


Human L1 Transposition Dynamics Unraveled with Functional Data Analysis

Di Chen,^{†,1} Marzia A. Cremona,^{†,2,3} Zongtai Qi,⁴ Robi D. Mitra,⁴ Francesca Chiaromonte,^{*,2,5,6} and Kateryna D. Makova ^{*,6,7}

¹Intercollege Graduate Degree Program in Genetics, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA

²Department of Statistics, The Pennsylvania State University, University Park, PA

³Department of Operations and Decision Systems, Université Laval, Québec, Canada

⁴Department of Genetics and Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO

⁵EMbeDS, Sant'Anna School of Advanced Studies, Pisa, Italy

⁶The Huck Institutes of the Life Sciences, Center for Medical Genomics, The Pennsylvania State University, University Park, PA

⁷Department of Biology, The Pennsylvania State University, University Park, PA

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: fxc11@psu.edu; kdm16@psu.edu.

Associate editor: Amanda Larracuent

Abstract

Long Interspersed Elements-1 (L1s) constitute >17% of the human genome and still actively transpose in it. Characterizing L1 transposition across the genome is critical for understanding genome evolution and somatic mutations. However, to date, L1 insertion and fixation patterns have not been studied comprehensively. To fill this gap, we investigated three genome-wide data sets of L1s that integrated at different evolutionary times: 17,037 de novo L1s (from an L1 insertion cell-line experiment conducted in-house), and 1,212 polymorphic and 1,205 human-specific L1s (from public databases). We characterized 49 genomic features—proxyming chromatin accessibility, transcriptional activity, replication, recombination, etc.—in the ± 50 kb flanks of these elements. These features were contrasted between the three L1 data sets and L1-free regions using state-of-the-art Functional Data Analysis statistical methods, which treat high-resolution data as mathematical functions. Our results indicate that de novo, polymorphic, and human-specific L1s are surrounded by different genomic features acting at specific locations and scales. This led to an integrative model of L1 transposition, according to which L1s preferentially integrate into open-chromatin regions enriched in non-B DNA motifs, whereas they are fixed in regions largely free of purifying selection—depleted of genes and noncoding most conserved elements. Intriguingly, our results suggest that L1 insertions modify local genomic landscape by extending CpG methylation and increasing mononucleotide microsatellite density. Altogether, our findings substantially facilitate understanding of L1 integration and fixation preferences, pave the way for uncovering their role in aging and cancer, and inform their use as mutagenesis tools in genetic studies.

Key words: transposable elements, LINE-1, transposition, fixation, integration.

Introduction

More than 45% of the human genome consists of transposable elements (TEs), including >17% occupied by Long Interspersed Element type 1, abbreviated as LINE-1 or L1 (Singer 1982; Cordaux and Batzer 2009). L1's youngest copies are the only active LINE transposons in our genomes (Penzkofer et al. 2017; Feusier et al. 2019). L1s facilitate activity of Short Interspersed Elements (SINEs) (Goodier and Kazazian 2008; Meyer et al. 2016; Scott and Devine 2017). Moreover, the L1 transposition machinery can be utilized

by noncoding and messenger RNAs and thus contributes to generating processed pseudogenes (Konkel et al. 2010; Beck et al. 2011). Altogether, L1-related transposition is thought to give rise to ~69% of the modern human genome (de Koning et al. 2011; Sotero-Caio et al. 2017). Therefore, studying L1 transposition dynamics should substantially advance our understanding of the evolution of genome structure.

L1 transposition follows a “copy-and-paste” mechanism (Kazazian and Moran 1998; Elbarbary et al. 2016). Full-

length human L1 elements are usually >6 kb long, yet the majority of L1s in the genome have experienced 5' truncations, inversions, or point mutations within their open reading frames, and thus became inactive (Ostertag and Kazazian 2001a; Beck et al. 2011). Recent advances in whole-genome sequencing (WGS) have enabled detection of L1 elements that are polymorphic among human populations and individuals (Ratcliffe et al. 2002; Konkel et al. 2007; Ewing and Kazazian 2011), and an increase in the number of identified human L1 elements has facilitated studies of L1 evolution and transposition mechanisms (Moran et al. 1996; Kazazian and Moran 1998; Ostertag and Kazazian 2001b; St. Laurent et al. 2010; Richardson et al. 2017). Meanwhile, WGS and transposon capture sequencing in human and other model organisms (e.g. mice) have revealed heritable L1 insertions in both the germline and early embryogenesis, suggesting their contribution to genomic diversification (Feusier et al. 2019). Moreover, it has been reported that de novo insertions of L1s and dysregulation of L1s (both polymorphic and fixed ones) in the human genome can lead to a variety of diseases including cancer (Goodier and Kazazian 2008; Belancio et al. 2009; Beck et al. 2011; Payer and Burns 2019) suggesting an important impact of L1 transposition on human health.

Previous studies have investigated the chromosomal distribution of L1 elements with respect to several genomic features. For instance, densities of fixed L1 elements of different evolutionary ages were found to vary by chromosome, and to be affected by local nucleotide composition and recombination rate (Graham and Boissinot 2006). It has also been reported that younger human L1s are abundant in AT-rich regions with low gene density (Boissinot 2004). Recent studies of de novo L1 integrations in cultured human cells have suggested a strong correlation between L1 insertion preferences and DNA replication (Sultana et al. 2019), whereas the distribution of recently inserted elements was found to be influenced by chromatin state (Singer 1982; Sultana et al. 2017, 2019). These findings imply that, whereas L1 activities shape the structure of the human genome, the genomic landscape may at least partially determine the dynamics of L1 transposition over the course of evolution (Beauregard et al. 2008). In agreement with this notion, L1 transposition was found to be affected by a wide range of molecular and cellular processes. For instance, such genes as *MORC2* and *p53* can restrain L1 activity through selective transcriptional silencing (Liu et al. 2018) and post-translational regulation via the piRNA (Piwi-interacting RNA) pathway (Wylie et al. 2016). However, to date, the genome-wide dynamics of human L1 transposition has not been studied within an evolutionary framework, through which the insertion and fixation preferences of these elements can be elucidated.

In addition to providing information on evolutionary processes in the genome, a detailed understanding of L1 transpositional activity and integration preferences can facilitate the use of L1s as a mutagenesis tool in molecular genetic studies. Indeed, L1 retrotransposition provides a powerful platform for mutagenesis screens with successful applications in mammalian systems—including mouse and human cells

(An et al. 2006). There are many advantages to using L1 retrotransposons as a mutagenesis tool; for instance, they provide stable donor copies and enable RNA-level manipulation (Ivics et al. 2009). Knowing what genomic landscape may attract L1 insertions, one can engineer L1s to target-specific locations and to avoid genomic regions prone to structural rearrangements (Graham and Boissinot 2006).

With the development of multiple high-throughput experimental approaches (e.g. ChIP-seq, DNA footprinting, and bisulfite sequencing), genomic landscape features can be investigated at increasingly high resolution (Hesselberth et al. 2009; Krueger et al. 2012; Landt et al. 2012) and can provide critical information for studying L1 integration and fixation dynamics. In particular, genomic landscape measurements in consecutive subregions can be treated as “curves” along each chromosome. On one hand, this enables comparisons of landscape features among different genomic regions, revealing not only the presence, but also the location and scale of significant differences. On the other hand, this allows one to take into account the ordered nature of the measurements, hence gaining power in characterizing differences. We can analyze genomic features as curves using Functional Data Analysis (FDA) (Ramsay and Silverman 2007), a branch of statistics specifically developed to study data described as curves (mathematical functions), which was only recently introduced into genomics research (Zhang et al. 2014; Campos-Sánchez et al. 2016; Cremona et al. 2018, 2019; Guiblet et al. 2018).

In this study, we applied FDA to the genome-wide analysis of L1 transposition dynamics, considering three genome-wide data sets of human L1s representing newly integrated, polymorphic, and human-specific L1s, together with 49 genomic landscape features collated from other studies. To the best of our knowledge, we performed the first genome-wide analysis of L1 transposition dynamics in an evolutionary framework and using FDA to leverage an extensive list of genomic landscape features at high resolution. We demonstrated that the genomic distribution of human L1 elements is not random and is strongly associated with the local genomic landscape. Our analyses revealed potential mechanisms through which local genomic features have influenced L1 transposition dynamics and, in turn, L1 transposition has shaped the genomic landscape over the course of evolution.

Results

L1 Data Sets

To investigate the relationship between L1 distribution and local genomic landscape in an evolutionary framework, we considered three data sets comprising integrations of L1 elements at different evolutionary time points; namely, de novo, polymorphic, and human-specific L1s (supplementary table S1, Supplementary Material online). De novo L1s experienced minimal selection. Human-specific L1s could have been subject to selection for millions of years. Polymorphic L1s experienced levels of selection somewhere between those of de novo and human-specific L1s. Thus, studying de novo L1s should inform integration preferences, contrasting

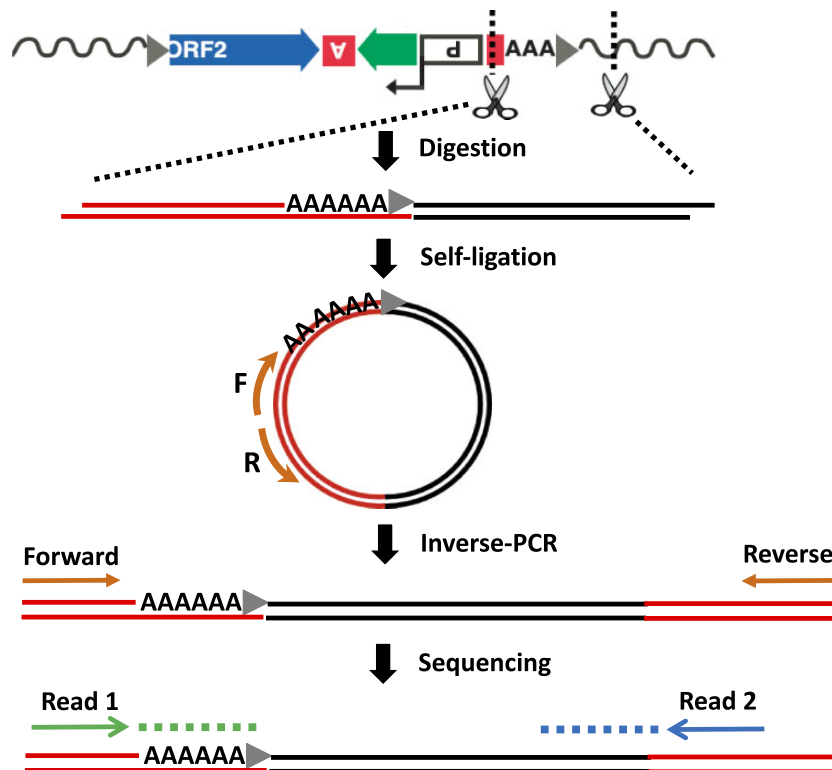


Fig. 1. Identification of in vivo de novo L1 insertions by inverse PCR. Vectors containing both a synthetic human L1 element (full-length synthetic ORFeus-Hs, see Materials and Methods) and GFP were transfected into cultured cells. The vectors were marked by two restriction enzyme sites (*MspI* and *TaqI*) and 14 different barcodes of four to six nucleotides. Although the successful de novo L1 integration events are captured by GFP expression, the genomic DNA along with a stretch of the L1 element (its poly-A tail end) is obtained by restriction enzyme digestion. The positions of L1 insertions are acquired by inverse PCR and paired-end Illumina sequencing.

distributions of human-specific versus de novo L1s should highlight fixation preferences, and investigating polymorphic L1s might provide additional insights on the interplay between integration and fixation.

For de novo L1s, we harvested L1s from an induced L1 insertion experiment conducted in the cultured human kidney stem cell line HEK-293T (fig. 1 and supplementary fig. S1, Supplementary Material online), which allows efficient vector amplification and high levels of expression with transient transfection (Rio et al. 1985; Lin et al. 2014). Positions of L1 insertions were captured by inverse PCR followed by Illumina sequencing (see Materials and Methods section). By analyzing sequencing data from this experiment, we identified 17,037 de novo L1 insertions. To the best of our knowledge, this is one of the largest collections of de novo L1 insertions in human cells. Next, we obtained 1,012 polymorphic L1s from a cross-referenced study of human polymorphic L1s (Ewing and Kazazian 2011)—the ones present in some but not all human genomes examined. The polymorphic L1 data set we have chosen for our analysis (Ewing and Kazazian 2011) is well-balanced in terms of sample size (1,012 polymorphic L1s) and population representation (310 individuals from 13 populations), while also reflecting insertion rates and allele frequency spectra similar to those in other studies of polymorphic L1s (Stewart et al. 2011; Yu et al. 2017) (supplementary table S5, Supplementary Material online). Finally, we obtained 1,205 human-specific L1HSs using the

RepeatMasker (Smit et al. 2015) track of GRCh37/hg19 from the UCSC Genome Browser (Karolchik et al. 2004) and performing the following filtering: we conservatively selected only those L1HSs that were absent from the genomes of nonhuman great apes (Boissinot et al. 2000; Ovchinnikov et al. 2002; Philippe et al. 2016) and were not annotated as polymorphic in (Ewing and Kazazian 2011).

L1 Elements Are Not Randomly Distributed

To assess whether L1 elements are randomly distributed across the genome, we analyzed their positions and the distances between subsequent L1s within and between our three data sets. Karyotype plots (supplementary fig. S2, Supplementary Material online) and chromosome-specific element densities (supplementary fig. S3, Supplementary Material online) did not suggest any obvious enrichment or depletion of de novo, polymorphic, or human-specific L1s on specific chromosomes, in agreement with previous studies (Sultana et al. 2019). However, within each of the three L1 data sets considered, the distribution of distances between L1 elements was far from random (fig. 2A and supplementary fig. S4, Supplementary Material online). In particular, L1 elements from the same data set were closer to each other compared with random expectation ($P = 10^{-16}$ for de novo L1s, $P = 1.5 \times 10^{-5}$ for polymorphic L1s, and $P = 9.7 \times 10^{-11}$ for human-specific L1s, Kolmogorov–Smirnov test; see Materials and Methods section). Furthermore, the analysis

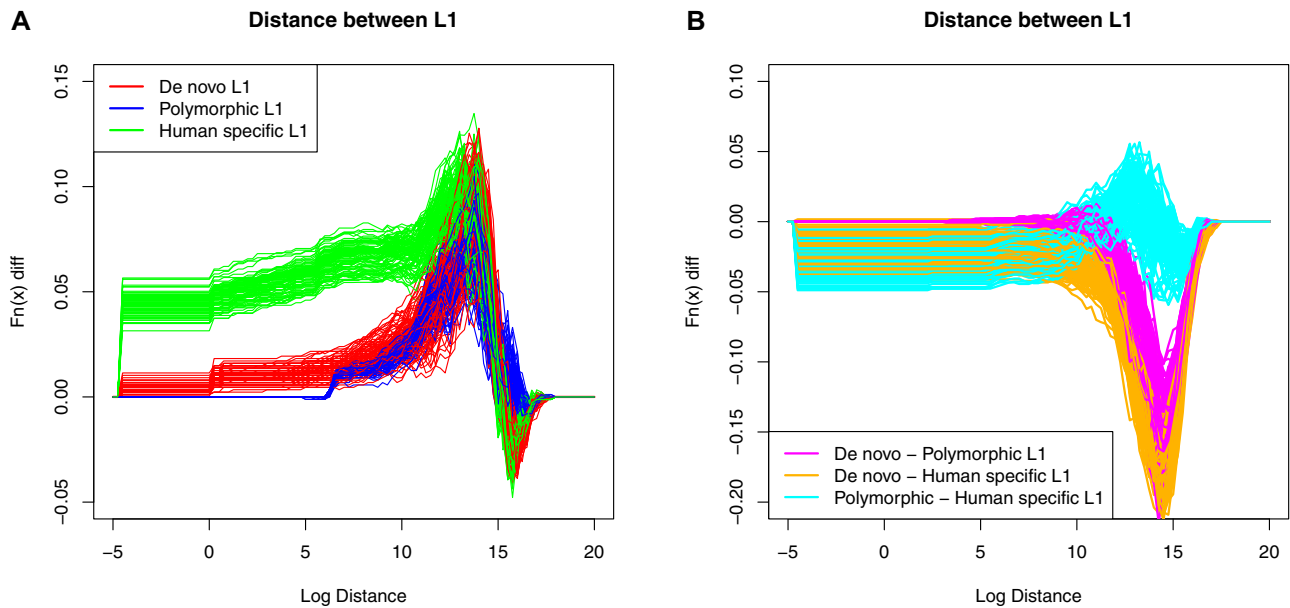


Fig. 2. Distribution of distances between L1 elements. (A) Differences between observed and expected cumulative distributions of the distances between L1 elements of the same type (de novo, polymorphic, or human-specific). (B) Differences between observed and expected cumulative distributions of the distances between L1 elements of different types. Each line shows results based on a random sample of 900 L1s of each type (100 random samples in total). Distances are reported on a log scale. Positive differences indicate smaller distance between L1s compared with random expectation, negative differences indicate larger distance compared with random expectation.

of distances between L1s from different data sets (fig. 2B) revealed distinct patterns for de novo, polymorphic, and human-specific L1s. In particular, de novo L1s were generally located further than expected from the other two types of L1s (fig. 2B and supplementary fig. S5, Supplementary Material online). Notably, the distribution of de novo L1 insertions appeared nonrandom also when considering de novo L1 data sets generated in other recent studies (Flasch et al. 2019; Sultana et al. 2019) (supplementary table S6 and fig. S16, Supplementary Material online).

Genomic Landscape Features Analyzed

To understand the determinants of the (nonrandom) distributions observed for L1s along the genome, we quantitated the genomic landscape surrounding L1 elements and studied its association with L1 integration and fixation. Specifically, using publicly available sources (e.g. ENCODE [ENCODE Project Consortium 2012] and UCSC Genome Browser [Karolchik et al. 2004]) and results from previous studies (see Materials and Methods section), we collected data on 49 quantitative genomic features that may influence L1 integration and fixation dynamics (table 1 and supplementary table S2, Supplementary Material online). These included features related to chromatin structure, transcription regulation, DNA methylation, nucleotide composition, non-B DNA structures, non-L1 transposons, gene expression in human embryonic stem cells (hESCs), replication, recombination, and selection. In general, we strived to be consistent regarding the sources of genomic features, which was an important component in our study design. Specifically, 22 features (e.g. GC content, exon coverage, and most conserved elements) were not cell-line specific, and we extracted most of the other

features (e.g. histone modifications and DNA methylation) from hESCs. Since our de novo L1 data set was generated in HEK-293T cells, we also examined epigenetic features available for hg19 in the HEK-293T cell line (or in HEK-293 when not available in HEK-293T), and compared them with the same features generated in hESC lines. The results indicated substantial genome-wide correlation between the features from HEK-293T (or HEK-293) and hESC (supplementary table S4 and fig. S15, Supplementary Material online). Therefore, the genomic feature data sets employed in our study are generally representative of the genomic landscape of HEK-293T cells.

We constructed 100-kb flanking genomic regions surrounding each L1 insertion (± 50 kb), as well as 10,037 100-kb control regions with minimal L1 element coverage ($<7\%$; fig. 3; see Materials and Methods section). We excluded regions overlapping with unsequenced gaps (Kent et al. 2002) and repeats with artifactual ChIP-seq or DNase-seq signals (supplementary table S3, Supplementary Material online; Materials and Methods section), as well as sex chromosomes, given the lack of genomic feature data available for them. Forty-four features were measured at 1-kb resolution ("high-resolution features"), providing 100 measurements per L1-flanking (or control) region. Five additional features—telomere hexamers, distance to the telomere, distance to the centromere, replication timing profile, and sex-averaged recombination rate—were measured at 100-kb resolution ("low-resolution features"), providing a single measurement per L1-flanking (or control) region. Features were extracted as coverage (percentage of the window covered by a feature), average value weighted by window coverage ("weighted average"), count, or average signal, per 1-kb window (or per

Table 1. Genomic landscape features and their contributions in single and mFLRs.

Group	Name	Format	Resolution	Source	De Novo L1 vs. Control		Human-Specific L1 vs. De Novo L1	
					Pseudo-R ² for sFLR (%)	RCDE for mFLR (%)	Pseudo-R ² for sFLR (%)	RCDE for mFLR (%)
Chromatin	DNase hyper. sites	Signals	High	ENCODE	1.00	5.03	18.12	1.89
Chromatin	RNA Pol II	Coverage	High	Barski et al. (2007)	0.23	1.59	5.72	Not sel.
Chromatin	CTCF	Signals	High	ENCODE	Not sign.	Not sign.	13.22	Not sel.
Transcription	H3K4me2	Signals	High	ENCODE	2.51	Not sel.	15.56	0.70
Transcription	H3K9ac	Signals	High	ENCODE	2.38	1.10	15.64	Not sel.
Transcription	H3K4me3	Signals	High	ENCODE	1.48	1.42	11.09	Not sel.
Transcription	H3K79me2	Signals	High	ENCODE	Not sign.	Not sign.	4.10	Not sel.
Transcription	H3K27ac	Signals	High	ENCODE	2.69	Not sel.	12.62	Not sel.
Transcription	H4K20me1	Signals	High	ENCODE	1.21	Not sel.	9.57	Not sel.
Transcription	H3K4me1	Signals	High	ENCODE	4.20	0.93	12.32	2.64
Transcription	H3K36me3	Signals	High	ENCODE	1.48	0.71	7.55	Not sel.
Transcription	H3K9me3	Signals	High	ENCODE	Not sign.	Not sign.	1.50	4.46
Transcription	H3K27me3	Signals	High	ENCODE	0.76	Not sel.	9.18	Not sel.
Transcription	H2AFZ	Signals	High	ENCODE	0.54	Not sel.	1.91	Not sel.
Transcription	Gene expression	W. aver.	High	UCSC Genome Browser	1.19	Not sel.	3.84	Not sel.
DNA methylation	Sperm hypometh	Count	High	Molaro et al. (2011)	2.04	1.34	3.22	2.12
DNA methylation	CpG methylation	W. aver.	High	Lister et al. (2009)	0.27	Not sel.	0.32	Not sel.
DNA methylation	5-hMc	Count	High	Szulwach et al. (2011)	Not sign.	Not sign.	11.28	Not sel.
DNA methylation	CHH methylation	W. aver.	High	Lister et al. (2009)	Not sign.	Not sign.	Not sign.	Not sign.
DNA methylation	CHG methylation	W. aver.	High	Lister et al. (2009)	Not sign.	Not sign.	1.95	Not sel.
Non-B DNA	G-quadruplex	Coverage	High	Cer et al. (2011)	1.16	1.64	9.43	Not sel.
Non-B DNA	A-phased repeats	Coverage	High	Cer et al. (2011)	Not sign.	Not sign.	9.29	Not sel.
Non-B DNA	Direct repeats	Coverage	High	Cer et al. (2011)	0.44	Not sel.	4.78	Not sel.
Non-B DNA	Inverted repeats	Coverage	High	Cer et al. (2011)	Not sign.	Not sign.	1.78	Not sel.
Non-B DNA	Mirror repeats	Coverage	High	Cer et al. (2011)	2.18	Not sel.	0.31	Not sel.
Non-B DNA	Z DNA motifs	Coverage	High	Cer et al. (2011)	Not sign.	Not sel.	2.30	Not sel.
Microsatellites	Mononucl. microsat	Coverage	High	Genome screening	0.16	Not sel.	1.73	Not sel.
Microsatellites	Di-, tri-, and tetranucl.	Coverage	High	Genome screening	0.22	Not sel.	Not sign.	Not sign.
Nucl. composition	GC content	Percent	High	Genome screening	1.23	13.99	17.13	1.92
L1 target motifs	L1 target motifs	Count	High	Genome screening	4.43	16.22	3.75	Not sel.
Other TEs	Alu	Coverage	High	Genome screening	0.81	3.72	12.61	5.44
Other TEs	MIR	Coverage	High	UCSC Genome Browser	6.56	Not sel.	1.23	Not sel.
Other TEs	L2 and L3	Coverage	High	UCSC Genome Browser	4.94	8.52	Not sign.	Not sign.
Other TEs	DNA Transposons	Coverage	High	UCSC Genome Browser	Not sign.	Not sign.	Not sign.	Not sign.
Other TEs	LTR elements	Coverage	High	UCSC Genome Browser	0.67	Not sel.	3.38	Not sel.
Replication	Replication origins	Count	High	Besnard et al. (2012)	0.45	0.95	11.75	Not sel.
Recombination	Recomb. hotspots	Count	Low	Myers et al. (2008)	0.05	Not sel.	1.25	Not sel.
Selection	Most cons. elements	Coverage	High	UCSC Genome Browser	8.74	2.17	3.45	Not sel.
Selection	CpG islands	Coverage	High	UCSC Genome Browser	2.54	4.04	13.68	2.75
Selection	Exons	Coverage	High	UCSC Genome Browser	0.54	1.21	8.98	2.77
Selection	Introns	Coverage	High	UCSC Genome Browser	2.65	Not sel.	1.08	Not sel.
Chr. location	Dist. to centromere	Distance	Low	Genome screening	Not sign.	Not sign.	0.09	Not sel.
Chr. location	Distance to telomere	Distance	Low	Genome screening	Not sign.	Not sign.	1.54	Not sel.
Chr. location	Telomere hexamer	Count	Low	Phohl et al. (2002)	9.96	0.85	1.00	Not sel.
Replication	Replication timing	W. aver.	Low	Ryba et al. (2010)	0.33	3.47	11.74	Not sel.
Recombination	Recombination rate	W. aver.	Low	Kong et al. (2010)	Not sign.	Not sign.	Not sign.	Not sign.
Total pseudo-R ²						31.97		26.97

NOTE.—Testis gene expression (Brawand et al. 2011), exon expression, and transcript expression (UCSC Genome Browser) were excluded from the analysis due to their high correlations with other features (supplementary fig. S6, Supplementary Material online).

Chromatin: chromatin structure; "Not sel.", features that were not selected in the final mFLR models (potentially due to interdependencies among features, see text); "Not sign.", features that showed no significant differences in IWTomics tests (see text); Transcription: transcription regulation and gene expression; "W. aver.", weighted average.

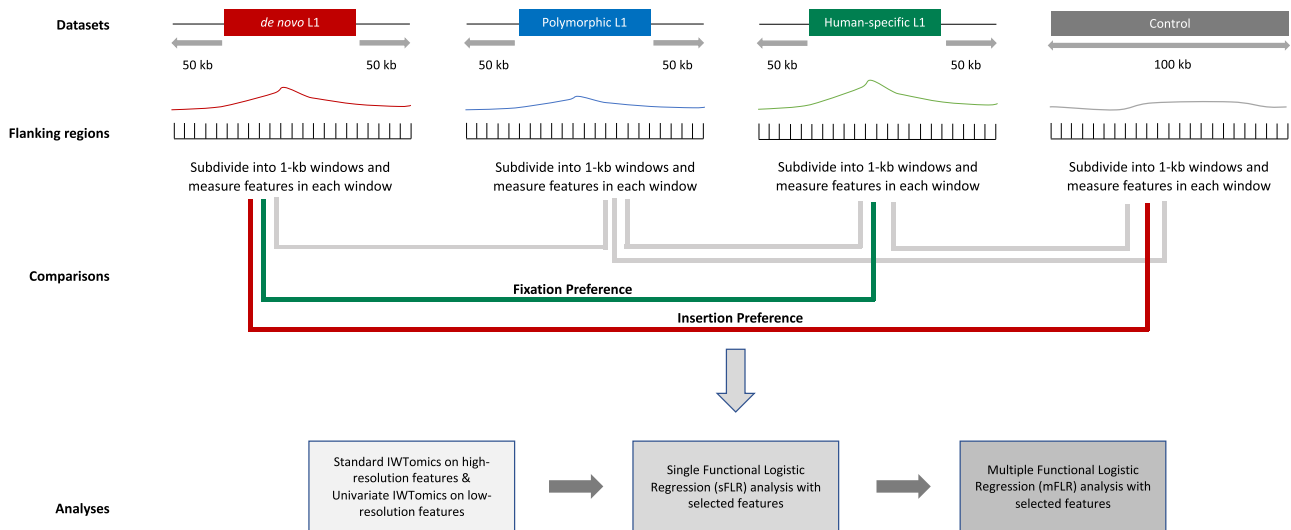


Fig. 3. FDA workflow. Illustration of the FDA workflow used in the study. The 100-kb L1 regions were constructed taking 50-kb in each direction of the insertion sites, and the control regions were constructed as 100-kb nonoverlapping intervals with low coverage ($<7\%$) of L1s. High-resolution genomic features were measured within each 1-kb window of the 100-kb regions, and treated as functional data (i.e. curves) for FDAs. Curves in different groups (different types of L1s, or each L1 type vs. controls) were then compared using IWTomics and FLR. The control regions in this study contain less than 7% coverage by all annotated L1 elements.

100-kb in the case of low-resolution features) in each L1-flanking (or control) region. Then, for each feature, values were averaged across all L1 elements belonging to the same data set, producing 100 mean values and thus mean curves (or a single mean value in the case of low-resolution features). Three high-resolution features were highly correlated (Spearman's correlation coefficient >0.8) with other features (supplementary fig. S6, Supplementary Material online) and were excluded from subsequent analyses. Thus, a total of 41 high-resolution and five low-resolution features were retained.

Functional Data Analysis

To capture multiscale (up to 100 kb) differences in local genomic landscape features among L1s from the three data sets and control regions, we utilized four FDA approaches. *First*, to identify differences in low-resolution features between L1s from the three data sets and control regions, we used the univariate version of Interval-Wise Testing for omics data (IWTomics) (Cremona et al. 2018). Considering the low-resolution features one at a time, the test focuses on a mean value for every 100-kb region and evaluates the difference in means between two sets of L1-flanks, or one set of L1-flanks and a set of controls. We compared low-resolution features between de novo L1s and controls; human-specific L1s and de novo L1s; human-specific L1s and controls; polymorphic L1s and controls; polymorphic L1s and de novo L1s; and finally, human-specific L1s and polymorphic L1s (a total of six comparisons). *Second*, to investigate differences in high-resolution features, we used IWTomics in its standard (i.e. functional) version, running the same six comparisons for each feature (again one at a time). Standard IWTomics allows one to contrast two sets of curves composed of contiguous values. In our case, we tested for differences between curves composed of 100 mean values (one per 1-kb window) for

each genomic feature, for the three L1 data set and the controls (the same six comparisons). *Third*, to quantify the impact of each specific feature (independent of the effects of other features) on distributions of L1s at different evolutionary time points, we ran *single* Functional Logistic Regressions (sFLRs) (Ramsay and Silverman 2007; Febrero Bande and Oviedo de la Fuente 2012), using the low- and high-resolution features that were significant according to IWTomics test and the same six comparisons (fig. 3). The discriminatory strength of each feature was quantified with pseudo- R^2 s from these sFLRs. *Fourth*, to quantify joint effects of multiple features, many of which can interact and are correlated according to our clustering analysis (supplementary fig. S6, Supplementary Material online), we built multiple Functional Logistic Regressions (mFLRs), again using the same genomic landscape feature data and the same six comparisons. mFLRs take into account multiple features at a time. For each pairwise comparison of L1 flanks and controls, we identified a subset of relevant features among the (low- and high-resolution) ones that were significant according to IWTomics, using a functional variable selection method based on group lasso (Meier et al. 2008; Matsui 2014), and then ran the corresponding mFLR with this subset. The mFLR provided quantification of the total impact (total deviance explained by the selected features taken together), as well as the impact of each individual feature (Relative Contribution to the Deviance Explained, or RCDE) when considered with others (table 1 and supplementary table S2, Supplementary Material online). Notably, due to the functional (i.e. curve) nature of the data, neither sFLR nor mFLR provides a sign for the effect of each feature on the differences between L1 flanks and/or controls (effect estimates are themselves curves). However, this information can be retrieved from the IWTomics analysis.

Here, we present results (for all four FDA approaches) from comparisons of de novo L1 flanks versus controls (fig. 4A and

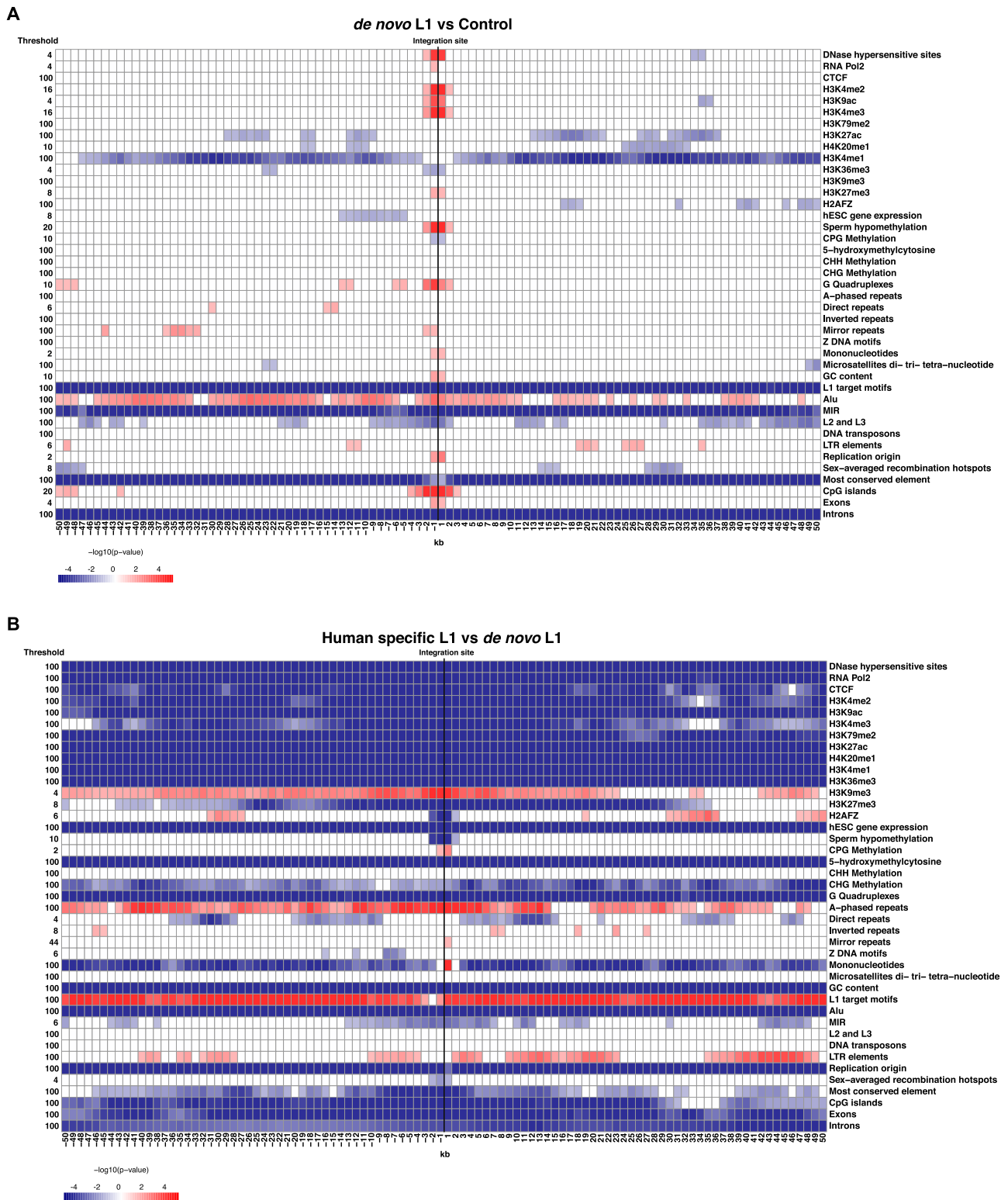


FIG. 4. Summary of IWTomics results for individual high-resolution features. (A) De novo L1 flanking regions versus control regions. (B) Human-specific L1 versus de novo L1 flanking regions. The X-axis represents the position analyzed within the 100-kb flanking regions of L1 elements (or 100-kb control regions); each unit is a 1-kb window. The black vertical line across the center marks the insertion site. Each row represents one genomic feature and reports the adjusted P value curve on a \log_{10} scale. White: nonsignificant difference (P value > 0.05). Red: significant difference, with overrepresentation of the feature. Blue: significant difference, with underrepresentation of the feature. The selected scale thresholds corresponding to the adjusted P value curves are noted on the left (column “Threshold”). The control regions contain less than 7% coverage by all annotated L1 elements.

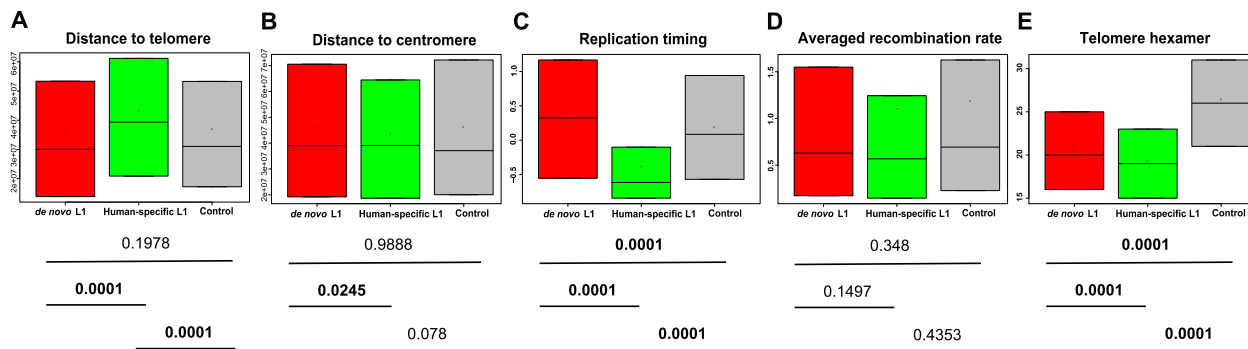


Fig. 5. Summary of IWTOmics results for individual low-resolution features. (A) Distance to the telomere. (B) Distance to the centromere. (C) Replication timing. (D) Sex-averaged recombination rate. (E) Count of telomere hexamers. Each panel presents the boxplots of the feature in the flanking regions of de novo and human-specific L1s and in control regions. Black dot: mean; bold horizontal line: median; box limits: 25th and 75th percentiles (whiskers and outliers not shown). The *P* values of pairwise IWTOmics tests are noted at the bottom; significant ones (*P* value < 0.05) are in bold. An extended summary comprising also the flanking regions of polymorphic L1s is provided in [supplementary figure S9, Supplementary Material](#) online. The control regions contain less than 7% coverage by all annotated L1 elements.

[supplementary fig. S7, Supplementary Material](#) online; [table 1](#)) and of human-specific versus de novo L1 flanks ([fig. 4B](#) and [supplementary fig. S7, Supplementary Material](#) online; [table 1](#)). They should reflect, respectively, L1 integration and selection preferences with respect to different genomic features—and are thus particularly informative. Results for the other four comparisons are included in the Supplement ([supplementary figs. S8–S10 and table S2, Supplementary Material](#) online).

De Novo L1 Insertion Landscape

To investigate insertion preferences, we compared genomic features in the flanks of de novo L1s versus control regions. The univariate IWTOmics analysis ([fig. 5](#)) contrasting low-resolution features suggested that de novo L1 insertions are significantly and positively associated with early replication timing ($P = 0.0001$; [fig. 5C](#)), and significantly and negatively associated with telomere hexamers ($P = 0.0001$; [fig. 5E](#)).

The standard (functional) IWTOmics analysis revealed 17 high-resolution genomic features that were significantly overrepresented at de novo L1 flanks, suggesting their positive association with L1 insertions ([fig. 4A](#) and [supplementary fig. S7A and B, Supplementary Material](#) online). Among these features, 13 had highly localized signals centered at the L1 integration site. These included seven features with particularly strong overrepresentation at the L1 integration site: DNase hypersensitive sites (DHS), H3K4me2, H3K4me3, and H3K9ac histone marks, sperm hypomethylation, CpG islands, and G-quadruplexes. In contrast, *Alu* density was significantly overrepresented across almost the entire 100-kb flanks of de novo L1s ([fig. 4A](#)). In addition, IWTOmics identified 12 high-resolution features with underrepresented signals at de novo L1 flanks, suggestive of their negative influence on L1 insertion preferences ([fig. 4A](#)). Among them, H3K36me3 histone marks and CpG methylation had underrepresented signals localized at the L1 integration site, whereas most conserved elements, introns, MIRs, and L1 target sites were significantly underrepresented across the entire de novo L1 flanks analyzed. Interestingly, H3K4me1 histone marks were significantly

underrepresented starting at ± 2 kb from L1 integration sites, but not closer to them ([fig. 4A](#)).

The sFLR models estimated the strength of each genomic feature (not considering other features) in explaining de novo L1 integration preferences ([table 1](#)). Most conserved elements, MIRs, and telomere hexamer were the strongest predictors, each explaining deviance above 5% (pseudo- $R^2 = 8.74\%$, pseudo- $R^2 = 6.56\%$, and pseudo- $R^2 = 9.96\%$ respectively). Other strong predictors were H3K4me1 histone marks, L1 target sites, and L2 and L3 (pseudo- $R^2 = 4.20\%$, pseudo- $R^2 = 4.43\%$, and pseudo- $R^2 = 4.94\%$, respectively).

The mFLR model comparing de novo L1 flanks with controls selected 18 genomic features ([table 1](#)). Taken together, these features explained 31.97% of the total deviance. Based on their relative contributions (here evaluated in the context of the mFLR), several features had a particularly strong effect (RCDE > 5%) on L1 integration preferences ([table 1](#)), including L1 target sites (RCDE = 16.2%), GC content (RCDE = 14.0%), and DHS (RCDE = 5.03%).

L1 Fixation Landscape

To investigate fixation preferences, we compared the distribution of genomic features in the flanks of human-specific versus de novo L1s. The univariate IWTOmics analysis contrasting low-resolution features ([fig. 5](#)) suggested that L1 fixation is significantly and negatively associated with early replication timing ($P = 0.0001$), telomere hexamers ($P = 0.0001$), and distance to centromere ($P = 0.0245$).

The standard (functional) IWTOmics ([fig. 4B](#)) identified six high-resolution features that were significantly overrepresented at human-specific L1 flanks versus those of de novo L1s. These included three features that were overrepresented over most of the 100-kb flanks analyzed—H3K9me3 histone marks, A-phased repeats, and L1 target motifs; two features that had localized overrepresentation at the L1 integration site—CpG methylation (stronger effect) and mirror repeats (weaker effect); and LTR elements that displayed a “patchy” overrepresentation. IWTOmics also identified as many as 27 features that were underrepresented at human-specific L1-

flanks versus those of de novo L1s (fig. 4B), suggesting that the regions might undergo selection against L1 fixation, and thus lack fixed L1 elements. Although most of them were underrepresented over the entire 100-kb flank length, H2AZF histone marks, sperm hypomethylation, and sex-averaged recombination hotspots were underrepresented only in the vicinity of the L1 integration site. Interestingly, mononucleotide microsatellites were enriched close to the integration site but underrepresented along the remainder of the flanks (fig. 4B), suggesting distinct associations of this feature at different scales.

Also here, the sFLR models allowed us to evaluate the strength of each genomic feature in explaining de novo L1 fixation preferences. Features such as DHS, GC content, and H3K9ac and H3K4me2 histone marks had strong effects, each explaining more than 15% of the deviance (table 1). The next tier of predictors each explained 10–15% of deviance and included CpG islands, CTCF, H3K27ac, H3K4me1 and H3K4me3 histone marks, *Alus*, replication origins, replication timing profile, and 5hMC methylation. Several other predictors each explained 5–10% of deviance. These included H3K4me1, H3K27me3 and H3K36me3 histone marks, G-quadruplexes, A-phased repeats, exons, and RNA Pol II.

The mFLR model comparing human-specific and de novo L1 flanks selected nine predictors and explained 26.97% of the deviance (table 1). Among the strongest predictors (with RCDE >2%) were *Alus* (RCDE = 5.44%), H3K9me3 (RCDE = 4.46%) and H3K3me1 (RCDE = 2.64%) histone marks, exons (RCDE = 2.77%), CpG islands (RCDE = 2.75%), and sperm hypomethylation (RCDE = 2.12%).

Discussion

Our analysis of 49 genomic landscape with FDA suggested that de novo, polymorphic, and human-specific L1s in the human genome are characterized by unique genomic landscapes, with different features exhibiting associations at specific locations and scales. In general, de novo L1 integrations tend to occur in regions with open chromatin structure, elevated transcriptional activities, and high GC content (fig. 4A). In contrast, after accounting for their integration preferences, human-specific L1s tend to concentrate in regions with relatively low exon content, enriched transcriptional repression marks and conserved elements (fig. 4B). The genomic landscape for polymorphic L1s is generally similar to that of human-specific L1s, yet their comparison with control suggests less significant, weaker associations (supplementary fig. S8A and table S2, Supplementary Material online). This is consistent with our results showing that, in the genome, polymorphic L1s are located closer to human-specific than de novo L1s (fig. 2). Below we discuss the results from our analyses, and relate the L1 transposition dynamics with different biological processes represented by genomic landscape features.

Biological Processes and Features Associated with L1 Integration and Fixation

Chromatin Structure

Our results suggest that L1 integration and fixation are associated with open and condensed chromatin structure,

respectively. Three chromatin structure features were considered in our analysis: 1) DHSs, which are open chromatin regions accessible to *trans*-factors and other regulatory elements (Wallrath et al. 1994; Tsompana and Buck 2014); 2) RNA Pol II-binding sites, which are positively correlated with open chromatin structure and gene expression (Barski et al. 2007; Kines and Belancio 2012; Sun et al. 2015); and 3) CTCF motifs, which facilitate interactions between transcription regulatory sequences and are hypothesized to facilitate boundaries between topologically associated domains (TADs) (Kim et al. 2007; Schmidt et al. 2012; Ong and Corces 2014; Ghirlando and Felsenfeld 2016). We found that DHS and RNA Pol II sites were enriched at integration sites of de novo L1s (fig. 4A and supplementary fig. S7A, Supplementary Material online), with relatively weak signals identified in sFLRs, but stronger signals in the mFLR (table 1; CTCF was not significant in any of our analyses). Thus, chromatin structure features may play an important role in L1 integration, even when considered in the context of other genomic features. In contrast, DHS, RNA Pol II, and CTCF sites were underrepresented over the whole 100 kb surrounding L1s in the comparison of human-specific versus de novo elements (fig. 4B). These effects were strong in sFLRs (all three predictors had pseudo- R^2 above 5%), but weaker in the mFLR (only DHSs were selected; table 1), suggesting that effects of chromatin features might be partially masked by other features included in this model. We hypothesize that open chromatin structure can provide better accessibility for the L1 integration machinery, in line with other studies (Cost and Boeke 1998; Sultana et al. 2019). In contrast, L1 elements that inserted into genome regions with condensed chromatin structure are more likely to become fixed, likely due to the lack of regulatory units and lower transcription output in these regions of the genome (ENCODE Project Consortium 2012; Ward and Kellis 2012).

Transcriptional Regulation and Gene Expression

Our investigation of 11 epigenetic marks from ENCODE (ENCODE Project Consortium 2012) and gene expression profiles in hESCs (gene expression) (Karolchik et al. 2004) indicated a strong correlation between transcriptional regulation and L1 transposition dynamics. Epigenetic marks of active transcription landscape (Zhou et al. 2011; EpiGenie Epigenetics Background, Tools and Database 2020)—H3K4me2 (active promoters), H3K9ac (transcription activation; transition between transcription initiation and elongation) (Gates et al. 2017), and H3K4me3 (transcriptional elongation)—were all overrepresented specifically at the insertion sites of de novo L1s (fig. 4A and supplementary fig. S7B, Supplementary Material online). The associations of these features with L1 integration were confirmed by their significance in both sFLR and mFLR models (except for H3K4me2, which was not selected in the mFLR). This suggests a localized positive effect (at the scale of several kilobases) of active transcriptional activities on L1 insertion.

In contrast, a comparison of the landscape between human-specific and de novo L1s revealed significantly

decreased hESC gene expression levels, as well as underrepresented histone marks of active transcription (H3K4me2, H3K9ac, H3K4me3, K79me2), elevated transcription activities (H3K27ac, H3K20me1, H3K4me1), and open chromatin (H3K36me3) over the whole 100-kb L1 flanking region (fig. 4B). Moreover, the transcription repression mark H3K9me3 was significantly overrepresented over most of the 100-kb region, and this overrepresentation was particularly strong within ± 8 kb from L1 insertion site (fig. 4B and supplementary fig. S7C, Supplementary Material online). However, the transcription shutdown mark H3K27me3, which is also linked to high-CpG promoters (Zhou et al. 2011) due to “bivalent domains”, was underrepresented over the whole 100-kb region. The heterochromatin mark H2AFZ (Rangasamy et al. 2004; Nishida et al. 2005) was underrepresented in the immediate vicinity of integration sites comparing human-specific versus de novo L1s. sFLR models indicated particularly strong effects of hESC gene expression and of active transcription marks, but few histone marks were selected in the mFLR model, highlighting their interdependencies with other genomic features.

In summary, our results suggest that de novo L1 insertions are facilitated by active transcription marks, whereas human-specific L1s are fixed in nonheterochromatic regions—where transcription is inactive or repressed and levels of gene expression are low, suggesting a potentially strict regulation of fixed L1s (Philippe et al. 2016). Moreover, epigenetic marks act at larger scales on L1 fixation preferences (e.g. 100 kb) and at smaller scales on L1 insertion preferences (e.g. 1–2 kb), arguing for different molecular and evolutionary mechanisms.

DNA Methylation

Our analysis revealed significant but contrasting effects of DNA methylation on L1 insertion and fixation. Five DNA methylation features were analyzed: 1) sperm hypomethylation (at CpG sites), which reflects genomic regions with low methylation levels in sperm (Molaro et al. 2011); 2) CpG methylation (in H1-hESC), which silences gene expression (Weber et al. 2007; Lister et al. 2009; Straussman et al. 2009) and limits TE transcription thus controlling their expansion in the genome (Rodriguez et al. 2008; Oliver and Greene 2009); 3) 5-hMc methylation, the first oxidative product in the active demethylation of 5-methylcytosine, which is preferentially established at CpG dinucleotides (Szulwach, Li, Li, Song, Wu, et al. 2011; Branco et al. 2012) and silences gene expression (Szulwach, Li, Li, Song, Han, et al. 2011; Mooijman et al. 2016); 4 and 5) CHH and CHG methylation, which is enriched in exons of highly expressed genes (Lister et al. 2009; He and Ecker 2015). In the immediate vicinity (± 1 kb) of de novo L1 insertions, CpG methylation was depleted, whereas sperm hypomethylation was enriched (fig. 4A); sFLRs showed a weak effect of CpG methylation, and stronger effect of sperm hypomethylation, which was also selected in the mFLR (table 1). In contrast, after subtracting the effects of de novo insertions, in the immediate vicinity of fixed L1s CpG methylation was enriched and sperm hypomethylation was depleted (fig. 4B); sperm hypomethylation had again a strong

effect according to sFLRs and was selected in the mFLR (table 1). L1 fixation preferences were also associated with underrepresented 5-hMc and CHG methylation across the whole 100-kb flanking region analyzed (fig. 4B); these two features showed strong and weak effects, respectively, in sFLRs, but were not selected in the mFLR (table 1).

We hypothesize that genomic regions with low CpG methylation (and high hypomethylation) have elevated transcription, and thus are more accessible to the L1 transposition machinery. Besides, the underrepresented CpG methylation signals both upstream and downstream of the L1 insertion site may act as barriers to prevent the expansion of L1s. In agreement with this, hypomethylation was associated with young and active L1 subfamilies in previous studies (Khan et al. 2005; Molaro et al. 2011). Regarding fixation preferences, our results point towards a paucity of fixed L1s in regions with actively expressed genes (we observe increased CpG methylation and decreased sperm hypomethylation). Moreover, L1s are usually not fixed in regions with highly expressed genes, explaining the negative association with CHG methylation. Increased CpG methylation near fixed L1s might also limit their own transcriptional activity (Zemach et al. 2010; Huang et al. 2017).

Non-B DNA Motifs and Microsatellites

Based on our results, non-B DNA motifs and microsatellites have significant associations with the insertion and fixation preferences of L1s. Specifically, we examined six types of non-B DNA: G-quadruplexes, A-phased repeats, direct repeats, inverted repeats, mirror repeats, and Z-DNA motifs—all potentially altering the DNA structure relative to the most common B form (Zhao et al. 2010; Cer et al. 2013; Sahakyan et al. 2017). We also examined coverage of mononucleotide microsatellites and combined coverage of di-, tri-, and tetranucleotide microsatellites, many of which also form non-B DNA (Guiblet et al. 2018). We found that G-quadruplexes, mirror repeats and mononucleotide microsatellites were enriched in the immediate vicinity of L1 insertion sites (fig. 4A); however, only G-quadruplexes were selected by the mFLR (table 1). In the comparison of human-specific versus de novo L1s flanks, G-quadruplexes were underrepresented, and A-phased repeats were overrepresented, over the whole 100-kb region, and mononucleotide microsatellites were enriched at the fixation site but underrepresented away from it (fig. 4B). The three features were not selected in the mFLR (table 1). G-quadruplexes, mirror repeats, and mononucleotide microsatellites might attract new L1 integrations by inducing DNA stability (Li et al. 2002; Kejnovský et al. 2013) and/or by changing chromatin structure (Li et al. 2002; Bochman et al. 2012; Lexa et al. 2014; Hou et al. 2019). The mononucleotide microsatellites enrichment observed in the immediate vicinity of L1 integration sites persisted for fixed elements. The depletion of mononucleotide microsatellites observed across the entire flanks of fixed L1s, which are enriched at poly-A tails of retrotransposed genes and TEs, could reflect gene scarcity in the broader vicinity of fixed elements. Underrepresentation of G-quadruplexes and overrepresentation of adenine-rich A-

phased repeats (Cer et al. 2011) might reflect the overall low GC content of the flanks of fixed L1s.

Nucleotide Composition and L1 Target Motifs

We found that nucleotide composition (i.e. GC content) and L1 target motifs exhibit major associations with L1 insertion and fixation preferences. Specifically, GC content was elevated in the immediate vicinity of de novo L1 insertion sites (fig. 4A) and was a strong predictor in both sFLR and mFLR comparing de novo L1 flanks with controls (table 1). In contrast, GC content was globally lower in the flanks of human-specific L1s compared with de novo L1s (fig. 4B); also here, it was a strong predictor in both sFLR and mFLR (table 1). These results are in agreement with previous findings that fixed L1 elements are usually found in AT-rich regions of the genome (Lander et al. 2001; Medstrand et al. 2002; Kvikstad and Makova 2010). We also ruled out the potential experimental bias from the two restriction enzymes *MspI* and *TaqI* used for the de novo L1 insertion assay, by analyzing the genome-wide distance distribution of *MspI* and *TaqI* sites (supplementary fig. S18, Supplementary Material online) as well as comparing their enrichment against different genomic features, including GC content (supplementary note S1, Supplementary Material online). L1 target motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAG, CTA AAA, and TCAAAA) (Feng et al. 1996; Jurka 1997; Zhao et al. 2019) were under- and overrepresented in the 100-kb regions surrounding de novo and fixed L1 elements, respectively; this feature effect was strong in both sFLRs, but was selected only in the de novo versus control mFLR.

The underrepresentation of L1 target motifs in the flanks of de novo L1s is at first sight counterintuitive. However, because its signal extends over the whole 100-kb flanking region, it might reflect the overall AT-richness of L1 target motifs, as de novo L1s prefer integrating into GC-rich regions abounding in transcribed genes. Specifically, we observed a depletion of L1 target sites in the whole 100-kb flanking regions of de novo L1s (fig. 4A), and not at smaller resolution. Thus, depletion may be largely driven by the resolution used—which we selected because it is preferable for most other genomic features. To further study the presence of L1 target site motifs near the de novo L1s at small scales along with the potential bias from the L1 insertion assay (supplementary note S2, Supplementary Material online), we also analyzed the distribution of distances between the consensus L1 target site motifs and de novo L1 elements, using both the complete de novo L1 data set and a stringently filtered subset (supplementary note S2, Supplementary Material online). The results revealed that the majority of the de novo L1s have at least one target site motif within a distance of 1 kb for both cases. This observation was also supported by contrasting L1 target motifs between the de novo L1 data sets and our L1-depleted control regions with IWTomics (supplementary note S2, Supplementary Material online). The comparison revealed consistent signals of L1 target motifs before and after filtering; in both cases the mean motif counts (per 1 kb window) in the de novo L1 regions were between 2 and

3, which does not indicate a complete depletion of L1 target site motifs. Additional potential explanations for this counterintuitive observation include 1) the suboptimal scale analyzed for L1 target motifs (they are 6-bp long, while we analyzed scales starting from 1 kb); and 2) the lack of specificity of the L1 endonuclease, as the majority of L1s were found to insert into sites that differ from the exact consensus L1 target motif (TTAAAA) (Feng et al. 1996; Cost and Boeke 1998; Boissinot 2004; Zhao et al. 2019).

Interestingly, the separate effects of L1 target motifs and GC content in the sFLRs comparing de novo L1 flanks versus controls were not particularly strong, but increased drastically when the two features were considered together in the mFLR (table 1). We hypothesize that this might be due to GC content correlating with many genomic features in the genome, including L1 target motifs (Kvikstad and Makova 2010). This was supported by our comparisons of L1 target motif counts between L1 flanking regions and controls matched for GC content. Specifically, we computed the quartiles of mean GC content considering all regions simultaneously, and plotted L1 target counts in L1 regions versus controls for each level of GC content (supplementary fig. S11, Supplementary Material online). The results revealed more prominent differences in L1 target motif counts between L1 flanks and control at GC-poor (0–25% and 25–50% quantiles) than GC-rich (higher quantiles) regions (supplementary fig. S11B–D, Supplementary Material online), suggesting interactions between GC content and L1 target motifs.

Chromosomal Location

Location on the chromosome, which we characterized considering distance to the nearest centromere, distance to the nearest telomere, and count of telomere hexamers, is also associated with integration and fixation preferences of L1s. Fixed L1s were generally located further from telomeres compared with de novo L1s, suggesting that telomeric regions are less tolerant of L1 fixation. However, telomere hexamers were significantly underrepresented in de novo L1 flanks versus controls (strong effect in sFLR, selected in mFLR), and in the flanks of fixed versus de novo L1s (weaker effect in sFLR, not selected in mFLR). This observation might be explained by the negative impact of telomere hexamers on L1 activities possibly due to the Telomere Position Effect, according to which heterochromatin is formed and gene expression is repressed near the telomeres (Pedram et al. 2006; Calado and Dumitriu 2013; Venkatesan et al. 2017). Alternatively, this observation may be due to the difficulty in mapping L1 sequences to regions close to telomeres and enriched with hexamer repeats (Pohl et al. 2002; Treangen and Salzberg 2011; Lee et al. 2014). Thus, these results should be treated with caution. We also observed that human-specific L1s are located closer to centromeres than de novo L1s (fig. 5). Although this effect was weak (table 1), pericentromeric regions have decreased GC content (Duret and Arndt 2008) and experience relaxed selection (Horvath and Slotte 2017), potentially explaining an enrichment of fixed, human-specific L1s close to centromeres.

Transposition of Other TEs

Investigating the distributions of five types of TEs—*Alus*, MIRs, L2/L3 elements, DNA transposon, and LTR elements—revealed important associations between some such elements and L1 transposition dynamics. Specifically, *Alus* were overrepresented, whereas MIRs and L2s/L3s were underrepresented, over 100 kb analyzed for de novo L1 flanks versus controls (the underrepresentation of L2s/L3s was “patchy”; fig. 4A). All three effects were strong in sFLRs, and *Alus* and L2s/L3s were selected by the mFLR. The underrepresentation of L2/L3 elements in de novo L1 flanks may be explained by 1) the fact that L2 and L3 elements have lost mobility and are common in conserved genomic regions (Silva et al. 2003; Meyers 2006), which lack de novo L1 insertions (fig. 4A); and/or 2) an observation that regions enriched with L2 elements, especially those involved in regulatory networks via miRNAs, may have nucleotide composition or DNA structures repelling insertion of new L1 elements (Petri et al. 2019). This is in line with proposed differences between L1 and L2 elements in structural and functional characteristics, as well as in host defense systems developed by the genome (Rebollo et al. 2012; Lindič et al. 2013; McLaughlin et al. 2014). The overrepresentation of *Alus* in the flanking regions of de novo L1s can be related to the fact that fixed *Alu* elements are frequently found in the GC-rich regions of the genome, which might also be preferred by new L1 insertions (Soriano et al. 1983; Jurka 2004; Wagstaff et al. 2013) (fig. 4A). Also, such enriched *Alu* signals near de novo L1s can in part be explained by the dependency of *Alu* activity on the L1 transposition machinery and the associated endonuclease cleavage sites (Boeke 1997; Deininger 2011; Wimmer et al. 2011; Elbarbary et al. 2016).

In the human-specific versus de novo L1 flanks comparison, *Alus* were globally underrepresented, and MIRs and LTRs were under- and overrepresented, respectively, but in a more “patchy” fashion. *Alus* had a very strong effect in sFLR, and were selected by the mFLR. Higher coverage of LTR elements in the flanks of human-specific versus de novo L1s is consistent with the depletion of both L1 and LTR elements in gene-rich regions, due to negative selection (Deininger and Batzer 2002; Medstrand et al. 2002). MIR-rich regions do not tolerate L1 fixations likely due to the potential regulatory functions of MIRs and their positive correlation with the presence of gene enhancers (Matassi et al. 1998; Jjingo et al. 2014). The paucity of *Alus* in human-specific L1 flanking regions could be explained by their dearth in AT-rich genomic regions, which are favored by L1 fixation (Wagstaff et al. 2012) (fig. 4B).

Replication and Recombination

Our results suggest that replication and recombination profiles have significant but weak associations with the insertion and fixation preferences of L1 elements. We analyzed two replication-associated features—replication timing profile (Ryba et al. 2010) and replication origins (Besnard et al. 2012), and two recombination-associated features—recombination rate (Kong et al. 2010) and recombination hotspots (Myers et al. 2008). We found that early-replicating regions

were positively associated with L1 insertion, but with limited effects (pseudo- $R^2 < 0.5\%$ in sFLRs, both features selected by the mFLR). At the same time, early-replicating regions, replication origins, and recombination hotspots were negative predictors of L1 fixation; all three features had strong effects according to sFLRs, but not selected by the mFLR.

Our results on the association between L1 integration and early replication timing are consistent with the S-phase bias of L1 transposition suggested by other studies (Mita et al. 2018; Sultana et al. 2019). Genomic regions rich in early replicating domains might allow earlier access to sites of less compact chromosomal folding, which are exploited by new L1 integrations (Ryba et al. 2010; Xie et al. 2013; Flasch et al. 2019; Sultana et al. 2019). High density of replication origins might facilitate this process. The negative association of L1 fixation with early replication timing and replication origins might be due to potential effects of replication on the deletion of inserted elements (Yehuda et al. 2018). This is consistent with a potential crosstalk between L1 insertion and other activities and DNA replication, especially during cell division (Ryba et al. 2010). In addition, different replicating domains might not only influence the retrotransposition of L1s, but also affect the DNA replication of L1 genomic sequences (Koren et al. 2012; Zaratiegui 2017), which also suggests additional contribution of the replication process to the L1 life cycle. The negative association of L1 fixation with recombination hotspots might also be due to recombination effects on L1 deletion (Boissinot et al. 2001; Song and Boissinot 2007; Belancio et al. 2009; Bourgeois and Boissinot 2019), as well as to the fact that human-specific L1 regions are located closer to the centromere (fig. 5), where recombination rates are low (Mahtani and Willard 1998; Myers et al. 2005; Croll et al. 2015).

Selection

Here, we focus on the associations of L1 integration and fixation with most conserved elements, CpG islands, exons, and introns—which all act as proxies for purifying selection in the genome. Particularly informative for selection inference are associations between these features and L1 fixation preferences, as gleaned from the comparison of human-specific versus de novo L1 flanks. All four features considered were underrepresented across the whole 100-kb flanks studied (with most conserved elements underrepresented more strongly in the ± 15 kb surrounding the elements; fig. 4B). CpG islands and exons were also selected in the mFLR. These results indicate strong selection against fixation of L1 elements in these functionally constrained parts of the genome (Bejerano et al. 2004; Asthana et al. 2005; Kines and Belancio 2012; Yang et al. 2014).

Integrative Models of L1 Transposition Dynamics

To summarize how different genomic features are correlated with L1 transposition dynamics, we combined the results from IWTomics and FLR analyses (figs. 4 and 5 and table 1) and developed two integrative biological models relating the local genomic landscape with L1 insertion and fixation

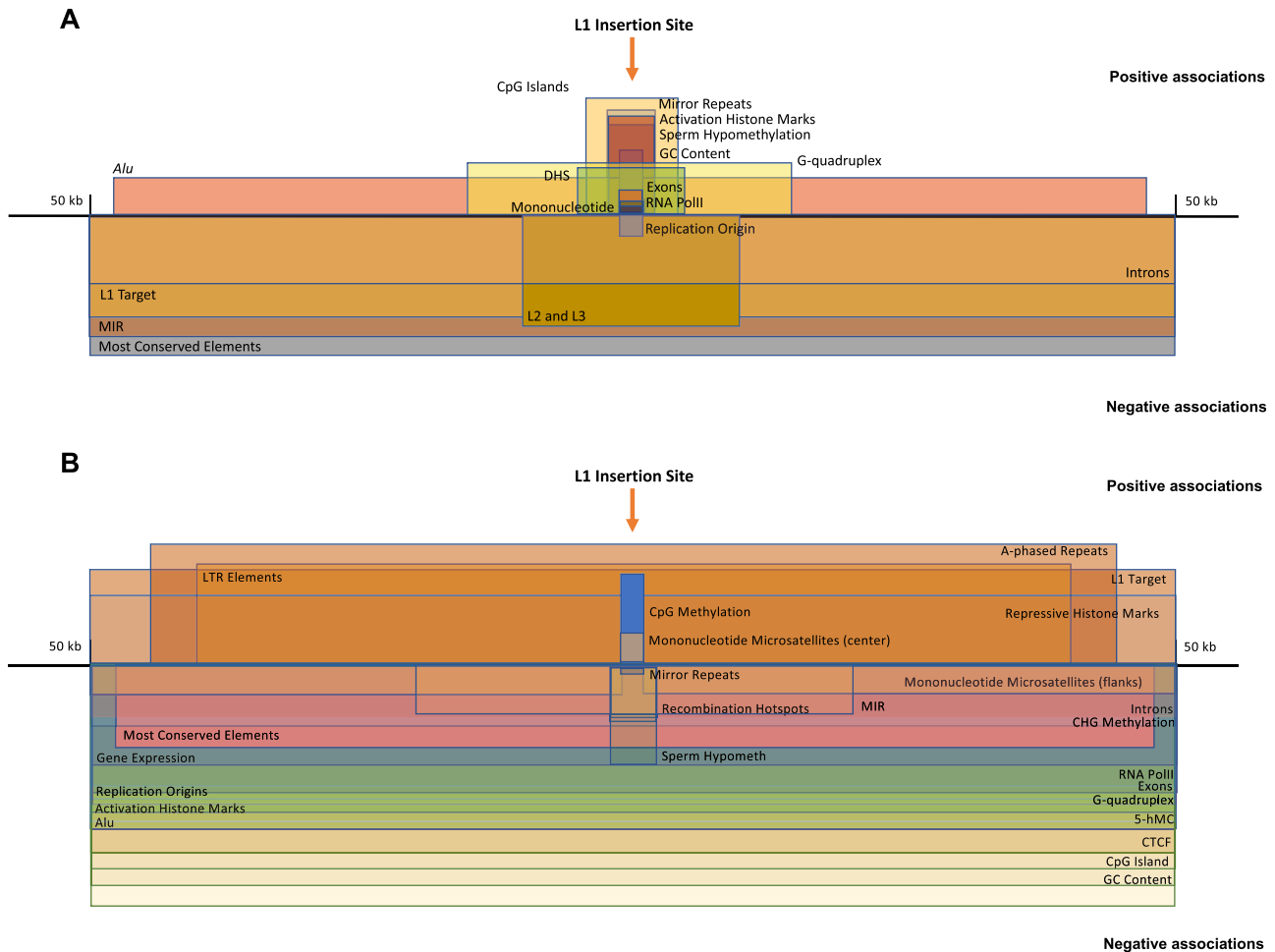


FIG. 6. Integrative models of L1 transposition dynamics based on IWTomics and sFLR results. (A) A model for insertion preferences. (B) A model for fixation preferences. The horizontal black line represents the linear genome structure, with boundaries marking the 100-kb flanking region centered at the L1 insertion site. Each rectangle represents a genomic feature. The placement (above or under the horizontal black line) of the rectangle indicates the sign of a feature's effect (positive or negative), whereas the location and width of the rectangle indicates the location and scale of the effect within the 100-kb flanking region, respectively (based on IWTomics). The height of the rectangle indicates the strength of effect (based on sFLR). Features not included due to unlocalized signals or negligible contributions are: gene expression, direct repeats, mirror repeats, di-, tri-, and tetranucleotide microsatellites, LTRs, recombination hotspots, and five low-resolution features (insertion model); and direct repeats, inverted repeats, Z DNA, and five low-resolution features (fixation model).

preferences (fig. 6). In these models, the scale and the direction (enrichment vs. depletion) of the signal originate from IWTomics results (fig. 4) and are depicted by the width and positive versus negative location in the model schematics, respectively. The strength of the signals originates from the pseudo- R^2 based on the sFLRs (left part of table 1) and is depicted by the bars in the schematics (proportional to bar height; fig. 6).

A Model of L1 Insertion (fig. 6A)

We found that de novo L1s preferentially integrate into actively transcribed, hypomethylated, open-chromatin, and early-replicating regions of the genome. These regions are also enriched in G-quadruplex motifs and mononucleotide microsatellites, which can form non-B DNA (Sinden 2012). These signals are evident at the scale of a few kilobases from the integration site. The potential underlying mechanism is that the genomic regions with actively

transcribed genes usually have higher chromatin accessibility, which facilitates the insertion of L1 elements. Also, unstable non-B DNA might provide opportunities for L1 insertions. Because actively transcribed regions are usually GC-rich (Eyre-Walker and Hurst 2001; Vinogradov 2003), we also observed increased GC content and *Alu* content in regions enriched for de novo L1 insertions. *Alu* elements, particularly older ones, are usually enriched in GC-rich regions (Smit 1999; Gu et al. 2000; Jurka et al. 2004; Kvikstad and Makova 2010). In addition, early-replicating domains and regions with higher transcriptional activities, found to be associated in previous studies (Rivera-Mulia et al. 2015; Fu et al. 2018). However, regions enriched with old inactive TEs (ancient L2/L3 and MIR elements) are usually GC-poor (Matasi et al. 1998; Medstrand et al. 2002), and most conserved elements as a rule are present in nongenic (i.e. AT-rich) regions, explaining why they appear to be negative predictors over large regions in fig. 6A.

A Model of L1 Fixation (fig. 6B)

In contrast to L1 integration, L1 fixation occurs in genomic regions depleted of exons, introns, CpG islands, gene expression, and most conserved elements (this is observed across the 100-kb flanks considered in our analysis). This pattern suggests strong effects of purifying selection acting against fixing L1s in these functional (or putatively functional) regions of the genome (Medstrand et al. 2002; Lowe et al. 2007; Elbarbary et al. 2016). Because genes are usually GC-rich and many of them are actively transcribed from DNA with open chromatin, L1 fixation is negatively associated with GC content, transcription activation histone marks, and other predictors of open chromatin (e.g. DHS and Pol II sites), and positively associated with repressive histone marks (again with effects over the whole 100-kb region analyzed). Therefore, we propose that L1 fixation tends to occur in AT-rich regions with low gene content, low levels of transcription activities, and closed chromatin structure, likely due to the relaxed selection pressure in such regions.

Consequences of L1 Transposition on the Genomic Landscape

Based on our results, the genomic landscape influences L1 transpositional activities and, in turn, fixed L1s modify the genomic landscape surrounding them. For instance, we found an enrichment in CpG methylation ± 1 kb from the insertion site of human-specific L1s (fig. 6B). L1s themselves are prone to DNA methylation (possibly as a genome-defense system to control the expression and spread of the elements) (Yoder et al. 1997; Cohen et al. 2011; Noshay et al. 2019), and methylation may spread to the neighboring region—potentially altering the expression pattern of genes located nearby (Elbarbary et al. 2016). This is consistent with suggestions that L1s can fine-tune transcriptional activities via the genome-wide inhibition of transcriptional elongation (Han et al. 2004) and that L1s can affect gene structure, transcriptional activities, and translation (Belancio et al. 2006; Chuong et al. 2017).

Somatic L1 insertions have also been reported to modulate local DNA methylation levels in the mouse genome by carrying CpG islands that can be subsequently hypermethylated (Grandi et al. 2015). In contrast, an opposite effect was previously observed for germ-line L1 insertions, which often introduce hypomethylated CpG islands and have a localized influence on the neighboring CpG sites (Lees-Murdock et al. 2003; Rosser and An 2012; Grandi et al. 2015). These findings might further explain the enriched CpG methylation close to the insertion sites of human-specific L1s, which result from germ-line insertions.

In addition, when transcribed as part of a larger transcript context, LINEs and SINEs can also affect mRNA stability and thus further influence the translation process (Boissinot et al. 2006; Elbarbary et al. 2016; Petri et al. 2019). We also detected an enrichment in mononucleotide microsatellites ± 1 kb from the insertion site of human-specific L1s (fig. 6B). L1 sequences themselves are known to be hotbeds of AT-rich microsatellites, which constitute the majority of mononucleotide

microsatellites (Kelkar et al. 2011), and it is possible that this process “spills over” to the genomic regions in the vicinity of fixed L1s.

Limitations of the Current Study and Future Directions

Utilizing a comprehensive list of genomic features (the largest list considered to date in this type of studies), we built mFLR models that explain as much as $\sim 30\%$ of the variability in L1 insertion and fixation behavior (table 1). This strong explanatory power allowed us to gain important insights, but we should also ask what may be behind the substantial share of variability that we did *not* explain. *First*, some genomic features affecting L1 integration and fixation dynamics might still be missing from our list. Additional features, once information on them becomes available, should be incorporated in future studies. *Second*, our mFLR models did not comprise explicitly interactions between two or more features. Although interactions between functional predictors can be included in mFLR (Usset et al. 2016; Greven and Scheipl 2017), coefficient estimation becomes more complex and interpretation of the interaction terms is not straightforward. mFLRs with interactions, too, may be leveraged in future studies. *Third*, anticipated advances in statistical methods, particularly in the domain of functional variable selection, are likely to provide better models; an effective algorithm to select functional predictors from a large pool will permit us to include all available genomic features simultaneously and reduce the need of preselecting features based on individual tests (see Materials and Methods section).

Also, de novo L1 insertions included in our study were harvested from a cell line experiment not reflecting germ-line events, in contrast to the polymorphic and human-specific elements. This caveat might influence some of our findings regarding the influence of local genomic features on TE integration, particularly the ones that are cell-type specific, for example, DNA methylation and replication timing profiles (Lees-Murdock et al. 2003; Ryba et al. 2010; Rosser and An 2012; Grandi et al. 2015).

Here, we studied integration preferences for de novo L1 insertions using engineered L1 sequences from kidney stem cells (HEK-293T). This represents a useful model system, and our results about how different genome characteristics influence L1 insertions have important implications for future studies of somatic L1 transposition and its impact on human health and disease. However, HEK-293 cells have previously been reported to be aneuploid, with different levels of structural variation found in several lines, including the HEK-293T line (Lin et al. 2014; Binz et al. 2019) we utilized here. Although this may lead to copy number changes in some genomic regions in the cell line we used, our conclusions are still robust for several reasons. First, while capturing the L1 insertion events, we retained only the unique L1 insertions in each genomic region using the co-occurrence of barcode markers and restriction sites as criteria for successful insertions (supplementary fig. S1, Supplementary Material online and Materials and Methods section). Second, our results on the

chromosome-wide distribution of de novo L1 insertions revealed a strong linear correlation between the number of insertions and chromosomal size (supplementary fig. S3, Supplementary Material online), suggesting minimal effects of potential changes in copy number on target sites. Third, we have contrasted the density of de novo L1 insertions between “aneuploid hotspots” in HEK-293T cells obtained from the literature (Lin et al. 2014; Binz et al. 2019) and other, randomly selected genomic regions. No significant differences were found (supplementary fig. S14, Supplementary Material online), again suggesting a minimal impact of potential target sites duplications on our L1 insertion assay. Fourth, we performed an additional IWTomics analysis of de novo L1 insertion hotspots, defined as multiple overlapping de novo L1 flanking regions (i.e. two, three, or more than three overlapping regions). We observed increasingly stronger signals of genomic features contributing positively to L1 insertions in our model (such as DHS and H3K4me2) in regions where close de novo L1 insertions were found (supplementary fig. S13, Supplementary Material online), suggesting that multiple insertion events were likely driven by local genomic landscape features instead of by amplified regions in the genome of HEK-293T cells.

Importantly, since cell lines might not represent the same karyotype and genomic landscape for TE integration as regular cells, our findings should be validated in future large-scale trio resequencing studies, when such large data sets become available. L1 insertions are reported to be highly frequent in somatic tissues, and can potentially play important roles in developmental processes (Muotri et al. 2005; Kano et al. 2009) and behavior learning (Baillie et al. 2011; Bedrosian et al. 2018). L1 activities can also assist in forming brain plasticity in response to environmental stress via somatic variations in the genome (Baillie et al. 2011; Bedrosian et al. 2018), suggesting potential roles for L1s in the regulation of neurons. Somatic L1 retrotransposition has also been found to occur in different cancer types, including lung and colon cancers, suggesting a potential role of somatic L1 insertions in carcinogenesis (Miki et al. 1992; Scott and Devine 2017).

Finally, we compared our findings with those of two recent de novo L1 integration data sets generated in hESC (Flasch et al. 2019) and HeLa (Sultana et al. 2019) cells (supplementary table S6, Supplementary Material online). Regardless of the differences in experimental design, genomic scales analyzed, and statistical methods used, we still found many features having similar effects on L1 insertion (supplementary table S7, Supplementary Material online). For instance, active histone marks and early replicating domains contributed positively to L1 integration (though with different strengths) across all three studies. However, some other findings were inconsistent among the studies (e.g. for DHS and H3K27me3; supplementary table S7, Supplementary Material online). These discrepancies were *not* due to different statistical approaches, as we still observed them when we reran a substantial part of our IWTomics analyses on the data sets from (Flasch et al. 2019; Sultana et al. 2019) (supplementary figs. S16 and S17, Supplementary Material online) but might be explained by differences in cell lines and genomic scales used.

Future studies applying the same experimental design and analysis framework across different cell lines should be able to pinpoint the causes of these inconsistencies with more confidence.

Conclusions

We presented the first high-resolution, genome-wide analysis of L1 transposition dynamics in an evolutionary framework. We demonstrated that insertion and fixation preferences, and thus the genomic distribution of L1s in the human genome, are affected by the local genomic landscape. The use of FDA statistical tools allowed us to shed light on the potential mechanisms through which regional genomic characteristics influence L1 transposition dynamics. Moreover, our results suggest that L1 transpositional activities, in turn, re-shape the genomic landscape over the course of evolution. This study extends our understanding of L1 transposition dynamics, provides insights into the structure and evolution of the human genome, and illustrates how powerful FDA methodology can aid in extracting information from high-resolution genomic data (Cremona et al. 2019). These tools could be utilized in a variety of genomic studies in the future.

Materials and Methods

In Vivo L1 Insertion Experiment

The positions of de novo L1 insertions were retrieved from an L1 integration experiment in HEK-293T cells according to the following steps. First, vectors containing both a synthetic human L1 element (full-length synthetic ORFeus-Hs) (An et al. 2011) and Green Fluorescent Protein (GFP) were transfected into cultured cells. The vectors were marked with two restriction enzyme sites (*MspI*: CCGG and *TaqI*: TCGA) and 14 different 4- to 6-nucleotide barcodes, which enabled the identification of unique insertion events in the downstream analysis. The high genome-wide densities of the two restriction sites minimized potential bias in detecting the insertion events (supplementary fig. S18, Supplementary Material online). Second, the successful de novo L1 integration events were captured by the expression of GFP. Finally, the positions of L1 insertions were revealed using inverse PCR followed by Illumina sequencing (fig. 1).

Cell Transfection and Fluorescence-Activated Cell Sorting

The plasmid pld225 containing the L1 element was contributed by the lab of Jef Boeke (An et al. 2011). The plasmid DNA was extracted using EndoFree Plasmid Maxi Kit (Qiagen) following the manufacturer's protocol and then prepared for cell transfection. The de novo retrotransposition of L1 was performed in human embryonic kidney cell line HEK-293T, which was maintained in Dulbecco's Modified Eagle Media (Gibco) supplemented with 10% fetal bovine serum, penicillin (100 units/ml), and streptomycin (100 µg/ml). HEK-293T cells were first seeded at 2×10^5 cells per well in six-well plates and grown overnight. The next day, transfections were performed with 1 µg plasmid and 2.5 µl transfection reagent (Fugene HD; Roche) according to the manufacturer's protocol. The day after transfection, cells were treated with trypsin and

transferred to 60-mm plates with complete medium containing puromycin at 1 $\mu\text{g}/\text{ml}$. After 3 days of puromycin selection, cells were washed in 1 \times phosphate-buffered saline and sorted by fluorescence-activated cell sorting. The gating for GFP positive cells was determined by analyzing cells transfected with a puromycin-resistant but GFP-negative control plasmid. A minimum of 500,000 cells were sorted for genomic DNA extraction.

Inverse PCR and Illumina Sequencing

Genomic DNA was extracted using DNeasy blood and tissue kit (Qiagen) following the manufacturer's protocol. Each DNA sample was divided into three 2-mg aliquots, each digested by *Msp* I or *Taq* I individually (New England Biolabs). Digested DNA was ligated overnight at 16 °C in dilute solution to encourage self-ligation. Following ligase inactivation, the ligation pool was then concentrated with either Microcon YM-100 or Amicon Ultra 10K columns (Millipore), and the volume was adjusted to 30 μl with water (when necessary). One microliter was used for inverse PCR with primers (iPCR_F_fixORFeus: AATGATACGGCGACCGCCGAGATCTACACAGCTCTGTAACCATTAGCTGCAATAAA CAAGTTAAC; iPCR_R_fixORFeus: CAAGCAGAAGACGGC ATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGC) that anneal at a complementary region of the pld225 plasmid to amplify the genomic regions flanking L1 insertion loci (fig. 1 and supplementary fig. S1A, Supplementary Material online). The adapter sequences (Adapter [P5] added on the forward iPCR primer: AATGATACGGCGACCGCCGAGATCTACAC; adapter [P7] added on the reverse iPCR primer: CAAGCA GAAGACGGCATAACGAGAT), which allow the PCR products to be sequenced on the Illumina genome analyzer, were added to the inverse PCR primers. The inverse PCR products were then purified using the QIAquick PCR purification kit (Qiagen) and diluted to 10-nM concentration. For each sample, the same amount of PCR product from digestion with each restriction endonuclease was pooled and submitted for Illumina MiSeq sequencing.

Sequencing Analysis of De Novo L1 Insertions

We estimated the insertion locus of each de novo L1 as the 3'-end of read 2 (fig. 1); read 2 should lead to a more precise location than read 1, since it does not need to sequence the entire poly-A tail to reach the insertion locus. In particular, we first filtered the fastq reads by barcode and restriction sites (i.e. we only retained reads with both barcode and at least one of the restriction sites, which correspond to successful L1 insertion events), trimmed the 5' end of the retained reads (keeping the two restriction sites as part of the reads, but not the L1 element, fig. 1 and supplementary fig. S1, Supplementary Material online), and separately stored barcodes and restriction sites. We then trimmed the poly-Ts at the 3'-end of the reads that reached the poly-A tail using Sequence Content Trimmer on Galaxy (Afgan et al. 2018) (parameters: window size 10; frequency threshold 0.89; minimum read length 15), and subsequently using PRINSEQ

0.20.4 (Schmieder and Edwards 2011) (parameters: minimum tail length to trim poly-A/T at 3'-end 4; minimum sequence length in base pairs 15; set output data as FASTQ and Both). Next, we aligned the processed reads to the hg19 reference genome using BWA aligner (with default parameters), and filtered aligned reads with the cut-off parameter $q \geq 1$ using samtools and bedtools. Next, we retrieved the barcode and restriction site information by matching the sequencing read IDs, and annotated the strand information for all of the de novo L1 insertions (supplementary fig. S1, Supplementary Material online). Finally, we collapsed the insertions at the same location by merging reads containing the same barcode and with start (for the positive strand) or end (for the negative strand) positions at a distance less than 4 bp—since it is very unlikely to obtain two very close insertions with the same barcode. As a result, we retrieved 17,037 unique de novo L1 insertions. In addition, we examined the potential bias from genomic poly(A/T) sequences on de novo L1 detection, which might create false positive signals or shift the estimated insertion site, but did not find any significant effect from the genomic poly(A/T) sequences (supplementary note S3, Supplementary Material online).

L1 Data Sets

We first collected the 17,037 de novo L1s from the L1 integration experiment described above. Next, we collected 1,012 polymorphic L1s from a public data set cross-referenced from five different studies (Ewing and Kazazian 2011), and we converted their genomic coordinates from hg18 to hg19 using the LiftOver utility (Casper et al. 2018). Finally, we obtained 1,205 human-specific L1s (annotated as L1HSs) from the RepeatMasker (Smit et al. 2015) annotation of the hg19 genome, as available at the UCSC Genome Browser (Karolchik et al. 2004). For each of these three L1 data sets, we only considered elements on autosomes and chromosome X for the subsequent analyses (supplementary table S1, Supplementary Material online).

Analysis of L1 Distance Distribution

To investigate whether the genomic distribution of L1s is random, we compared the distribution of distances between L1 elements of the same type with a random expectation. We also compared the distribution of distances between L1 elements of two different types with a random expectation. In particular, for each of the three L1 data sets (de novo L1s, polymorphic L1s, and human-specific L1s), we computed distances between each element and the closest element of the same type (on either strand, and either upstream or downstream). We then compared the resulting distance distribution with the distance distribution obtained by randomly shuffling L1 genomic positions (produced considering a data set with the same number of elements and element lengths, but randomized positions). In particular, we performed a bootstrap Kolmogorov–Smirnov test (with 100 resamplings) to test for differences between the empirically observed and the randomized distance distributions, using the “ks.boot” function from the R package

“Matching” (Sekhon 2011). The comparison was visualized using cumulative distribution plots (supplementary fig. S4A–C, Supplementary Material online) and quantile–quantile plots (supplementary fig. S4D–F, Supplementary Material online). In addition, to compare distance distributions across the three L1 data sets, we considered a “normalized” cumulative distribution of the distances between L1 elements. Specifically, we first subsampled 900 elements from each L1 data set, and used these subsamples to compute the cumulative distributions of the distances between L1 elements of the same type. We then normalized these distributions by subtracting the corresponding expected cumulative distribution, and plotted results based on 100 subsamples (fig. 2A). We also analyzed the distances between L1 elements from different data sets using the same procedure and plots (supplementary fig. S5A–F, Supplementary Material online), and compared the distance distributions across the three pairs of data sets (de novo L1 and human-specific L1; de novo L1 and polymorphic L1; polymorphic L1 and human-specific L1; fig. 2B).

Generation of a Comprehensive Blacklist

With the wide use of functional genomics experiments such as ChIP-seq and DNase-seq, it was observed that certain regions of the genome frequently produce artifactual signals, mainly due to the erroneous mapping of reads originating from repetitive regions (ENCODE Project Consortium 2012; Amemiya et al. 2019). These regions are frequently found at certain types of sequences such as centromeres, telomeres, and satellite repeats. Since in our genomic landscape analysis we considered functional genomics features measured by ChIP-seq and DNase-seq, it was essential to remove these artifactual regions. First, we considered the ENCODE blacklist for hg19 (ENCODE Project Consortium 2012; Amemiya et al. 2019), a set of problematic regions in the genome that show artificially high signal in several ENCODE experiments, independently of the cell line and experiment type. We then expanded this blacklist to include problematic regions specific to H1-human embryonic stem cell line (H1-hESC, the cell line we are considering for most of the functional genomic experiments in this study). In particular, we added to the blacklist the genomic regions that showed extreme signal in the H1-hESC ChIP-Seq control sample. The bam files of this control experiment were retrieved from the ENCODE portal (ID: ENCSTR000AMI), and the two replicates were merged into a single control file with samtools. We then employed two approaches to identify regions with extreme signals. First, we called peaks in the control file using MACS2 with default parameters (Zhang et al. 2008; Feng et al. 2012). Second, we screened the genome based on the strength of the control ChIP-Seq signal using a script originally developed by Chris Morrissey and Belinda Giardine from Ross Hardison’s Lab at Penn State University (Morrissey 2013; Cheng et al. 2014). In particular, we considered a 5,000-bp sliding window, and blacklisted all regions with signal 4 standard deviations greater than average, with at least 8-fold change in spikes. The two approaches revealed 2,094 and 519 blacklisted regions, respectively (supplementary table S3, Supplementary Material

online). Our comprehensive blacklist was obtained by merging the ENCODE blacklist with the genomic regions of extreme H1-hESC ChIP-Seq control signals, and it contained 861 regions for a total size of 11.8 Mb (supplementary table S3, Supplementary Material online).

Construction of L1 Flanking and Control Regions

Given the low quality of the sequencing data on sex chromosomes for several genomic features, only the L1 elements on autosomes were considered when we constructed flanking regions for the FDA workflow. This reduced our data sets to 16,322 de novo L1s, 954 polymorphic L1s, and 1,094 human-specific L1s (supplementary table S1, Supplementary Material online). We constructed the flanking regions of the 16,322 autosomal de novo L1 insertions by taking the 50-kb upstream and 50-kb downstream sequences centered at the insertion sites. Overlaps between flanking regions might affect subsequent analyses, assigning more weight to genomic regions covered by multiple L1 flanks; hence we removed part of the overlapping regions, to obtain a data set of non-overlapping regions that maximized the number of regions retained (for a pair of overlapping regions, we kept only the first one; for a group of three overlapping windows we kept the first one and the third one, if they did not overlap, etc.). After filtering out genome assembly gaps and blacklisted regions, we retained a total of 7,981 de novo L1 regions. The 954 autosomal polymorphic L1s (Ewing and Kazazian 2011) are not annotated in the reference genome, hence we used the sites of polymorphic L1 directly and constructed 100-kb flanking regions centered at these sites for each polymorphic L1. After removing overlapping windows, genome assembly gaps, and blacklisted regions, 836 polymorphic L1 regions were retained. For the 1,094 autosomal human-specific L1s (Karolchik et al. 2004; Smit et al. 2015), we first merged the overlapping and adjacent elements and then constructed the regions by flanking 50 kb upstream and 50 kb downstream of each element—the element sequences were not included. This resulted in 834 nonoverlapping human-specific 100-kb L1 flanking regions, after removing genome assembly gaps and blacklisted regions (supplementary table S3, Supplementary Material online). In addition, when constructing the flanking regions we annotated the L1 elements strand information (whether they were inserted on the positive or negative strand) whenever possible. The strand was annotated for all 7,981 de novo L1s regions, but only for 670 polymorphic L1 regions and 725 human-specific L1 regions. This was due to the lack of information about insertion directions for a subset of polymorphic L1s (Ewing and Kazazian 2011) and to the merging of overlapping/adjacent human-specific L1s on opposite strands. We considered the strand information in our FDA (see below).

To construct our controls, we partitioned the hg19 human genome into 100-kb consecutive regions, and excluded those that overlapped with genomic gaps (Kent et al. 2002) or blacklisted regions (as described below). We then filtered out regions overlapping with any of the three L1 100-kb flanking region data sets. In addition, we filtered out regions overlapping 100-kb regions flanking polymorphic L1s from

dbRIP (Wang et al. 2006). These L1s were not included in our polymorphic L1 data set because of their heterogeneity (some of them are in the reference genome while some are not, hence merging them with the Ewing and Kazazian's data set [Ewing and Kazazian 2011] might introduce bias). Yet we excluded them and their flanks to obtain cleaner controls. Finally, to minimize the “noise” from older L1 elements in the genome, we filtered the control regions based on their coverage of all referenced L1 elements in the hg19 genome assembly (except for human-specific L1s since they were already removed). Only control regions with less than 7% coverage by (all referenced) L1 element were kept, leading to a final set of 1,034 “clean” control regions. The 7% threshold was chosen to obtain a number of control regions of the same order of magnitude as in each of the three L1 data sets.

We also considered the fact that some of the 100-kb flanking regions from different L1 data sets (e.g. de novo L1s and human-specific L1s) might overlap, making the data sets not completely independent. We performed IWTomics analysis (see “Interval-Wise Testing with IWTomics” section) both on the complete data sets and after removing all the overlapping regions among different data sets (this left us with 7,517 de novo L1 regions, 332 polymorphic L1 regions, and 357 human-specific L1 regions). Since results were similar (not shown), we kept the overlapping regions among different L1 data sets in our analyses, to maximize the number of considered L1s and thus our statistical power.

Extraction of Genomic Landscape Features

We extracted genomic features in the flanking regions of de novo L1s, polymorphic L1s, human-specific L1s, and in control regions. A total of 49 features were collected from various sources (table 1), among which 44 high-resolution features measured at 1-kb resolution over the 100-kb regions, and five low-resolution features (telomere hexamers, distance to the telomere, distance to the centromere, replication timing, and recombination rate) measured at 100-kb resolution, providing a single measurement per region.

All features obtained from ChIP-Seq experiments (histone modifications, DNase hypersensitive sites, and CTCF motifs) were measured as “signals,” that is as the average number of reads aligned in each 1-kb window. For the features measured as “coverage” (table 1), we computed the proportion of the window covered by the feature using bedtools 2.25.0 (Quinlan 2014). For the features measured as “weighted averages,” the extraction was performed on the Galaxy platform, using the function “Assign Weighted Average Values” (Goecks et al. 2010; Afgan et al. 2018). The extraction of “count” features was performed via bedtools 2.25.0 (Quinlan 2014) and the Galaxy platform (Afgan et al. 2018). While extracting the high-resolution genomic features in the L1 flanking regions, we also considered strand information by reversing the order of 1-kb windows when the element was on the negative strand.

For the high-resolution features, we performed a clustering based on Spearman's correlation. In detail, we considered all 1-kb windows corresponding to L1 flanking regions and control regions and performed a hierarchical clustering using 1-

|Spearman's correlation| as dissimilarity and complete linkage (supplementary fig. S6, Supplementary Material online). At a cutoff of 0.2 (corresponding to a Spearman's correlation of ± 0.8), we identified two tight clusters of features. One comprised three expression profiles (testis expression, gene expression, transcript expression), and the other exon-related (exon coverage and exon expression). We selected only one representative feature for each cluster, and thus excluded three features (testis expression, transcript expression, and exon expression) to reduce multicollinearity issues in the multiple regression analysis (see below).

Interval-Wise Testing with IWTomics

To compare the profiles described by high-resolution features along the 100-kb flanking regions of different L1s, as well as between L1 flanks and control regions, we employed the IWTomics (Pini and Vantini 2016; Cremona et al. 2018). IWTomics is a nonparametric inference procedure that tests for differences between the distributions of two sets of curves. In particular, IWTomics tests the null hypothesis that the distributions of the two sets of curves are equal against the alternative hypothesis that they differ. Importantly, if a significant difference is detected, it provides also the locations (i.e. the 1-kb windows) where such difference is observed. This is achieved by first computing pointwise P values (i.e. a P value for each 1-kb window), and then by adjusting them for multiple comparison, taking into consideration the ordered nature of the measurements (i.e. of the 100 1-kb windows). In addition, the extended version of the test that we employed—implemented in the R package *IWTomics* (Cremona et al. 2018)—also provides the scales (i.e. lengths of the subintervals) at which significant differences unfold (see supplementary fig. S7, Supplementary Material online for an example of IWTomics complete output). The test is fully nonparametric and based on permutations, so it requires no assumption on the curve distributions; this characteristic makes it particularly advantageous for testing the heterogeneous genomics features used in our study.

We employed IWTomics to analyze each of the 41 high-resolution genomic features measured in contiguous 1-kb windows along the 100-kb flanks of different groups (de novo L1s, polymorphic L1s, human-specific L1s), and along the 100-kb control regions. We considered six pairwise comparisons: de novo L1 versus control, polymorphic L1 versus control, human-specific L1 versus control, polymorphic L1 versus de novo L1, human-specific L1 versus de novo L1, and polymorphic L1 versus human-specific L1 (fig. 4 and supplementary fig. S8, Supplementary Material online). Specifically, each curve was defined in the interval $[-50$ kb, 50 kb], where 0 represents the L1 or the center of a control region, with values over a grid of 100 points corresponding to the 100 1-kb windows where the genomic features were measured. To denoise and turn these discrete measurements into functional data, we slightly smoothed each curve using Nadaraya–Watson kernel smoothing with Gaussian kernel and bandwidth = 2. We used a higher level of smoothing (bandwidth = 3) for CpG islands, since the sparsity and uneven distribution of this feature induced massive zero-

inflation (less than 10% of the 1-kb windows had nonzero original measurements). Smoothing was performed via the *smooth* function in the IWTomics package. All curves corresponding to the same feature and to regions of the same type were then aligned over their $[-50 \text{ kb}, 50 \text{ kb}]$ domain, and the four groups of curves were treated as samples from four underlying stochastic functions, each with its distribution. For each genomic feature and each of six pairwise comparison, we tested the null hypothesis that the two stochastic functions have the same distribution, against the alternative hypothesis that their distributions differ. We tested all possible scales, from the 1-kb window to the entire 100-kb region, detecting both the scales and the locations at which the distributions differ. We employed IWTomics with three different test statistics—mean difference, median difference, and multiquantile difference (the sum of the 5th, 25th, 50th, 75th, and 95th quantile differences)—to focus on different characteristics of the distributions. The results with mean differences captured group differentiation quite efficiently, and were thus used for further analysis (multiquantile differences produced similar results, whereas median differences detected less differentiation). IWTomics' empirical P values were computed using 10,000 random permutations. The five low-resolution features were analyzed considering the same six pairwise comparisons and employing the univariate version of IWTomics, where one single value is considered for each 100-kb region (fig. 5).

Since the de novo L1 data set was substantially larger than the polymorphic L1, human-specific L1, and control data sets, we randomly subsampled 1,000 de novo L1 regions to achieve a comparable sample size across all groups analyzed. IWTomics tests involving de novo L1s were run ten times, using 10 independent random subsamples of 1,000 de novo L1 regions. The ten runs produced similar results (e.g. significance, location, and scale; data not shown) which we summarized using pointwise medians of the adjusted P value curves (fig. 4 and supplementary figs. S7 and S8, Supplementary Material online; pointwise medians were computed for each comparison and each possible adjustment scale, from the 1-kb window to the entire 100-kb region).

Single Functional Logistic Regression Analysis

For genomic features that showed significant differences in some of the IWTomics comparisons, we quantified individual effects using sFLR models. For each of the six pairwise comparisons (de novo L1 vs. control, polymorphic L1 vs. control, human-specific L1 vs. control, polymorphic L1 vs. de novo L1, human-specific L1 vs. de novo L1, and polymorphic L1 vs. human-specific L1), we identified significant features (according to IWTomics, at any location and scale), and for each significant feature we fitted a sFLR with the two groups as binary response and the feature as predictor. For example, in the comparison between de novo L1 and control, we fitted single logistic regression models on each of the 33 genomic features (31 high-resolution features and two low-resolution

features) identified by IWTomics in the same comparison, using as response the binary variable denoting de novo L1 flanking regions as $Y = 1$ and control regions as $Y = 0$. Prior to fitting the sFLRs, we examined the distribution of each genomic feature (considering all 1-kb windows for high-resolution features, and all 100-kb regions for low-resolution features) and performed a transformation by taking a shifted logarithm if the distribution was skewed. In detail, we computed the natural logarithm after adding a positive shift parameter s , that is we used the transformation $\log(x + s)$, and we selected $s \in \{1, 10^{-1}, \dots, 10^{-10}\}$ to maximize the P values of the Shapiro–Wilk normality test on the transformed data in all groups (except for replication timing, that had both positive and negative values, where we considered $s \in \{2, 4, \dots, 22\}$). Each genomic feature was then included in a sFLR as either functional or scalar predictor [indicated as $x(t)$ and x in the following equations, respectively]—with the model reducing to an ordinary single logistic regression in the latter case. In symbols, we fitted the models

$$\text{logit}(E[Y|x(t)]) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \int_{-50}^{50} \beta(t)x(t)dt$$

$$\text{logit}(E[Y|x]) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta x$$

for functional and scalar predictors, respectively, where P represents the probability of being in the group denoted by $Y = 1$ conditionally to the observed predictor. A high-resolution feature was treated as a functional predictor in a given comparison if it showed significant, localized differences in IWTomics results and in pointwise boxplots. In contrast, a high-resolution feature was considered as a scalar predictor in a given comparison if IWTomics results suggested a significant but nonlocalized (i.e. global) difference across the entire 100-kb interval, and pointwise boxplots showed flat signals. In this case, the high-resolution feature was summarized by computing its average over the 100 1-kb measurements in each 100-kb region. The five low-resolution features (recombination rate, replication timing, distance from the telomere, distance from the centromere, and telomere hexamers), when significant, were also treated as scalar predictors. The R function *glm* was employed to fit the models for scalar predictors, using the *binomial* family and the *logit* link function. The sFLR for functional predictors were fitted with the function *frege.glm* from the R package *fda.usc* (Febrero Bande and Oviedo de la Fuente 2012), using again the *binomial* family and the *logit* link function. A quadratic B-spline basis (order 3) with six equispaced breaks was employed for representing both $\beta(t)$ and $x(t)$ (we used the function *create.bspline.basis* from the R package *fda*).

For each sFLR model (in each comparison), we measured the discriminatory strength of the predictor with the pseudo- R^2 , which indicates the proportion of Deviance Explained by the model, that is with

$$DE = R_{\text{psuedo}}^2 = \frac{D_{\text{null}} - D_{\text{model}}}{D_{\text{null}}},$$

where D_{null} is the null deviance and D_{model} is the model residual deviance.

In comparisons involving de novo L1s, also the sFLR analysis was performed ten times—using the same ten random subsamples of de novo L1 flanking regions generated for the IWTomics analysis. Again, results from the ten random subsamples revealed similar signals (pseudo- R^2 , significance, beta coefficients, etc.). We then compared the pseudo- R^2 values for each predictor across all 10 random subsamples and selected the subsample (random 1) with the least extreme values for downstream analyses (supplementary fig. S10, Supplementary Material online).

Multiple Functional Logistic Regression Analysis

For each of the six pairwise comparisons (de novo L1 vs. control, polymorphic L1 vs. control, human-specific L1 vs. control, polymorphic L1 vs. de novo L1, human-specific L1 vs. de novo L1, and polymorphic L1 vs. human-specific L1), we employed a mFLR model to quantify the joint effects of different genomic landscape features on the insertion and fixation preferences of the L1 elements. Similarly to what was done in the above sFLR analysis, we considered the genomic features that showed significant differences in some of the IWTomics comparisons, and we included each of them in the mFLR model either as a functional or as a scalar predictor [indicated as $x_j(t)$ and x_j in the following equations, respectively]. If a feature showed a skewed distribution, we transformed it with a shifted logarithm in the same way we did for sFLRs (see details in previous Subsection). As response, we used a binary indicator for the two types of regions being compared (e.g. in the comparison between de novo L1 and control we indicated de novo regions with $Y = 1$ and control regions with $Y = 0$). In symbols, for each comparison we fitted the model:

$$\begin{aligned} & \text{logit}\left(E[Y|x_1, \dots, x_r, x_{r+1}(t), \dots, x_{r+s}(t)]\right) \\ &= \ln\left(\frac{P}{1-P}\right) \\ &= \beta_0 + \sum_{j=1}^r \beta_j x_j + \sum_{j=r+1}^{r+s} \int_{-50}^{50} \beta_j(t) x_j(t) dt, \end{aligned}$$

where x_1, \dots, x_r are the r scalar predictors, $x_{r+1}(t), \dots, x_{r+s}(t)$ are the s functional predictors, and P represents the probability of being in the group denoted by $Y = 1$ conditionally to the observed predictors.

Even omitting features that were nonsignificant in the IWTomics analysis, and reducing to scalar predictors high-resolution features that showed significant but flat signals, each mFLR model included several predictors. For example, the mFLR model to compare de novo L1 and control included 13 scalar and 20 functional predictors. To reduce the complexity of the mFLR models and retain only relevant predictors (i.e. only those genomic features that are useful in differentiating among the two compared groups),

we employed a variable selection method for generalized functional regression models based on group lasso (Matsui 2014). In particular, we standardized each predictor and expressed each of the functional predictors $x_j(t)$ via a quadratic B-spline basis expansion (order 3) with six equispaced breaks (we used the function *create.bspline.basis* from the R package *fda*):

$$x_j(t) = \sum_{k=1}^6 w_{j,k} \phi_k(t) = \mathbf{w}_j^T \boldsymbol{\phi}(t).$$

The same basis was employed for representing each coefficient curve $\beta_j(t)$, obtaining:

$$\beta_j(t) = \sum_{k=1}^6 b_{j,k} \phi_k(t) = \mathbf{b}_j^T \boldsymbol{\phi}(t).$$

The mFLR model could therefore be rewritten as:

$$\begin{aligned} & \text{logit}\left(E[Y|x_1, \dots, x_r, x_{r+1}(t), \dots, x_{r+s}(t)]\right) \\ &= \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{j=1}^r \beta_j x_j + \sum_{j=r+1}^{r+s} \mathbf{b}_j^T \mathbf{J}_\phi \mathbf{w}_j, \end{aligned}$$

where

$$\mathbf{J}_\phi = \int_{-50}^{50} \boldsymbol{\phi}(t) \boldsymbol{\phi}_j^T(t) dt$$

is the cross-product matrix of the B-spline basis. The vector of parameters

$$\mathbf{b} = [\beta_0, \beta_1, \dots, \beta_r, \mathbf{b}_{r+1}^T, \dots, \mathbf{b}_{r+s}^T]^T$$

was then estimated using the group lasso penalty for logistic regression (Yuan and Lin 2006; Meier et al. 2008), treating the parameters corresponding to the expansion of the same predictor as a group. In symbols, the vector of parameters was estimated by minimizing the penalized log-likelihood function

$$l_\lambda(\mathbf{b}) = -l(\mathbf{b}) + \lambda(|\beta_0| + \sum_{j=1}^r |\beta_j| + \sum_{j=r+1}^{r+s} \sqrt{6} \|\mathbf{b}_j\|),$$

where $l(\mathbf{b})$ is the log-likelihood function, $\|\cdot\|$ indicates the Euclidean norm, and λ is a regularization parameter. This minimization was performed using an R in-house script based on Matsui's code (Matsui 2014). The regularization parameter λ was selected using the BIC (see supplementary fig. S12, Supplementary Material online).

To conclude, for each comparison we fitted a final mFLR comprising only the variables selected by the group lasso. Also here, we employed the function *fregre.glm* from the R package *fda.usc* (Febrero Bande and Oviedo de la Fuente 2012), with *binomial* family, *logit link* function and a quadratic B-spline basis (order 3) with six equispaced breaks for representing each $\beta_j(t)$ and $x_j(t)$ (we used again the function *create.bspline.basis* from the R package *fda*).

We measured the total discriminatory power of each final mFLR model with the total pseudo- R^2 , which corresponds to the proportion of Deviance Explained by the model:

$$DE = R_{\text{psuedo}}^2 = \frac{D_{\text{null}} - D_{\text{model}}}{D_{\text{null}}},$$

where D_{null} is the null deviance and D_{model} is the model's residual deviance. In addition, we measured the contribution of each individual feature to the final mFLR model with the RCDE:

$$\text{RCDE} = \frac{(D_{\text{null}} - D_{\text{model}}) - (D_{\text{null}} - D_{\text{red_model}})}{(D_{\text{null}} - D_{\text{model}})},$$

where D_{null} is the null deviance, D_{model} is the model's residual deviance and $D_{\text{red_model}}$ is the residual deviance of a reduced model obtained by removing the predictor whose contribution is being measured.

Data and Code Availability

We have set up a github repository (https://github.com/makovalab-psu/L1_Project; last accessed August 26, 2020) and shared the chromosomal coordinates of de novo, polymorphic, and human-specific L1s analyzed in this study (https://github.com/makovalab-psu/L1_Project/tree/master/Datasets; last accessed August 26, 2020). The repository also contains the computational pipelines and code, along with the corresponding intermediate files (.RData) used to generate the results. The raw sequencing reads from the de novo L1 insertion experiment were uploaded to the Short Read Archive (SRA) under the accession number PRJNA640178.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to Rebeca Campos-Sanchez, Ross Hardison, Belinda Giardine, Anton Nekrutenko, Martin Cech, Dave Bouvier, and Dan Blankenberg for their assistance. We thank Hidetoshi Matsui for providing R code for variable selection in Functional Logistic Regression models, Jef Boeke for providing the plasmid pld225 containing the synthetic full-length ORFeus-Hs element, and Wilfried Guiblet and Debmalaya Nandy for helpful discussions.

This study was supported by the funds made available through the Clinical and Translational Sciences Institute, the Institute for Computational and Data Sciences, and the Eberly College of Sciences—at Penn State. Additional support was provided under grants from the Pennsylvania Department of Health using Tobacco Settlement and CURE Funds. The department specifically disclaims any responsibility for any analyses, responsibility, or conclusions. R.D.M. was supported by NIH (RF1 MH117070 and R01GM123203).

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. 2018. The Galaxy platform

- for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46(W1):W537–544.
- Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep.* 9(1):9354.
- An W, Dai L, Niewiadomska AM, Yetil A, O'Donnell KA, Han JS, Boeke JD. 2011. Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs. *Mobile DNA* 2(1):2.
- An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, Ye P, Triplett C, Boeke JD. 2006. Active retrotransposition by a synthetic L1 element in mice. *Proc Natl Acad Sci U S A.* 103(49):18662–18667.
- Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of Sequence Conservation at Nucleotide Resolution. *PLoS Comput Biol.* 3(12):e254.
- Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479(7374):534–537.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–837.
- Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet.* 42(1):587–617.
- Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet.* 12(1):187–215.
- Bedrosian TA, Quayle C, Novaresi N, Gage FH. 2018. Early life experience drives structural variation of neural genomes in mice. *Science* 359(6382):1395–1399.
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304(5675):1321–1325.
- Belancio VP, Deininger PL, Roy-Engel AM. 2009. LINE dancing in the human genome: transposable elements and disease. *Genome Med.* 1(10):97.
- Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* 34(5):1512–1521.
- Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin J-M, Lemaitre J-M. 2012. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol.* 19(8):837–844.
- Binz RL, Tian E, Sadhukhan R, Zhou D, Hauer-Jensen M, Pathak R. 2019. Identification of novel breakpoints for locus- and region-specific translocations in 293 cells by molecular cytogenetics before and after irradiation. *Sci Rep.* 9(1):1–0.
- Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet.* 13(11):770–780.
- Boeke JD. 1997. LINEs and Alus—the polyA connection. *Nat Genet.* 16(1):6–7.
- Boissinot S. 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res.* 14(7):1221–1231.
- Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol.* 17(6):915–928.
- Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. 2006. Fitness cost of LINE-1 (L1) activity in humans. *Proc Natl Acad Sci U S A.* 103(25):9590–9594.
- Boissinot S, Entezam A, Furano AV. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol.* 18(6):926–935.
- Bourgeois Y, Boissinot S. 2019. On the population dynamics of junk: a review on the population genomics of transposable elements. *Genes* 10(6):419.
- Branco MR, Ficuz G, Reik W. 2012. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet.* 13(1):7–13.
- Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The

- evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
- Calado RT, Dumitriu B. 2013. Telomere dynamics in mice and humans. *Semin Hematol.* 50(2):165–174.
- Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. 2016. Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. *PLoS Comput Biol.* 12:1–41.
- Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2018. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* 46(D1):D762–769.
- Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfvsky N, Luke BT, Bacolla A, Collins JR, Stephens RM. 2011. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* 39:383.
- Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starnier NJ, Halusa GN, Volfvsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* 41:94–100.
- Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* 515:371–375.
- Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet.* 18(2):71–86.
- Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* 145(5):773–786.
- Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 10(10):691–703.
- Cost GJ, Boeke JD. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* 37(51):18081–18093.
- Cremona MA, Pini A, Cumbo F, Makova KD, Chiaromonte F, Vantini S. 2018. IWTomics: testing high-resolution sequence-based “Omics” data at multiple locations and scales. *Bioinformatics* 34(13):2289–2291.
- Cremona MA, Xu H, Makova KD, Reimherr M, Chiaromonte F, Madrigal P. 2019. Functional data analysis for computational biology. *Bioinformatics* 35(17):3211–3213.
- Croll D, Lendenmann MH, Stewart E, McDonald BA. 2015. The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics* 201(3):1213–1228.
- Deininger P. 2011. Alu elements: know the SINEs. *Genome Biol.* 12(12):236.
- Deininger PL, Batzer MA. 2002. Mammalian retroelements. *Genome Res.* 12(10):1455–1465.
- de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7(12):e1002384.
- Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4(5):e1000071.
- Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. *Science* 351(6274):aac7247.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
- EpiGenie epigenetics background, tools and database. 2020. Available from: <https://epigenie.com/epigenie-learning-center/>. Accessed August 26, 2020.
- Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* 21(6):985–990.
- Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet.* 2(7):549–555.
- Febrero Bande M, Oviedo de la Fuente M. 2012. Statistical Computing in Functional Data Analysis: The R Package Fda.usc. *J Stat Softw.* 5128(4):1.
- Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc.* 7(9):1728–1740.
- Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87(5):905–916.
- Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* 29(10):1567–1577.
- Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, Wilson TE, Moran JV. 2019. Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. *Cell* 177(4):837–851.e28.
- Fu H, Baris A, Aladjem MI. 2018. Replication timing and nuclear structure. *Curr Opin Cell Biol.* 52:43–50.
- Gates LA, Shi J, Rohira AD, Feng Q, Zhu B, Bedford MT, Sagum CA, Jung SY, Qin J, Tsai M-J, et al. 2017. Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J Biol Chem.* 292(35):14456–14472.
- Ghirlando R, Felsenfeld G. 2016. CTCF: making the right connections. *Genes Dev.* 30(8):881–891.
- Goecks J, Nekrutenko A, Taylor J, Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11(8):R86.
- Goodier JL, Kazazian HH Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* 135(1):23–35.
- Graham T, Boissinot S. 2006. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J Biomed Biotechnol.* 2006:1–5.
- Grandi FC, Rosser JM, Newkirk SJ, Yin J, Jiang X, Xing Z, Whitmore L, Bashir S, Ivics Z, Izsvák Z, et al. 2015. Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites. *Genome Res.* 25(8):1135–1146.
- Greven S, Scheipl F. 2017. A general framework for functional regression modelling. *Stat Model.* 17(1–2):1–35.
- Gu Z, Wang H, Nekrutenko A, Li W-H. 2000. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* 259(1–2):81–88.
- Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejniovská I, Kejniovsky E, Eckert K, Chiaromonte F, Makova KD. 2018. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res.* 28(12):1767–1778.
- Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429(6989):268–274.
- He Y, Ecker JR. 2015. Non-CG methylation in the human genome. *Annu Rev Genomics Hum Genet.* 16(1):55–77.
- Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6(4):283–289.
- Horvath R, Slotte T. 2017. The role of small RNA-based epigenetic silencing for purifying selection on transposable elements in *Capsella grandiflora*. *Genome Biol Evol.* 9(10):2911–2920.
- Hou Y, Li F, Zhang R, Li S, Liu H, Qin ZS, Sun X. 2019. Integrative characterization of G-quadruplexes in the three-dimensional chromatin structure. *Epigenetics* 14(9):894–911.
- Huang J, Lynn JS, Schulte L, Vendramin S, McGinnis K. 2017. Epigenetic control of gene expression in maize. *Int Rev Cell Mol Biol.* 328:25–48.
- Ivics Z, Li MA, Mátés L, Boeke JD, Nagy A, Bradley A, Izsvák Z. 2009. Transposon-mediated genome manipulation in vertebrates. *Nat Methods* 6(6):415–422.
- Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak VV, Jordan IK. 2014. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA* 5(1):14.

- Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A*. 94(5):1872–1877.
- Jurka J. 2004. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev*. 14(6):603–608.
- Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A*. 101(5):1268–1272.
- Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH Jr. 2009. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev*. 23(11):1303–1312.
- Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 32(Database issue):D493–496.
- Kazazian HH, Moran JV. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet*. 19(1):19–24.
- Kejnovský E, Michalovova M, Steflava P, Kejnovska I, Manzano S, Hobza R, Kubat Z, Kovarik J, Jamilena M, Vyskot B. 2013. Expansion of microsatellites on evolutionary young Y chromosome. *PLoS One* 8(1):e45519.
- Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res*. 21(12):2038–2048.
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res*. 12(6):996–1006.
- Khan H, Smit A, Boissinot S. 2005. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res*. 16(1):78–87.
- Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6):1231–1245.
- Kines KJ, Belancio VP. 2012. Expressing genes do not forget their LINES: transposable elements and gene expression. *Front Biosci*. 17(1):1329–1344.
- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–1103.
- Konkel MK, Walker JA, Batzer MA. 2010. LINES and SINEs of primate evolution. *Evol Anthropol*. 19(6):236–249.
- Konkel MK, Wang J, Liang P, Batzer MA. 2007. Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. *Gene* 390(1–2):28–38.
- Koren A, Polak P, Nemes J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet*. 91(6):1033–1040.
- Krueger F, Kreck B, Franke A, Andrews SR. 2012. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* 9(2):145–151.
- Kvikstad EM, Makova KD. 2010. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res*. 20(5):600–613.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
- Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 22(9):1813–1831.
- Lee M, Hills M, Conomos D, Stutz MD, Dagg RA, Lau LMS, Reddel RR, Pickett HA. 2014. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res*. 42(3):1733–1746.
- Lees-Murdock DJ, De Felici M, Walsh CP. 2003. Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* 82(2):230–237.
- Lexa M, Steflava P, Martinek T, Vorlickova M, Vyskot B, Kejnovsky E. 2014. Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* 15(1):1032.
- Lin Y-C, Boone M, Meuris L, Lemmens I, Van Roy N, Soete A, Reumers J, Moisse M, Plaisance S, Drmanac R, et al. 2014. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun*. 5:4767.
- Lindič N, Budić M, Petan T, Knisbacher BA, Levanon EY, Lovšin N. 2013. Differential inhibition of LINE1 and LINE2 retrotransposition by vertebrate AID/APOBEC proteins. *Retrovirology* 10(1):156.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
- Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol*. 11(12):2453–2465.
- Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J. 2018. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* 553(7687):228–232.
- Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A*. 104(19):8005–8010.
- Mahtani MM, Willard HF. 1998. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res*. 8(2):100–110.
- Matassi G, Labuda D, Bernardi G. 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. *FEBS Lett*. 439(1–2):63–65.
- Matsui H. 2014. Variable and boundary selection for functional data via multiclass logistic regression modeling. *Comput Stat Data Anal*. 78:176–185.
- McLaughlin RN Jr, Young JM, Yang L, Neme R, Wichman HA, Malik HS. 2014. Positive selection and multiple losses of the LINE-1-derived L1TD1 gene in mammals suggest a dual role in genome defense and pluripotency. *PLoS Genet*. 10(9):e1004531.
- Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res*. 12(10):1483–1495.
- Meier L, Van De Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *J R Stat Soc Ser B Stat Methodol*. 70(1):53–71.
- Meyer TJ, Held U, Nevonen KA, Klawitter S, Pirzer T, Carbone L, Schumann GG. 2016. The flow of the gibbon LAVA element is facilitated by the LINE-1 retrotransposition machinery. *Genome Biol Evol*. 8(10):3209–3225.
- Meyers RA ed. 2006. Anthology of human repetitive DNA. In: Encyclopedia of molecular cell biology and molecular medicine. Vol. 3. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. p. 370.
- Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*. 52:643–645.
- Mita P, Wudzinska A, Sun X, Andrade J, Nayak S, Kahler DJ, Badri S, LaCava J, Ueberheide B, Yun CY, et al. 2018. LINE-1 protein localization and functional dynamics during the cell cycle. *eLife* 7:e30058.
- Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD. 2011. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146(6):1029–1041.
- Mooijman D, Dey SS, Boisset JC, Crosetto N, Van Oudenaarden A. 2016. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat Biotechnol*. 34(8):852–856.
- Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* 87(5):917–927.
- Morrissey CS. 2013. Understanding the epigenetics of erythroid differentiation through the power of deep sequencing.

- Muotri AR, Nakashima K, Toni N, Sandler VM, Gage FH. 2005. Development of functional human embryonic stem cell-derived neurons in mouse brain. *Proc Natl Acad Sci U S A*. 102(51):18644–18648.
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310(5746):321–324.
- Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*. 40(9):1124–1129.
- Nishida H, Suzuki T, Ookawa H, Tomaru Y, Hayashizaki Y. 2005. Comparative analysis of expression of histone *H2a* genes in mouse. *BMC Genomics* 6(1):108.
- Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, Stitzer MC, Crisp PA, Hirsch CN, Zhang X, et al. 2019. Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet*. 15(9):e1008291.
- Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. *BioEssays* 31(7):703–714.
- Ong C-T, Corces VG. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*. 15(4):234–246.
- Ostertag EM, Kazazian HH Jr. 2001a. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res*. 11(12):2059–2065.
- Ostertag EM, Kazazian HH Jr. 2001b. Biology of mammalian L1 retrotransposons. *Annu Rev Genet*. 35(1):501–538.
- Ovchinnikov I, Rubin A, Swergold GD. 2002. Tracing the LINEs of human evolution. *Proc Natl Acad Sci U S A*. 99(16):10522–10527.
- Payer LM, Burns KH. 2019. Transposable elements in human genetic disease. *Nat Rev Genet*. 20(12):760–713.
- Pedram M, Sprung CN, Gao Q, Lo AWI, Reynolds GE, Murnane JP. 2006. Telomere Position Effect and silencing of transgenes near telomeres in the mouse. *MCB* 26(5):1865–1878.
- Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, Zemojtel T. 2017. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res*. 45(D1):D68–73.
- Petri R, Brattås PL, Sharma Y, Jönsson ME, Pircs K, Bengzon J, Jakobsson J. 2019. LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLoS Genet*. 15(3):e1008036.
- Philippe C, Vargas-Landin DB, Doucet AJ, Van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G. 2016. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* 5:1–30.
- Pini A, Vantini S. 2016. The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics* 72(3):835–845.
- Plohl M, Prats E, Martínez-Lage A, González-Tizón A, Méndez J, Cornudella L. 2002. Telomeric localization of the vertebrate-type hexamer repeat, (TTAGGG)_n, in the wedgeshell clam *Donax trunculus* and other marine invertebrate genomes. *J Biol Chem*. 277(22):19839–19846.
- Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinform*. 47(1):11.12.1–11.12.34.
- Ramsay J, Silverman BW. 2007. *Applied Functional data analysis: methods and case studies*. New York: Springer Press.
- Rangasamy D, Greaves I, Tremethick DJ. 2004. RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. *Nat Struct Mol Biol*. 11(7):650–655.
- Ratcliffe SJ, Heller GZ, Leader LR. 2002. Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional Logistic Regression. *Stat Med*. 21(8):1115–1127.
- Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet*. 46(1):21–42.
- Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea G-O, Muñoz-Lopez M, Jesuadian JS, Kempen M-JHC, Carreira PE, Jeddleloh JA, et al. 2017. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res*. 27(8):1395–1405.
- Rio DC, Clark SG, Tjian R. 1985. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* 227(4682):23–28.
- Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, Loring JF, Lian Z, Weissman S, Robins AJ, et al. 2015. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Res*. 25(8):1091–1103.
- Rodriguez J, Vives L, Jordà M, Morales C, Muñoz M, Vendrell E, Peinado MA. 2008. Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. *Nucleic Acids Res*. 36(3):770–784.
- Rosser JM, An W. 2012. L1 expression and regulation in humans and rodents. *Front Biosci*. E4(6):2203–2225.
- Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*. 20(6):761–770.
- Sahakyan AB, Murat P, Mayer C, Balasubramanian S. 2017. G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol*. 24(3):243–247.
- Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148(1–2):335–348.
- Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6):863–864.
- Scott EC, Devine SE. 2017. The role of somatic L1 retrotransposition in human cancers. *Viruses* 9(6):131.
- Sekhon JS. 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 42(7):1–52.
- Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashov AS. 2003. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res*. 82(1):1–18.
- Sinden RR. 2012. *DNA structure and function*. San Diego: Elsevier Press.
- Singer MF. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28(3):433–434.
- Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev*. 9(6):657–663.
- Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. Available from: <http://www.repeatmasker.org>. Accessed August 26, 2020.
- Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* 390(1–2):206–213.
- Soriano P, Meunier Rotival M, Bernardi G. 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A*. 80(7):1816–1820.
- Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol Evol*. 9(1):161–177.
- Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HYK, Lee W-P, et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet*. 7(8):e1002236.
- St. Laurent G, Hammell N, McCaffrey TA. 2010. A LINE-1 component to human aging: do LINE elements exact a longevity cost for evolutionary advantage? *Mech Ageing Dev*. 131(5):299–305.
- Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z, Cedar H. 2009. Developmental programming of {CpG} island methylation profiles in the human genome. *Nat Struct Mol Biol*. 16(5):564–571.
- Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioeger L, Nigumann P, Saccani S, Andrau J-C, et al. 2019. The landscape of L1 retrotransposons in the human genome is shaped by pre-insertion sequence biases and post-insertion selection. *Mol Cell* 74(3):555–570.e7.

- Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet.* 18(5):292–308.
- Sun J, Rockowitz S, Chauss D, Wang P, Kantorow M, Zheng D, Cvekl A. 2015. Chromatin features, RNA polymerase II and the comparative expression of lens genes encoding crystallins, transcription factors, and autophagy mediators. *Mol Vis.* 21:955–973.
- Szulwach KE, Li X, Li Y, Song C-X, Han JW, Kim S, Namburi S, Hermetz K, Kim JJ, Rudd MK, et al. 2011. Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet.* 7(6):e1002154.
- Szulwach KE, Li X, Li Y, Song C-X, Wu H, Dai Q, Irier H, Upadhyay AK, Gearing M, Levey AI, et al. 2011. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci.* 14(12):1607–1616.
- Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 13(1):36–46.
- Tsompana M, Buck MJ. 2014. Chromatin accessibility: a window into the genome. *Epigenet Chromatin* 7(1):33.
- Usset J, Staicu A-M, Maity A. 2016. Interaction models for functional regression. *Comput Stat Data Anal.* 94:317–329.
- Venkatesan S, Khaw AK, Hande MP. 2017. Telomere biology—insights into an intriguing phenomenon. *Cells* 6(2):15.
- Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31(7):1838–1844.
- Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, Roy-Engel AM. 2012. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS Genet.* 8(8):e1002842.
- Wagstaff BJ, Kroutter EN, Derbes RS, Belancio VP, Roy-Engel AM. 2013. Molecular reconstruction of extinct LINE-1 elements and their interaction with nonautonomous elements. *Mol Biol Evol.* 30(1):88–99.
- Wallrath LL, Lu Q, Granok H, Elgin SCR. 1994. Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. *BioEssays* 16(3):165–170.
- Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. 2006. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat.* 27(4):323–329.
- Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–1678.
- Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet.* 39(4):457–466.
- Wimmer K, Callens T, Wernstedt A, Messiaen L. 2011. The *NF1* gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet.* 7(11):e1002371.
- Wylie A, Jones AE, D'Brot A, Lu W-J, Kurtz P, Moran JV, Rakheja D, Chen KS, Hammer RE, Comerford SA, et al. 2016. p53 genes function to restrain mobile elements. *Genes Dev.* 30(1):64–77.
- Xie Y, Mates L, Ivics Z, Zsvák Z, Martin SL, An W. 2013. Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. *Mobile DNA* 4(1):10.
- Yang L, Brunsfeld J, Scott L, Wichman H. 2014. Reviving the dead: history and reactivation of an extinct L1. *PLoS Genet.* 10(6):e1004395.
- Yehuda Y, Blumenfeld B, Mayorek N, Makedonski K, Vardi O, Cohen-Daniel L, Mansour Y, Baror-Sebban S, Masika H, Farago M, et al. 2018. Germline DNA replication timing shapes mammalian genome composition. *Nucleic Acids Res.* 46(16):8299–8310.
- Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 13(8):335–340.
- Yu Q, Zhang W, Zhang X, Zeng Y, Wang Y, Wang Y, Xu L, Huang X, Li N, Zhou X, et al. 2017. Population-wide sampling of retrotransposon insertion polymorphisms using deep sequencing and efficient detection. *GigaScience* 6(9):1–11.
- Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J R Stat Soc B* 68(1):49–67.
- Zaratiegui M. 2017. Cross-regulation between transposable elements and host DNA replication. *Viruses* 9(3):57.
- Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916–919.
- Zhang F, Boerwinkle E, Xiong M. 2014. Epistasis analysis for quantitative traits by functional regression model. *Genome Res.* 24(6):989–998.
- Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9(9):R137.
- Zhao B, Wu Q, Ye AY, Guo J, Zheng X, Yang X, Yan L, Liu Q-R, Hyde TM, Wei L, et al. 2019. Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet.* 15(4):e1008043.
- Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci.* 67(1):43–62.
- Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet.* 12(1):7–18.