depth threshold increases. For hg19 and mm10 we chose a minimum coverage depth of 6 and 4 respectively.
(TIFF)

S5 Fig. Transposon family classification with linear discriminant analysis. Each rectangle represents the members of a transposon family under our prior recent and ancestral classification. For example, a rectangle coloured black represents the members of a particular transposon family that do not overlap ancestral elements. Rectangle width is the interquartile range of percent divergence from consensus and rectangle height is proportional to total genome coverage. The dotted line is the classification boundary determined by linear discriminant analysis. Rectangles above the line are transposon families classified as recent and rectangles below the line are transposon families classified as ancestral.
(TIFF)

S6 Fig. Transposon family period of activity and percent divergence from consensus. Transposons identified as recently active were classified as recent by our classifier and belong to families not shared between human and mouse. Transposons identified as active during divergence were classified as recent by our classifier and belong to families shared between human and mouse. Transposons identified as active within the ancestor were classified as ancestral by our classifier and belong to families shared between human and mouse. Transposons classified as ancestral by our classifier that belong to families not shared between human

and mouse are not shown.
(TIFF)

S7 Fig. Rank comparison of gap annotation methods per 200 kb bin in hg19 genomic background.
(TIFF)

S8 Fig. Rank comparison of gap annotation methods per 200 kb bin in mm10 genomic background.
(TIFF)

S9 Fig. DNA gain and loss rates surrounding human and mouse rearrangement breakpoints in the hg19 genomic background.
(TIFF)

S10 Fig. Sum of all gap annotation types per genomic bin for various bin sizes.
(TIFF)

S11 Fig. The impact of using different neighbour distances for calculating G_i^*. The neighbour distances is shown in red. A neighbour distance of 3 indicates that 3 bins upstream and 3 bins downstream of a particular bin are considered it's neighbours. R^2 is the coefficient of determination between our G_i^* values and the actual bin-wise density for a specific gap annotation. "roughness" is calculated as the standard deviation of the differences between adjacent bins, lower values indicate the degree of smoothing caused by increasing the neighbour distance.
(TIFF)

S12 Fig. Examples of DNA gain and loss hotspots at various neighbour distances. Neighbour distances are indicated in the top left corner of each plot in blue. Hotspots for human and mouse DNA gain and loss are shown in red.
(TIFF)

S13 Fig. Genomic distribution of gain and loss *hotspots* for hg19 and mm10 plotted against *mm10 synthetic genome*. Grey regions indicate bins with $\leq$ 150 kb of RBH nets and black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss hotspots. The percentages were calculated using the hotspots labelled in each row as the denominator. '*' and '**' represent p-values below .05 and .01 respectively based on the Fisher statistic.
(TIFF)

S14 Fig. Comparison of gap annotation methods in binned synthetic genomes. Regions identified as coldspots (Mb) for DNA gain or loss using the G_i^* statistic in each genome.
(TIFF)

S15 Fig. Genomic distribution of gain and loss *coldspots* for hg19 and mm10 plotted against *hg19 synthetic genome*. Grey regions indicate bins with $\leq$ 150 kb of RBH nets and black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss coldspots. The percentages were calculated using the coldspots labelled in each row as the denominator. '*' and '**' represent p-values below .05 and .01 respectively based on the Fisher statistic.
(TIFF)

S16 Fig. Genomic distribution of gain and loss *coldspots* for hg19 and mm10 plotted against *mm10 synthetic genome*. Grey regions indicate bins with $\leq$ 150 kb of RBH nets and

black vertical lines represent 50 Mb on non-synthetic genome. Inset table represents percent overlap of gain and loss coldspots. The percentages were calculated using the coldspots labelled in each row as the denominator. '*' and '**' represent p-values below .05 and .01 respectively based on the Fisher statistic.
(TIFF)

S17 Fig. Over representation of biological process GO terms in gain and loss hotspots in mm10. The axes are marked according to -log10 P-values. The size of points represents the total umber of annotations for each GO term. In red is the Pearson correlation coefficient.
(TIFF)

S18 Fig. Significant biological process GO terms in mm10 background. Parent terms were the top level biological process GO terms while child terms were those beneath each parent term. Only Parent terms whose children make up $>$ 5% of all terms in the genome are shown. Child terms were identified as significant at a FDR $<$ 0.05 based on a Fisher test using the 'classic' algorithm. The Y axis represents the proportion of significant child terms belonging to a particular parent (observed), divided by the proportion of all child terms in the genome that belong to that same parent term (expected). Also shown is the number of non-redundant GO terms and genes annotated with significant GO terms for each gap annotation.
(TIFF)

S19 Fig. Comparison of significant biological process GO terms and annotated genes. GO terms were identified as significant at a FDR $<$ 0.05 based on a Fisher test using the 'classic' algorithm. Annotated genes are genes that have been annotated with at least one of the significant GO terms. GO term lists and gene lists in each set are non-redundant.
(TIFF)

S20 Fig. Spatio-temporal DNA gain and loss. *Amount of DNA gain and loss in Mb* that occurred since the human (a) and mouse (b) divergence event. Levels of DNA turnover reflect the amount of DNA gain or loss that can be attributed to each branch pictured below in the phylogenetic trees. Divergence times for the phylogenetic trees were calculated using time tree's "estimated divergence time". Overlap between DNA gain and loss hotspots in the **mm10 genomic background** specific to each divergence event (**c**). Significant positive and negative associations based on Fisher's exact test (FDR $<$ 0.05) are coloured according to their log2 odds ratio (OR). Numbers on each plot's x and y axis represent time periods which individual DNA gains and losses were assigned to.
(TIFF)

S1 Table. hg19 and mm10 classification of transposon families. Transposon classification compares our LDA classifier against shared and lineage-specific transposon family names. Presented is the total Mb transposon coverage with number of families in brackets.
(TIFF)

S2 Table. List of outgroup genomes used to identify ancestral elements in hg19 and mm10.
(TIFF)

S3 Table. Top 10 biological process GO terms for genes located in hg19 gain hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in Methods).
(TIFF)

S4 Table. Top 10 biological process GO terms for genes located in hg19 loss hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four

separate algorithms that each take the GO hierarchy into account (described in Methods).
(TIFF)

S5 Table. Top 10 biological process GO terms for genes located in mm10 gain hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in Methods).
(TIFF)

S6 Table. Top 10 biological process GO terms for genes located in mm10 loss hotspots. P-values for each GO term were calculated using the fisher statistic combined with one of four separate algorithms that each take the GO hierarchy into account (described in Methods).
(TIFF)

S7 Table. Gene content of various genomic regions in human and mouse. mm10 loss developmental genes are genes in mouse loss hotspots that are annotated with developmental process GO terms. GRBs (gene regulatory blocks) are regions in the human genome that are enriched for conserved elements. Top 10% CCGs (commonly co-expressed genes) are the genes with the highest amount of co-expressed orthologs shared between human and mouse, these genes are likely to have conserved expression between both species. Bottom 10% CCGs are the genes with the least amount of co-expressed orthologs shared between human and mouse, these genes are likely to have divergent expression patterns between both species. Differentially connected genes are genes with the highest amount of differential connectivity between human and mouse, these are genes with non-conserved expression patterns. All human and mouse orthologs are one to one.
(TIFF)

S8 Table. Shared genes between overlapping regions. For each region comparison the percentage of genes found in gap annotation hotspots was reported. The statistical significance of each overlap was measured using a Fisher's exact test. Shown in bold font is the odds ratio and shown in brackets is the FDR.
(TIFF)

S1 Text. URLs for accessing the data that was used throughout the analysis.
(TXT)

Acknowledgments

We would like to thank Steve Pederson, Rick Tearle, Jonathan Henry Jacobsen, Lu Zeng and Zhipeng Qu for their helpful discussion throughout the research process and Catisha Coburn for help with editing the manuscript.

Author Contributions

Conceptualization: Reuben M. Buckley, R. Daniel Kortschak, David L. Adelson.

Formal analysis: Reuben M. Buckley.

Funding acquisition: David L. Adelson.

Methodology: Reuben M. Buckley.

Project administration: David L. Adelson.

Software: Reuben M. Buckley, R. Daniel Kortschak.

Supervision: R. Daniel Kortschak, David L. Adelson.

Writing – original draft: Reuben M. Buckley, R. Daniel Kortschak, David L. Adelson.

Writing – review & editing: Reuben M. Buckley, R. Daniel Kortschak, David L. Adelson.

References

1. Gregory TR. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. Annals of botany. 2005; 95(1):133–146. https://doi.org/10.1093/aob/mci009 PMID: 15596463
2. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Phil Trans R Soc B. 2015; 370(1678):20140331. https://doi.org/10.1098/rstb.2014.0331 PMID: 26323762
3. Gregory TR. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biological reviews. 2001; 76(1):65–101. https://doi.org/10.1111/j.1469-185X.2000.tb00059.x PMID: 11325054
4. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences. 2017; 114(8):E1460–E1469. https://doi.org/10.1073/pnas.1616702114
5. Consortium IHGS, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409 (6822):860. https://doi.org/10.1038/35057062
6. Lynch M, Walsh B. The origins of genome architecture. vol. 98. Sinauer Associates Sunderland (MA); 2007.
7. Whitney KD, Garland T Jr. Did genetic drift drive increases in genome complexity? PLoS genetics. 2010; 6(8):e1001080. https://doi.org/10.1371/journal.pgen.1001080 PMID: 20865118
8. Wright NA, Gregory TR, Witt CC. Metabolic engines of flight drive genome size reduction in birds. Proc R Soc B. 2014; 281(1779):20132780. https://doi.org/10.1098/rspb.2013.2780 PMID: 24478299
9. Vinogradov AE, Anatskaya OV. Genome size and metabolic intensity in tetrapods: a tale of two lines. Proceedings of the Royal Society of London B: Biological Sciences. 2006; 273(1582):27–32. https://doi.org/10.1098/rspb.2005.3266
10. Lynch M, Conery JS. The origins of genome complexity. science. 2003; 302(5649):1401–1404. https://doi.org/10.1126/science.1089370 PMID: 14631042
11. Hedges D, Deininger P. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2007; 616(1):46–59. https://doi.org/10.1016/j.mrfmmm.2006.11.021 PMID: 17157332
12. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Molecular biology and evolution. 2003; 20 (6):880–892. https://doi.org/10.1093/molbev/msg102 PMID: 12716993
13. Fan Y, Huang ZY, Cao CC, Chen CS, Chen YX, Fan DD, et al. Genome of the Chinese tree shrew. Nature communications. 2013; 4:1426. https://doi.org/10.1038/ncomms2416 PMID: 23385571
14. Kvikstad EM, Chiaromonte F, Makova KD. Ride the wavelet: a multiscale analysis of genomic contexts flanking small insertions and deletions. Genome research. 2009; 19(7):1153–1164. https://doi.org/10.1101/gr.088922.108 PMID: 19502380
15. Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS computational biology. 2007; 3(9):e176. https://doi.org/10.1371/journal.pcbi.0030176
16. Gasior SL, Preston G, Hedges DJ, Gilbert N, Moran JV, Deininger PL. Characterization of pre-insertion loci of de novo L1 insertions. Gene. 2007; 390(1):190–198. https://doi.org/10.1016/j.gene.2006.08.024 PMID: 17067767
17. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420(6915):520–562. https://doi.org/10.1038/nature01262 PMID: 12466850
18. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic acids research. 2016; 45(D1):D626–D634. https://doi.org/10.1093/nar/gkw1134 PMID: 27899642
19. Laurie S, Toll-Riera M, Radó-Trilla N, Albà MM. Sequence shortening in the rodent ancestor. Genome research. 2012; 22(3):478–485. https://doi.org/10.1101/gr.121897.111 PMID: 22128134
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002; 12(6):996–1006. https://doi.org/10.1101/gr.229102 PMID: 12045153

21. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences. 2003; 100(20):11484–11489. https://doi.org/10.1073/pnas.1932072100
22. Lee J, Hong Wy, Cho M, Sim M, Lee D, Ko Y, et al. Synteny Portal: a web-based application portal for synteny block analysis. Nucleic acids research. 2016; 44(W1):W35–W40. https://doi.org/10.1093/nar/gkw310 PMID: 27154270
23. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0.; 2013-2015.
24. Bivand R, Hauke J, Kossowski T. Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. Geographical Analysis. 2013; 45(2):150–179. https://doi.org/10.1111/gean.12008
25. Bivand R, Piras G. Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software. 2015; 63(18):1–36. https://doi.org/10.18637/jss.v063.i18
26. Getis A, Ord JK. Local spatial statistics: an overview. Spatial analysis: modelling in a GIS environment. 1996; 374:261–277.
27. Team TBD. BSgenome.Hsapiens.UCSC.hg19: Full genome sequences for Homo sapiens (UCSC version hg19); 2014.
28. Team TBD. BSgenome.Mmusculus.UCSC.mm10: Full genome sequences for Mus musculus (UCSC version mm10); 2014.
29. Pages H. BSgenome: Infrastructure for Biostrings-based genome data packages; 2017.
30. ENCODE Project Consortium, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57. https://doi.org/10.1038/nature11247 PMID: 22955616
31. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC genome browser database: update 2006. Nucleic acids research. 2006; 34(suppl_1):D590–D598. https://doi.org/10.1093/nar/gkj144 PMID: 16381938
32. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PloS one. 2012; 7(1):e30377. https://doi.org/10.1371/journal.pone.0030377 PMID: 22276185
33. International HapMap Consortium, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164):851. https://doi.org/10.1038/nature06258 PMID: 17943122
34. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005; 308(5718):107–111. https://doi.org/10.1126/science.1105322 PMID: 15705809
35. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004; 304(5670):581–584. https://doi.org/10.1126/science.1092500 PMID: 15105499
36. Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. Fine-scale maps of recombination rates and hotspots in the mouse genome. Genetics. 2012; 191(3):757–764. https://doi.org/10.1534/genetics.112.141036 PMID: 22562932
37. Kirby A, Kang HM, Wade CM, Cotsapas C, Kostem E, Han B, et al. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics. 2010; 185(3):1081–1095. https://doi.org/10.1534/genetics.110.115014 PMID: 20439770
38. Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics. 2011; 43(7):648–655. https://doi.org/10.1038/ng.847 PMID: 21623374
39. Carlson M. TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s); 2015.
40. Carlson M. TxDb.Mmusculus.UCSC.mm10.knownGene: Annotation package for TxDb object(s); 2016.
41. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLoS Computational Biology. 2013; 9. https://doi.org/10.1371/journal.pcbi.1003118 PMID: 23950696
42. Giordano J, Ge Y, Gelfand Y, Abrusán G, Benson G, Warburton PE. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS computational biology. 2007; 3(7):e137. https://doi.org/10.1371/journal.pcbi.0030137 PMID: 17630829
43. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453(7197):948. https://doi.org/10.1038/nature06947 PMID: 18463634
44. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular cell. 2010; 38(4):603–613. https://doi.org/10.1016/j.molcel.2010.03.016 PMID: 20513434

45. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology; 2016.
46. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006; 22(13):1600–1607. https://doi.org/10.1093/bioinformatics/btl140 PMID: 16606683
47. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. Bioinformatics. 2007; 23(22):3024–3031. https://doi.org/10.1093/bioinformatics/btm440 PMID: 17848398
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25. https://doi.org/10.1038/75556 PMID: 10802651
49. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22(23):2971–2972. https://doi.org/10.1093/bioinformatics/btl505 PMID: 17021158
50. R Core Team. R: A Language and Environment for Statistical Computing; 2015. Available from: http://www.R-project.org/.
51. Ooms J, James D, DebRoy S, Wickham H, Horner J. RMySQL: Database Interface and'MySQL' Driver for R; 2016. Available from: http://CRAN.R-project.org/package=RMySQL.
52. Wickham H, Francois R. dplyr: A Grammar of Data Manipulation; 2015. Available from: http://CRAN.R-project.org/package=dplyr.
53. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004; 5(10):R80. https://doi.org/10.1186/gb-2004-5-10-r80 PMID: 15461798
54. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome research. 2003; 13(1):13–26. https://doi.org/10.1101/gr.844103 PMID: 12529302
55. Kapitonov V, Jurkal J. The age of Alu subfamilies. Journal of molecular evolution. 1996; 42(1):59–65. https://doi.org/10.1007/BF00163212 PMID: 8576965
56. Smit AF, Tóth G, Riggs AD, Jurka J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. Journal of molecular biology. 1995; 246(3):401–417. https://doi.org/10.1006/jmbi.1994.0095 PMID: 7877164
57. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell. 2010; 141(6):956–969. https://doi.org/10.1016/j.cell.2010.04.042 PMID: 20550932
58. Murphy WJ, Larkin DM, Everts-Van Der Wind A, Bourque G, Tesler G, Auvil L, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 2005; 309 (5734):613–617. https://doi.org/10.1126/science.1111387 PMID: 16040707
59. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, et al. Comparative recombination rates in the rat, mouse, and human genomes. Genome research. 2004; 14(4):528–538. https://doi.org/10.1101/gr.1970304 PMID: 15059993
60. Ma J, Iannuccelli N, Duan Y, Huang W, Guo B, Riquet J, et al. Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. BMC genomics. 2010; 11(1):159. https://doi.org/10.1186/1471-2164-11-159 PMID: 20211033
61. Li G, Davis BW, Eizirik E, Murphy WJ. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). Genome research. 2016; 26(1):1–11. https://doi.org/10.1101/gr.186668.114 PMID: 26518481
62. Gebow D, Miselis N, Liber HL. Homologous and nonhomologous recombination resulting in deletion: effects of p53 status, microhomology, and repetitive DNA length and orientation. Molecular and cellular biology. 2000; 20(11):4028–4035. https://doi.org/10.1128/MCB.20.11.4028-4035.2000 PMID: 10805745
63. Lemaitre C, Tannier E, Gautier C, Sagot MF. Precise detection of rearrangement breakpoints in mammalian chromosomes. BMC bioinformatics. 2008; 9(1):286. https://doi.org/10.1186/1471-2105-9-286 PMID: 18564416
64. Makova KD, Hardison RC. The effects of chromatin organization on variation in mutation rates in the genome. Nature Reviews Genetics. 2015; 16(4):213–223. https://doi.org/10.1038/nrg3890 PMID: 25732611
65. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012; 489(7414):75. https://doi.org/10.1038/nature11232 PMID: 22955617

66. Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature genetics. 2010; 42(10):859–863. https://doi.org/10.1038/ng.658 PMID: 20818382

67. Berthelot C, Muffato M, Abecassis J, Crollius HR. The 3D organization of chromatin explains evolutionary fragile genomic regions. Cell reports. 2015; 10(11):1913–1924. https://doi.org/10.1016/j.celrep.2015.02.046 PMID: 25801028

68. Kamal M, Xie X, Lander ES. A large family of ancient repeat elements in the human genome is under strong selection. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(8):2740–2745. https://doi.org/10.1073/pnas.0511238103 PMID: 16477033

69. Nam K, Ellegren H. Recombination drives vertebrate genome contraction. PLoS genetics. 2012; 8(5): e1002680. https://doi.org/10.1371/journal.pgen.1002680 PMID: 22570634

70. Buckley RM, Kortschak RD, Raison JM, Adelson DL. Similar evolutionary trajectories for retrotransposon accumulation in mammals. Genome biology and evolution. 2017; 9(9):2336–2353. https://doi.org/10.1093/gbe/evx179 PMID: 28945883

71. Monaco G, van Dam S, Ribeiro JLCN, Larbi A, de Magalhães JP. A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. BMC evolutionary biology. 2015; 15(1):259. https://doi.org/10.1186/s12862-015-0534-7 PMID: 26589719

72. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nature communications. 2017; 8(1):441. https://doi.org/10.1038/s41467-017-00524-5 PMID: 28874668

73. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, et al. Reconstructing contiguous regions of an ancestral genome. Genome research. 2006; 16(12):1557–1565. https://doi.org/10.1101/gr.5383506 PMID: 16983148

74. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, et al. Rates and patterns of great ape retrotransposition. Proceedings of the National Academy of Sciences. 2013; 110 (33):13457–13462. https://doi.org/10.1073/pnas.1310914110

75. Blanchette M, Green ED, Miller W, Haussler D. Reconstructing large regions of an ancestral mammalian genome in silico. Genome research. 2004; 14(12):2412–2423. https://doi.org/10.1101/gr.2800104 PMID: 15574820

76. Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, et al. Reconstruction and evolutionary history of eutherian chromosomes. Proceedings of the National Academy of Sciences. 2017; 114(27):E5379–E5388. https://doi.org/10.1073/pnas.1702012114

77. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nature reviews Genetics. 2012; 13(5):303. https://doi.org/10.1038/nrg3186 PMID: 22456349

78. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, et al. Three periods of regulatory innovation during vertebrate evolution. science. 2011; 333(6045):1019–1024. https://doi.org/10.1126/science.1202702 PMID: 21852499

79. Graur D. An upper limit on the functional fraction of the human genome. Genome biology and evolution. 2017; 9(7):1880–1885. https://doi.org/10.1093/gbe/evx121 PMID: 28854598

80. Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS genetics. 2014; 10(7): e1004525. https://doi.org/10.1371/journal.pgen.1004525 PMID: 25057982

81. Gasior SL, Wakeman TP, Xu B, Deininger PL. The human LINE-1 retrotransposon creates DNA double-strand breaks. Journal of molecular biology. 2006; 357(5):1383–1393. https://doi.org/10.1016/j.jmb.2006.01.089 PMID: 16490214