The Pennsylvania State University

The Graduate School

GENOME ARCHITECTURE AND EVOLUTION SHAPED BY TRANSPOSABLE ELEMENTS

A Dissertation in

Genetics

by

Di Chen

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2021

The dissertation of Di Chen was reviewed and approved by the following:

Mark D. Shriver
Professor of Anthropology
Dissertation Advisor

Ross C. Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology
Chair of Committee

Stephen W. Schaeffer
Professor of Biology
Head, Department of Biology

Robert F. Paulson
Professor of Veterinary and Biomedical Sciences
Chair, Intercollege Graduate Program in Genetics

ABSTRACT

Transposable Elements (TEs) are important constituents of the human genome and are considered to play a critical role in shaping the genome architecture and evolution. In previous *in vivo* and *in vitro* studies, the genomic distribution of TEs has been investigated along with some of their functions in gene regulation and various cellular processes. However, to date, there has not been a high-resolution, genome-wide study of TEs in an evolutionary framework, through which the insertion and fixation preferences of the elements can be addressed in detail. Also, the interactions between TE activities and local genome landscape have not been fully revealed. The long-term goal of this study is to characterize the transposition dynamics of TEs and to further understand their contribution to the structure, function, and evolution of the human genome. In this dissertation, I focused on one specific TE family, namely the Long Interspersed Element-1 (LINE-1 or L1), which constitutes >17% of the human genome and still actively transpose in it. I studied the genome-wide insertion and fixation preferences of L1s at a high-resolution and investigated their interactions with different genomic landscape features such as histone modifications and DNA methylation. In detail, I analyzed three large datasets of L1s that integrated at different evolutionary time scales: 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). I also characterized 49 genomic features—proxying chromatin accessibility, transcriptional activity, replication, recombination, etc.—in the ±50 kb flanks of these elements. These features were contrasted between the three L1 datasets and L1-depleted regions using state-of-the-art Functional Data Analysis (FDA) statistical methods, which treat high-resolution data as mathematical functions. The 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 non-coding most conserved elements. Intriguingly, the results also suggest that L1 insertions modify local genomic landscape by extending CpG methylation and increasing mononucleotide microsatellite density. Altogether, the findings in this dissertation substantially improved our understanding of L1 integration and fixation preferences, and implied the critical role of TE activities in human health and diseases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Mentors and committee: I would like to thank my graduate advisor Mark Shriver, who has guided me through this journey, both in its good and tough times. You have been a wonderful mentor, I am truly glad that our paths crossed and hope they will overlap again in the future. I am also extremely lucky to have an amazing committee, and I could not possibly make it to the end of the tunnel without the generous support from Ross Hardison, Steve Schaeffer, and Bob Paulson. Lastly, I am truly proud to be part of the legendary Genetics program and the Huck Institutes of the Life Sciences. Thanks to Bob Paulson, Troy Ott and Melissa Rolls for their continuous admin support, as well as the admin assistance from Jean Pierce, Terrie Young, Jennifer Knecht, and Dana Coval-Dinant. I would also like to thank the Huck Institutes of the Life Sciences, Department of Biology, and Department of Anthropology for their funding support during my graduate study at Penn State.

Co-authors and contributors in the L1 project: Thanks to Kateryna Makova, Francesca Chiaromonte and Mark Shriver for their mentorship and guidance throughout this project. Special thanks to Marzia Cremona for being such a great collaborator and also a close friend. Thanks to Zongtai Qi and Robi Mitra for their important contributions to the L1 insertion assay. I would also like to thank Rebeca Campos-Sanchez, Ross Hardison, Belinda Giardine, Anton Nekrutenko, Martin Cech, Dave Bouvier, and Dan Blankenberg for their assistance. Thanks to Wilfried Guiblet and Debmalya Nandy for the helpful discussions.

Lab members: I would like to thank everyone from the Makova Lab and Shriver Lab for their generous support throughout my study at Penn State. I have been extremely lucky to be surrounded by so many talented colleagues, who also happened to be excellent friends – I have learned so much from all of you and hope my presence has created positive influence on you just as yours have affected mine.

Family and friends: I would like to thank my parents, Kexing Zhou and Lin Chen, who has taught me to always follow my heart and pursue happiness with a free spirit. And of course, my lovely wife, Lankun Ma, who has accompanied me for the past six-plus years. You have always been so wise and supportive, calming me down and picking me up for countless times. I could not possibly thank you enough for your contribution and sacrifice - I owe you at least a lifetime of love and joy. Finally, a shout-out to my dear friends: Lingyu Li, Haoyang Jiang, Mei Han, Ziyun Gao, Samarth Rangavittal, Wilfried Guiblet, Monika Cechova, Rahul Vegesna, Edmundo Torres-González, Arslan Zaidi, Lin An, Tao Yang, Hanshu Hong, Benjamin Frempah, and Tejas Mishra. Life is fantastic, and it is just way better when I get spend it with all of you.

Chapter 1

Introduction

Background

Human genome evolution and Transposable Elements

Mutations provide evolution with variants upon which natural selection and random genetic drift operate. Also, mutations are not evenly distributed across the human genome; instead, they vary in rates among different chromosomes and also among different regions of individual chromosomes (Hardison et al. 2003a; Don et al. 2013). Regional variation in mutation rates applies to multiple mutation types such as base substitutions, small insertions and deletions (indels), and insertions of Transposable Elements (TEs)(Hardison et al. 2003b; Makova and Hardison 2015). In particular, TEs are important constituents of the human genome and are considered to play a critical role in human genome evolution. More than 45% of the modern human genome consists of TEs or repetitive sequences that are derived from TEs (Lander et al. 2001a; Cordaux and Batzer 2009). Most TEs are inactive due to their truncated structure or due to the accumulation of mutations, while a small portion of them are still mobile and continue to reshape the landscape of our genome through ongoing transposition (Mills et al. 2007a; Kvikstad and Makova 2010; de Koning et al. 2011; Sotero-Caio et al. 2017). For instance, transpositional events in germline cells have direct mutagenesis effects at multiple scales and influence the genome structure of the offspring, triggering potential developmental disorders (Goodier and Kazazian 2008; Ivics et al. 2009). In addition, TE activities and their products can affect different cellular processes such as cell fate change and immune response. For instance, it has been reported that TE transcripts can trigger the

innate immune response in mammalian cells and regulate the formation of hematopoietic stem and progenitor cells (HSPC) through inflammatory signaling (Ahmad et al. 2018; Lefkopoulos et al. 2020). In previous in vivo and in vitro studies, the genomic distribution of TEs has been investigated along with their functions in different biological processes (Boissinot 2004; Kano et al. 2009; Elbarbary et al. 2016; Zhao et al. 2019). Therefore, TEs have been considered to continuously shape the genome architecture, including its structure and functions, through their sequences, transcripts, and insertion activities (Boissinot et al. 2006; Kejnovsky et al. 2015; Elbarbary et al. 2016).

Classification of Transposable Elements

TEs can be broadly classified into retrotransposons (Class I) and DNA transposons (Class II) according to their transpositional mechanisms (Graur and Li 2000; Bourque et al. 2018). The activity of retrotransposons follows a "copy-and-paste" mechanism, during which an RNA intermediate is reverse-transcribed and integrated to another genome locus (target site), while the original copy (donor site) is preserved (Boeke et al. 1985; Luan et al. 1993). In contrast, the DNA transposons follow a "cut-and-paste" mechanism, whereby a DNA intermediate is involved (Greenblatt and Brink 1963; Rubin et al. 1982). The majority of human TEs came from the activities of retrotransposons, particularly the non-LTR retrotransposons, in which the long terminal repeats (LTRs) are absent (Lander et al. 2001b; Liu et al. 2003). Non-LTR retrotransposons have a critical impact on the human genome at multiple levels, such as genome size expansion, genomic instability and rearrangements, and gene regulation (Lander et al. 2001a; Cordaux and Batzer 2009). While most of the non-LTR retrotransposons have lost mobility and remain inactive, some families are still active in the human genome (Kazazian et al. 1988; Deininger and Batzer 1999; Mills et al. 2007b; Belancio et al. 2008).

DNA Transposons

The DNA transposons occupy approximately 3% of the human genome, while they have lost mobility over 40 million years ago and are no longer active in humans (Lander et al. 2001b; Campos-Sánchez et al. 2014). In contrast, some DNA transposon families have recently been found active in the genome of bats and potentially other species (Pace and Feschotte 2007; Mitra et al. 2013). Several mutagenesis applications in mammalian cells have also been developed from the active DNA transposons, for instance, the Sleeping Beauty and piggyBac systems (Ding et al. 2005; Liu et al. 2005). In addition, the integration of DNA transposons has been studied previously and was shown to have a bias towards actively transcribed genomic regions (Campos-Sánchez et al. 2014).

LTR Retrotransposons

LTR retrotransposons possess two long terminal repeats (LTRs) and transpose via the reverse transcription of an RNA intermediate. LTR elements are commonly found in the eukaryote genome, for instance, they comprise 8% of the human genome and 10% of the mouse genome selection (Havecker et al. 2004; Zeng et al. 2017). Most of the human LTR-retrotransposons have lost mobility, except for a few Endogenous Retroviruses (ERVs), which might still be recently active (Belshaw et al. 2005; Campos-Sánchez et al. 2016). In contrast, the mouse genome harbors much higher LTR activities mainly from three groups: Intracisternal A Particle (IAP), Early Transposon family (ETn), and Mammalian apparent LTR-retrotransposons (MaLRs) (Deininger et al. 2003; Campos-Sánchez et al. 2016). In addition, LTR elements are usually found depleted in

gene-rich regions, potentially due to negative selection (Deininger and Batzer 2002; Medstrand et al. 2002).

SINEs

SINEs are non-LTR retrotransposons and have successfully accumulated in a wide range of mammalian genomes. For instance, *Alu*—a family of SINEs, comprises at least 10% of the human genome with over one million copies (Deininger 2011; Wagstaff et al. 2012a), *Alu* elements have high transpositional activities throughout the genome and have a wide range of effects on genome evolution and gene regulation (Deininger 2011; Wagstaff et al. 2012b). The activities of *Alu*s are facilitated by the transposition machinery of LINEs via shared endonuclease (EN) and reverse transcriptase (RT) (Boeke 1997; Deininger 2011; Wimmer et al. 2011; Elbarbary et al. 2016). Moreover, *Alu*s are frequently located in the GC-rich regions of the genome and have been found responsible for the majority of human diseases caused by TE activities (Deininger 2011; Wagstaff et al. 2012c). Another example is Mammalian-wide Interspersed Repeats (MIRs), which are positively correlated with the presence of gene enhancers and have been proposed to have potential regulatory functions (Matassi et al. 1998; Jjingo et al. 2014).

LINEs

LINE elements belong to the non-LTR retrotransposons and are autonomous, given that they code for the two enzymes (EN and RT) required in the transposition process. LINEs can be up to several kilobases in size, the elements usually contain an internal promoter for RNA polymerase II, a 5′ untranslated region (UTR), two open reading frames (ORFs), and a 3′ terminal poly-A site (Loeb

et al. 1986; Finnegan 1997). The ORF1 protein is an RNA binding protein, and the ORF2 encodes both the EN and RT (Finnegan 1997; Kolosha and Martin 1997; Weiner 2002). LINEs can be classified into five superfamilies (Jockey, L1, R2, RTE, and I), depending on their nature and location of the EN domain. They also have differences in other functional characteristics, as well as in the corresponding host defense systems developed by the genome (Rebollo et al. 2012; Lindič et al. 2013; McLaughlin et al. 2014). For instance, the LINE-2 (L2) and LINE-3 (L3) elements belong to the Jockey superfamily, they have lost mobility in human and are commonly found in conserved genomic regions (Silva et al. 2003; Meyers 2006). In particular, the L2 elements have been proposed to involve in the post-transcriptional gene regulatory networks via miRNAs(Petri et al. 2017). Among all the LINE superfamilies, the most notable group is the Long Interspersed Element-1, abbreviated as LINE-1 or L1. L1 elements are still active in human and have drawn increasing attention due to their critical roles in genome evolution, cellular functions, and human health (Singer 1982; Cordaux and Batzer 2009).

L1s in the human genome

More than 17% of the human genome is occupied by L1s (Singer 1982; Cordaux and Batzer 2009), and their youngest copies are the only active LINE retrotransposons in human (Penzkofer et al. 2017; Feusier et al. 2019). L1s facilitate the activity of 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 (Boeke et al. 1985; 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 2001; Beck et al. 2011). Recent advances in whole-genome sequencing (WGS) have enabled the 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; Eric M. Ostertag and Kazazian 2001; 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).

L1 transposition in germline and somatic cells

L1 transposition has mutagenesis effects as a result of both direct insertions and genetic variations induced by the insertional events (Cordaux and Batzer 2009; Payer and Burns 2019). It has been previously reported that germline de novo insertions of L1s and dysregulation of in the human genome can lead to a variety of genetic disorders(Goodier and Kazazian 2008; Belancio et al. 2009; Beck et al. 2011; Payer and Burns 2019). For instance, the L1 insertions in exon 14 of the factor VIII gene were found to cause Haemophilia A in patients (Kazazian et al. 1988). Another example is Duchenne/Becker muscular dystrophy, which can be caused by the partial exonization of L1 copy disrupting the open reading frame of the dystrophin gene (Gonçalves et al. 2017).

L1 insertions are also frequently found in somatic tissues, and can potentially play important roles in the developmental processes (Muotri et al. 2005; Kano et al. 2009) and behavior learning(Baillie

et al. 2011; Bedrosian et al. 2018). For instance, L1 activities can assist in forming brain plasticity in response to environmental stress via somatic variations in the neurons and other regulatory functions (Baillie et al. 2011; Bedrosian et al. 2018). Moreover, frequent somatic L1 retrotransposition events have also been found 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).

Use L1s as genetic tools

L1 transposition 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). Therefore, in addition to providing information on genome functions and evolution, a detailed understanding of L1 transpositional activity and integration preferences can further facilitate the use of L1s as a mutagenesis tool in molecular genetic studies. For example, 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).

Previous studies on L1 transposition

The chromosomal distribution of L1 elements in the human genome with respect to several genomic features has been investigated in some previous studies. For instance, densities of fixed L1 elements of different evolutionary ages were found to vary by chromosome (Kvikstad and Makova 2010), 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), while the distribution of recently inserted elements was found to be influenced by chromatin state (Singer 1982; Sultana et al. 2017; Sultana et al. 2019). These findings imply that, while 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).

Therefore, addressing the human L1 dynamics can aid in understanding the structure, function and evolution of our genome, which also has significant implications in human health. However, to date, there has not been a high-resolution, genome-wide study of L1s in an evolutionary framework, through which the insertion and fixation preferences of the elements can be elucidated. In addition, the interactions between L1s and local genome landscape have largely remained unclear.

Outline of Chapters

The long-term goal of the study is to characterize the transposition dynamics of TEs and to further understand their contribution to the structure, function, and evolution of the human genome. The specific objectives of this dissertation are to develop a framework to investigate the genome-wide distribution of L1s at a high-resolution and to address their interactions with local genomic landscape features. I address those objectives in the following three chapters, under the working hypothesis: *Different local genomic landscape features contribute in various ways to the L1 insertion and fixation preferences in the human genome*.

Chapters two focuses on the study design to investigate human L1 transposition dynamics using the methods of Functional Data Analysis (FDA). I introduced the experimental design and analytical framework of the study. The analysis leverages three large L1 datasets representing distinct evolutionary distances, as well as a comprehensive collection of high-resolution genomic features. I also present a successful application of FDA framework in the high-resolution genomics research (Cremona et al. 2018). The framework allows us to effectively address the scale and location of the features' effects on specific genomic intervals. It can be applied to a wide range of topics in future genomics research.

Chapter three focuses on the discussion of the correlation between human L1 activities and the local genomic landscape, as well as the biological models of L1 insertion and fixation. I examine the hypothesis that different local genomic landscape features contribute in various ways to the L1 insertion and fixation preferences in the human genome. In particular, I investigate the genome-wide distribution of L1s at different evolutionary time scales, and correlate the insertion and fixation preferences of the elements with local genomic features thus building an integrative model of L1 transposition dynamics. I demonstrate that the genomic distribution of human L1s is driven by the local genomic landscape, and our FDA analysis reveals the potential mechanisms through

which regional genomic characteristics influence new element insertions and their abilities to fix in the genome (Chen et al. 2020).

Chapter four extends the work in previous chapters by examining the robustness of the study design and reproducibility of the findings. Specifically, I validate different aspects of the study design in chapter two with computational experiments using multiple publicly available datasets. I also apply our analytical framework to several recently published *de novo* L1 datasets and different subsets of the polymorphic L1s to test the reproducibility and generalizability of the findings in previous chapters.

In the final chapter, I summarize the results from the previous chapters and further address the contribution of TEs to the architecture and evolution of the human genome. I also discuss the significance of this work and its broader impact on evolutionary biology, genomic research, and medicine. Finally, I point out several directions through which the methods and results from the current study can be extended in the future.

References

1. Ahmad S, Mu X, Yang F, Greenwald E, Park JW, Jacob E, Zhang C-Z, Hur S. 2018. Breaching Self-Tolerance to Alu Duplex RNA Underlies MDA5-Mediated Inflammation. Cell 172:797–810.e13.

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. 103:18662–18667.

3. 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 479:534–537.

4. Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. Annu. Rev. Genet. 42:587–617.

5. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. Annu. Rev. Genomics Hum. Genet. 12:187–215.

6. Bedrosian TA, Quayle C, Novaresi N, Gage FH. 2018. Early life experience drives structural variation of neural genomes in mice. Science 359:1395–1399.

7. Belancio VP, Deininger PL, Roy-Engel AM. 2009. LINE dancing in the human genome: transposable elements and disease. Genome Med. 1:97.

8. Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res. 18:343–358.

9. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): implications for present-day activity. J. Virol. 79:12507–12514.

10. Boeke JD. 1997. LINEs and Alus—the polyA connection. Nat. Genet. 16:6.

11. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. 1985. Ty elements transpose through an RNA intermediate. Cell 40:491–500.

12. Boissinot S. 2004. The Insertional History of an Active Family of L1 Retrotransposons in Humans. Genome Research 14:1221–1231.

13. 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. 103:9590–9594.

14. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. Genome Biol. 19:199.

15. 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:e1004956.

16. Campos-Sánchez R, Kapusta A, Feschotte C, Chiaromonte F, Makova KD. 2014. Genomic landscape of human, bat, and ex vivo DNA transposon integrations. Mol. Biol. Evol. 31:1816–1822.

17. Chen D, Cremona MA, Qi Z, Mitra RD, Chiaromonte F, Makova KD. 2020. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. Mol. Biol. Evol. Available from: http://dx.doi.org/10.1093/molbev/msaa194

18. Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10:691–703.

19. 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 34:2289–2291.

20. Deininger P. 2011. Alu elements: know the SINEs. Genome Biol. 12:236.

21. Deininger PL, Batzer MA. 1999. Alu repeats and human disease. Mol. Genet. Metab. 67:183–193.

22. Deininger PL, Batzer MA. 2002. Mammalian retroelements. Genome Res. 12:1455–1465.

23. Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. 2003. Mobile elements and mammalian genome evolution. Curr. Opin. Genet. Dev. 13:651–658.

24. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell 122:473–483.

25. Don PK, Ananda G, Chiaromonte F, Makova KD. 2013. Segmenting the human genome based on states of neutral genetic divergence. Proc. Natl. Acad. Sci. U. S. A. 110:14699–14704.

26. Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. Science 351:aac7247.

27. Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. 21:985–990.

28. 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. 29:1567–1577.

29. Finnegan DJ. 1997. Transposable elements: how non-LTR retrotransposons do it. Curr. Biol. 7:R245–R248.

30. Gonçalves A, Oliveira J, Coelho T, Taipa R, Melo-Pires M, Sousa M, Santos R. 2017. Exonization of an Intronic LINE-1 Element Causing Becker Muscular Dystrophy as a Novel Mutational Mechanism in Dystrophin Gene. Genes. Available from: http://dx.doi.org/10.3390/genes8100253

31. Goodier JL, Kazazian HH Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35.

32. 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.

33. Graur D, Li WH. 2000. Fundamentals of molecular evolution. Sinauer Associations. Inc. Sunderland MA.

34. Greenblatt IM, Brink RA. 1963. Transpositions of Modulator in maize into divided and undivided chromosome segments. Nature 197:412–413.

35. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003a. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

36. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003b. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

37. Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. Genome Biol. 5:225.

38. Ivics Z, Li MA, Mátés L, Boeke JD, Nagy A, Bradley A, Izsvák Z. 2009. Transposon-mediated genome manipulation in vertebrates. Nat. Methods 6:415–422.

39. 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. Mob. DNA 5:14.

40. 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. 23:1303–1312.

41. Kazazian HH Jr, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164–166.

42. Kazazian HH, Moran JV. 1998. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19:19–24.

43. Kejnovsky E, Tokan V, Lexa M. 2015. Transposable elements and G-quadruplexes. Chromosome Res. 23:615–623.

44. Kolosha VO, Martin SL. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc. Natl. Acad. Sci. U. S. A. 94:10155–10160.

45. 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:e1002384.

46. Konkel MK, Walker JA, Batzer MA. 2010. LINEs and SINEs of primate evolution. Evol. Anthropol. 19:236–249.

47. 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 390:28–38.

48. Kvikstad EM, Makova KD. 2010. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. Genome Res. 20:600–613.

49. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001a. Initial sequencing and analysis of the human genome. Nature 409:860–921.

50. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001b. Initial sequencing and analysis of the human genome. Nature 409:860–921.

51. Lefkopoulos S, Polyzou A, Derecka M, Bergo V, Clapes T, Cauchy P, Jerez-Longres C, Onishi-Seebacher M, Yin N, Martagon-Calderón N-A, et al. 2020. Repetitive Elements Trigger RIG-I-like Receptor Signaling that Regulates the Emergence of Hematopoietic Stem and Progenitor Cells. Immunity 53:934–951.e9.

52. Lindič N, Budič M, Petan T, Knisbacher BA, Levanon EY, Lovšin N. 2013. Differential inhibition of LINE1 and LINE2 retrotransposition by vertebrate AID/APOBEC proteins. Retrovirology 10:156.

53. Liu G, Geurts AM, Yae K, Srinivasan AR, Fahrenkrug SC, Largaespada DA, Takeda J, Horie K, Olson WK, Hackett PB. 2005. Target-site preferences of Sleeping Beauty transposons. J. Mol. Biol. 346:161–173.

54. Liu G, NISC Comparative Sequencing Program, Zhao S, Bailey JA, Sahinalp SC, Alkan C, Tuzun E, Green ED, Eichler EE. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res. 13:358–368.

55. 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 553:228–232.

56. Loeb DD, Padgett RW, Hardies SC. 1986. The sequence of a large L1Md element reveals a tandemly repeated 5'end and several features found in retrotransposons. and Cellular Biology. Available from: https://mcb.asm.org/content/6/1/168.short

57. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595–605.

58. Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 16:213–223.

59. Matassi G, Labuda D, Bernardi G. 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. FEBS Lett. 439:63–65.

60. 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:e1004531.

61. 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. 12:1483–1495.

62. 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.

63. 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. 8:3209–3225.

64. 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.

65. Mills RE, Andrew Bennett E, Iskow RC, Devine SE. 2007a. Which transposable elements are active in the human genome? Trends Genet. 23:183–191.

66. Mills RE, Andrew Bennett E, Iskow RC, Devine SE. 2007b. Which transposable elements are active in the human genome? Trends Genet. 23:183–191.

67. Mitra R, Li X, Kapusta A, Mayhew D, Mitra RD, Feschotte C, Craig NL. 2013. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc. Natl. Acad. Sci. U. S. A. 110:234–239.

68. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. 1996. High frequency retrotransposition in cultured mammalian cells. Cell 87:917–927.

69. 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. 102:18644–18648.

70. Ostertag EM, Kazazian HH Jr. 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. 11:2059–2065.

71. Ostertag EM, Kazazian HH Jr. 2001. Biology of Mammalian L1 Retrotransposons. Annu. Rev. Genet. 35:501–538.

72. Pace JK 2nd, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17:422–432.

73. Payer LM, Burns KH. 2019. Transposable elements in human genetic disease. Nat. Rev. Genet. Available from: http://dx.doi.org/10.1038/s41576-019-0165-8

74. 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:D68–D73.

75. Petri R, Brattås PL, Sharma Y, Jönsson ME, Pircs K, Bengzon J, Jakobsson J. 2017. LINE-2 transposable elements are a source for functional human microRNAs and target sites. Available from: http://dx.doi.org/10.1101/218842

76. 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. 21:1115–1127.

77. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager DL. 2012. Epigenetic interplay between mouse endogenous retroviruses and host genes. Genome Biol. 13:R89.

78. Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea G-O, Muñoz-Lopez M, Jesuadian JS, Kempen M-JHC, Carreira PE, Jeddeloh JA, et al. 2017. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res. 27:1395–1405.

79. Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell 29:987–994.

80. Scott EC, Devine SE. 2017. The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses. Available from: http://dx.doi.org/10.3390/v9060131

81. 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. 82:1–18.

82. Singer MF. 1982. SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28:433–434.

83. Sotero-Caio CG, Platt RN 2nd, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biol. Evol. 9:161–177.

84. 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. 131:299–305.

85. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger 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 74:555–570.e7.

86. Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat. Rev. Genet. 18:292–308.

87. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, Roy-Engel AM. 2012a. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. PLoS Genet. 8:e1002842.

88. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, Roy-Engel AM. 2012b. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. PLoS Genet. 8:e1002842.

89. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, Roy-Engel AM. 2012c. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. PLoS Genet. 8:e1002842.

90. Weiner AM. 2002. SINEs and LINEs: the art of biting the hand that feeds you. Curr. Opin. Cell Biol. 14:343–350.

91. Wimmer K, Callens T, Wernstedt A, Messiaen L. 2011. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS Genet. 7:e1002371.

92. 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. 30:64–77.

93. Zeng F-C, Zhao Y-J, Zhang Q-J, Gao L-Z. 2017. LTRtype, an Efficient Tool to Characterize Structurally Complex LTR Retrotransposons and Nested Insertions on Genomes. Front. Plant Sci. 8:402.

94. 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:e1008043.

Chapter 2

Study design of a genome-wide investigation of human L1 transposition

Most data in this chapter are published as a research article by Chen, D., Cremona, M.A., Qi, Z., Mitra, R.D., Chiaromonte, F. and Makova, K.D., 2020. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. *Molecular Biology and Evolution.* D.C. and M.C. contributed equally to this work. Reuse of content from the publication for thesis is in compliance of the journal policies.

Background

The main goal of this project is to perform a high-resolution, genome-wide study of L1s in an evolutionary framework. Specifically, we aim to address the insertion and fixation preferences of L1s with respect to the local genome landscape. 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 sub-regions can be treated as 'curves' along each chromosome. On the 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 2005), a branch of statistics specifically developed to study data described as curves (mathematical functions), that was only recently introduced into

genomics research (Zhang et al. 2014; Campos-Sánchez et al. 2016; Cremona et al. 2018; Guiblet et al. 2018; Cremona et al. 2019).

Design of study and methods

Design of study

To examine the working hypothesis that different local genomic landscape features contribute in various ways to the L1 insertion and fixation preferences in the human genome, we considered three datasets comprising integrations of L1 elements at different evolutionary time points; namely, *de novo*, polymorphic, and human-specific L1s (Table S1). *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 distributions of human-specific vs. *de novo* L1s should highlight fixation preferences, and investigating polymorphic L1s might provide additional insights on the interplay between integration and fixation. We then collected 49 genomic landscape features cross-referenced from other studies and analyzed them using the FDA statistical methods, which treat high-resolution data as mathematical functions. Specifically, we contrasted the genomic feature signals in the flanking regions of three L1 datasets and L1-depleted control regions via six pairwise comparisons (Fig. 2-1). Three advanced FDA methods (Interval-Wise Testing, single Functional Logistic Regression, and multiple Logistic Regression) were implemented in the workflow to identify the differences in genomic feature signals at high resolution, quantify the contribution of each feature, and address the joint effect from multiple features (Fig. 2-1). Through this study design, we performed the first

genome-wide analysis of L1 transposition dynamics in an evolutionary framework and used the FDA to leverage an extensive list of genomic landscape features at high resolution.

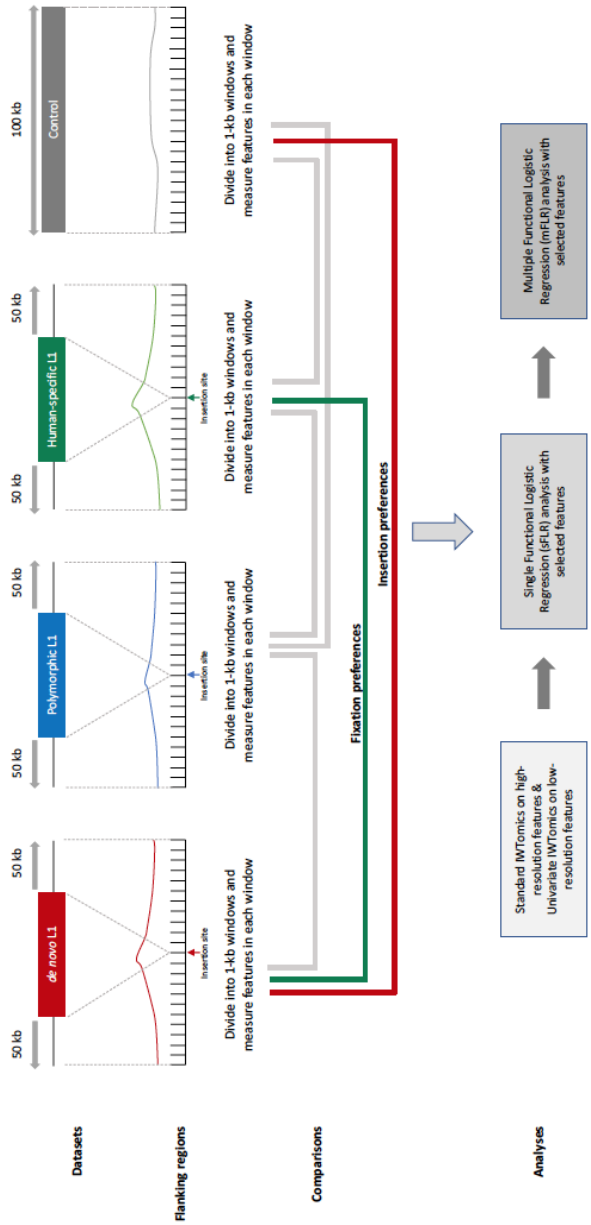


Figure 2-1. Functional Data Analysis (FDA) workflow. Illustration of the FDA workflow used in the study. The 100-kb L1 regions were constructed taking 50-kb in each direction of the insertion sites, and the control regions were constructed as 100-kb non-overlapping intervals with low coverage (<7%) of 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 FDA analyses. Curves in different groups (different types of L1s, or each L1 type vs. controls) were then compared using IWTomics (Interval-Wise Testing for omics data) and Functional Logistic Regression. The control regions in this study contain less than 7% coverage by all annotated L1 elements.

Experimental and computational methods

Collection of L1 datasets

We harvested the *de novo* L1s from an induced L1 insertion experiment conducted in the cultured human kidney stem cell line HEK-293T (Figs. 2-2 and S1), 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 (Details are shown in the section "*in vivo L1 insertion experiment*"). By analyzing sequencing data from this experiment, we identified 17,037 *de novo* L1 insertions (Figs. S2; S3A). 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 (Figs. S2; S3B). The polymorphic L1 dataset 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)(Table S5). We also 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) using the RepeatMasker (Smit et al. 2015) track of GRCh37/hg19 from the UCSC Genome Browser (Karolchik et al. 2004). We performed the following filtering: we conservatively selected only those L1HSs that were absent from the genomes of non-human 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) (Figs. S2; S3C).

For each of these three L1 datasets, we only considered elements on autosomes and chromosome X for the subsequent analyses (Table S1).

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 full-length ORFeus-Hs element (An et al. 2011) the human L1 element, 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 (Fig. S18). 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. 2-2).

Cell transfection and FACS: 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 Dubecco's Modified Eagle Media (DMEM; 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 μg/ml. After 3 days of puromycin selection, cells were washed in 1×phosphate-buffered saline and sorted by fluorescence-activated cell sorting (FACS). 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_fixORFeous: AATGATACGGCGACCGCCGAGATCTACACAGCTCTGTAACCATTAGCTGCAATAAAC AAGTTAAC;iPCR_R_fixORFeus:CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACT GGAGTTCAGACGTGTGC) that anneal at a complementary region of the pld225 plasmid to amplify the genomic regions flanking L1 insertion loci (Figs. 2-2; S1A). The adapter sequences (Adapter (P5) added on the forward iPCR primer: AATGATACGGCGACCGCCGAGATCTACAC; adapter (P7) added on the reverse iPCR primer: CAAGCAGAAGACGGCATACGAGAT), 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. 2-2); read 2 should lead to a more precise location that 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, Figs. 2-2; S1), 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 (Fig. S1). 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 bps–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 3).

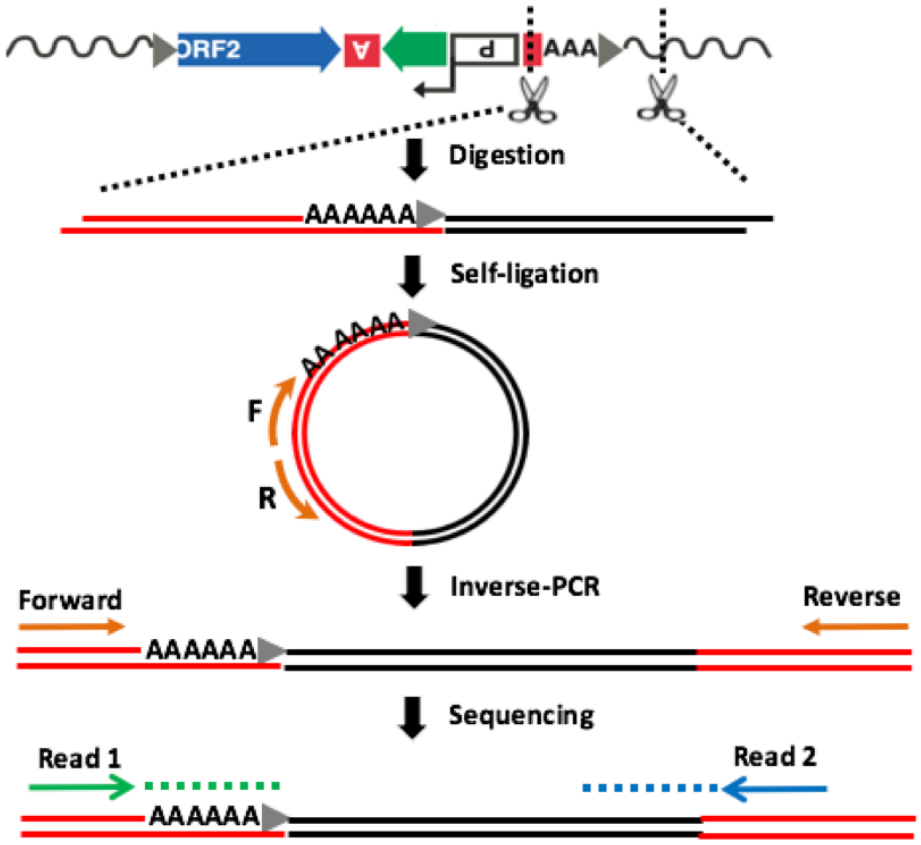


Figure 2-2. Identification of *in vivo de novo* L1 insertions by Inverse-PCR. Vectors containing both a synthetic human L1 element (full-length synthetic ORFeus-Hs, see Methods) and Green Fluorescent Protein (GFP) were transfected into cultured cells. The vectors were marked by two restriction enzyme sites (*MspI* and *TaqI*) and 14 different barcodes of four to six nucleotides. While 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 pair-end Illumina sequencing. Figure designed by Zongtai Qi.

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 (Fig. S4). We also compared the distribution of distances between L1 elements of two different types with a random expectation (Fig. S5). In particular, for each of the three L1 datasets (*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 dataset 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 purpose of bootstrapping was to provide comparable sample sizes across different L1 sets and the number of subsamples selected as a balance of sufficient statistical power and reasonable computational time. The comparison was visualized using cumulative distribution plots (Fig. S4A-C) and quantile-quantile (Q-Q) plots (Fig. S4D-F). In addition, to compare distance distributions across the three L1 datasets, we considered a 'normalized' cumulative distribution of the distances between L1 elements. Specifically, we first subsampled 900 elements from each L1 dataset, 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. We also analyzed the distances between L1 elements from different datasets using the same procedure and plots (Fig. S5A-F), 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).

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: ENCSR000AMI), 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 (Table S3). 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 (Table S3).

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 (Table S1). 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, in order 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 non-overlapping human-specific 100-kb L1 flanking regions, after removing genome assembly gaps and blacklisted regions (Table S3). 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 analysis (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 datasets. 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 dataset 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 dataset (Ewing and Kazazian 2011) might introduce bias). Yet we excluded them and their flanks to obtain cleaner controls. Finally, in order 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 in order to obtain a number of control regions of the same order of magnitude as in each of the three L1 datasets.

We also considered the fact that some of the 100-kb flanking regions from different L1 datasets (e.g. *de novo* L1s and human-specific L1s) might overlap, making the datasets not completely independent. We performed IWTomics analysis (see section 'Interval-wise testing with IWTomics') both on the complete datasets and after removing all the overlapping regions among different datasets (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 datasets in our analyses, in order 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 the features obtained from ChIP-Seq experiments (histone modifications, DNase hypersensitive sites, and CTCF motifs) were measured as 'signals', i.e. 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 (Fig. S6). 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) in order to reduce multicollinearity issues in the multiple regression analysis (see below).

Table **2-1**. Genomic landscape features and their contributions in single and multiple Functional Logistic Regressions.

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 sign.	2.30	Not sel.
Microsatellites	Mononucl. microsats	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	UCSC Genome Browser	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	(Plohl 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**

Chromatin = chromatin structure, Transcription = transcription regulation and gene expression, “Res.” = Resolution, “W.ave” = weighted average, “Not sign.” = features that showed no significant differences in IWTomics tests; “Not sel.” = features that were not selected in the final mFLR models (potentially due to interdependencies among features). 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 (Fig. S6).

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 Interval-Wise Testing for omics data (IWTomics) (Pini and Vantini 2016; Cremona et al. 2018). IWTomics is a non-parametric 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 Fig. S7 for an example of IWTomics complete output). The test is fully non-parametric 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 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 (Figs. 2-1 and S8). 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. In order 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 non-zero 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 kb, 50 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 multi-quantile difference (the sum of the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} quantile differences)–in order 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 (multi-quantile differences produced similar results, while 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 (Figs. 2-1 and S9).

Since the *de novo* L1 dataset was substantially larger than the polymorphic L1, human-specific L1 and control datasets, we randomly subsampled 1,000 *de novo* L1 regions in order 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, scale, etc.; data not shown) which we

summarized using pointwise medians of the adjusted *p*-value curves (Figs. S7 and S8; 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 single Functional Logistic Regression models (sFLR). 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 genomics 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, i.e. we used the transformation $log(x + s)$, and we selected $s \, \epsilon\{1, 10^{-1}, \ldots, 10^{-10}\}$ in order 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 $s \, \epsilon\{2,4, \ldots, 22\}$). 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

$$\text{logit}\left(\mathbb{E}\left[\text{Y} \mid x_F(t)\right]\right)=\ln\left(\frac{p_F}{1-p_F}\right)=\beta_0+\frac{1}{\sqrt{|\text{I}_F|}}\int_{\text{I}_F}\beta_F(t)x_F(t)\,dt$$

$$\text{logit}\left(\mathbb{E}\left[\text{Y} \mid Z_F\right]\right)=\ln\left(\frac{p_F}{1-p_F}\right)=\beta_0+\beta_F Z_F$$

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 non-localized (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 *fregre.glm* from the R package *fda.usc* (Febrero-Bande and 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, i.e. with

$$\mathrm{DE} = \mathrm{R}^2_{psuedo} = \frac{D_{null} - D_{model}}{D_{null}}$$

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 10 times—using the same 10 random subsamples of *de novo* L1 flanking regions generated for the IWTomics analysis. Again, results from the 10 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 (Fig. S10).

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 multiple Functional Logistic Regression (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 (for example, 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:

$$\text{logit}(\mathbb{E}[Y|x_1, \ldots, x_r, x_{r+1}(t), \ldots, x_{r+s}(t)]) = \ln\left(\frac{p}{1-p}\right)$$

$$= \beta_0 + \sum_{j=1}^{r} \beta_j x_j + \sum_{j=r+1}^{r+s} \int_{-50}^{50} \beta_j(t) x_j(t) dt$$

where $x_1, \ldots, x_r$ are the r scalar predictors, $x_{r+1}(t), \ldots, x_{r+s}(t)$ 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 non-significant 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. In order 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*):

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

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

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

The mFLR model could therefore be rewritten as:

$$\text{logit}(\mathbb{E}[Y|x_1, \ldots, x_r, x_{r+1}(t), \ldots, x_{r+s}(t)]) = \ln\left(\frac{p}{1-p}\right)$$

$$= \beta_0 + \sum_{j=1}^{r} \beta_j x_j + \sum_{j=r+1}^{r+s} \boldsymbol{b}_j^T \boldsymbol{J}_{\boldsymbol{\phi}} \boldsymbol{w}_j$$

where

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

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

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

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

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

where $l(b)$ 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 Fig. S12).

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 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:

$$\text{DE} = \text{R}^2_{psuedo} = \frac{D_{null} - D_{model}}{D_{null}}$$

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 Relative Contribution to the Deviance Explained (RCDE):

$$RCDE = \frac{(D_{null} - D_{model}) - (D_{null} - D_{red\ model})}{(D_{null} - D_{model})}$$

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

Data availability and contribution

We have set up a github repository (https://github.com/makovalab-psu/L1_Project) and included the chromosomal coordinates of de novo, polymorphic, and human-specific L1s analyzed in this study. The repository also contains the computational pipelines and code, along with the corresponding intermediate files (.RData) used to generate the results. Marzia Cremona helped with the customization of *IWTomics* package and developed the R code for L1 distance analysis and mFLR. Zongtai Qi and Robi Mitra performed the *de novo* L1 integration experiment. Thanks to Hidetoshi Matsui for providing R code for variable selection in mFLR models, and Jef Boeke for providing the plasmid pld225 containing the synthetic full-length ORFeus-Hs element. The raw sequencing reads from the *de novo* L1 insertion experiment were uploaded to the Short Read Archive (SRA) under the accession number PRJNA640178.

Funding

This project was supported by the funds made available through the Clinical and Translational Sciences Institute, Institute for CyberScience, and 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. The work from Robert Mitra's lab was supported by NIH grants RF1 MH117070 and R01GM123203.

References

1. 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:W537–W544.
2. Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9:9354.
3. 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. Mob. DNA 2:2.
4. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.
5. 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. 19:837–844.
6. Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. 17:915–928.
7. 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 478:343–348.
8. 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.

9. 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:D762–D769.
10. Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky 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:D383–D391.
11. 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.
12. Consortium T 1000 GP, The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68–74. Available from: http://dx.doi.org/10.1038/nature15393
13. 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 34:2289–2291.
14. Cremona MA, Xu H, Makova KD, Reimherr M, Chiaromonte F, Madrigal P. 2019. Functional data analysis for computational biology. Bioinformatics 35:3211–3213.
15. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.
16. Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. 21:985–990.
17. Febrero-Bande M, de la Fuente MO. 2012. Statistical Computing in Functional Data Analysis: TheRPackagefda.usc. Journal of Statistical Software 51. Available from: http://dx.doi.org/10.18637/jss.v051.i04

18. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. Nat. Protoc. 7:1728–1740.
19. 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:R86.
20. Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky 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. 28:1767–1778.
21. 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 6:283–289.
22. 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:D493–D496.
23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.
24. 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 467:1099–1103.
25. Krueger F, Kreck B, Franke A, Andrews SR. 2012. DNA methylome analysis using short bisulfite sequencing data. Nat. Methods 9:145–151.
26. 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. 22:1813–1831.

27. 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.

28. 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 462:315–322.

29. Matsui H. 2014. Variable and boundary selection for functional data via multiclass logistic regression modeling. Comput. Stat. Data Anal. 78:176–185.

30. Meier L, Van De Geer S, Bühlmann P. 2008. The group lasso for logistic regression. J. R. Stat. Soc. Series B Stat. Methodol. 70:53–71.

31. 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 146:1029–1041.

32. Morrissey CS. 2013. Understanding the epigenetics of erythroid differentiation through the power of deep sequencing.

33. 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. 40:1124–1129.

34. Ovchinnikov I, Rubin A, Swergold GD. 2002. Tracing the LINEs of human evolution. Proc. Natl. Acad. Sci. U. S. A. 99:10522–10527.

35. 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.

36. Pini A, Vantini S. 2016. The interval testing procedure: A general framework for inference in functional data analysis. Biometrics 72:835–845.

37. Plohl M, Prats E, Martinez-Lage A, Gonzalez-Tizon A, Mendez 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. 277:19839–19846.

38. Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics 47:11.12.1–11.12.34.

39. Ramsay J, Silverman BW. 2005. Functional Data Analysis. Springer Science & Business Media

40. Rio DC, Clark SG, Tjian R. 1985. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. Science 227:23–28.

41. 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. 20:761–770.

42. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

43. Sekhon JS. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: TheMatchingPackage forR. J. Stat. Softw. Available from: http://dx.doi.org/10.18637/jss.v042.i07

44. Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013--2015.

45. 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:e1002154.

46. Walker JA, Konkel MK, Adrian M, Stewart C, Kural D, Stro MP, Urban AE, Grubert F, Lam HYK, Lee W-P, et al. 2011. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. Available from: http://dx.doi.org/10.1371/journal.pgen.1002236

47. 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. 27:323–329.

48. Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Series B Stat. Methodol. 68:49–67.

49. 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 6:1–11.

50. Zhang F, Boerwinkle E, Xiong M. 2014. Epistasis analysis for quantitative traits by functional regression model. Genome Res. 24:989–998.

51. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9:R137.

Chapter 3

Human L1 transposition dynamics unraveled with Functional Data Analysis

Most data in this chapter are published as a research article by Chen, D., Cremona, M.A., Qi, Z., Mitra, R.D., Chiaromonte, F. and Makova, K.D., 2020. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. *Molecular Biology and Evolution.* D.C. and M.C. contributed equally to this work. Reuse of content from the publication for thesis is in compliance of the journal policies.

Summary of Analysis

As described in the previous chapter, we designed a genome-wide study of human L1 transposition dynamics, using three large datasets of L1s 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 also analyzed an extensive list of high-resolution genomic features to characterize the landscapes correlated with L1 integration and fixation. In this chapter, 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. Here we summarize the results and discuss the biological models of human L1 transpositional dynamics in detail.

Results

L1 elements are not randomly distributed in the human genome

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 datasets. Karyotype plots (Fig. S2) and chromosome-specific element densities (Fig. S3) 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 datasets considered, the distribution of distances between L1 elements was far from random (Figs. 3-1A and S4). In particular, L1 elements from the same dataset were closer to each other compared to random expectation ($p=10^{-16}$ for *de novo* L1s, $p=1.5\times10^{-5}$ for polymorphic L1s, and $p=9.7\times10^{-11}$ for human-specific L1s, Kolmogorov–Smirnov test; see Materials and Methods). Furthermore, the analysis of distances between L1s from different datasets (Fig. 3-1B) 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 (Figs. 3-1B and S5). Notably, the distribution of *de novo* L1 insertions appears non-random also when considering *de novo* L1 datasets generated in other recent studies (Flasch et al. 2019; Sultana et al. 2019) (Table S6; Fig. S16).

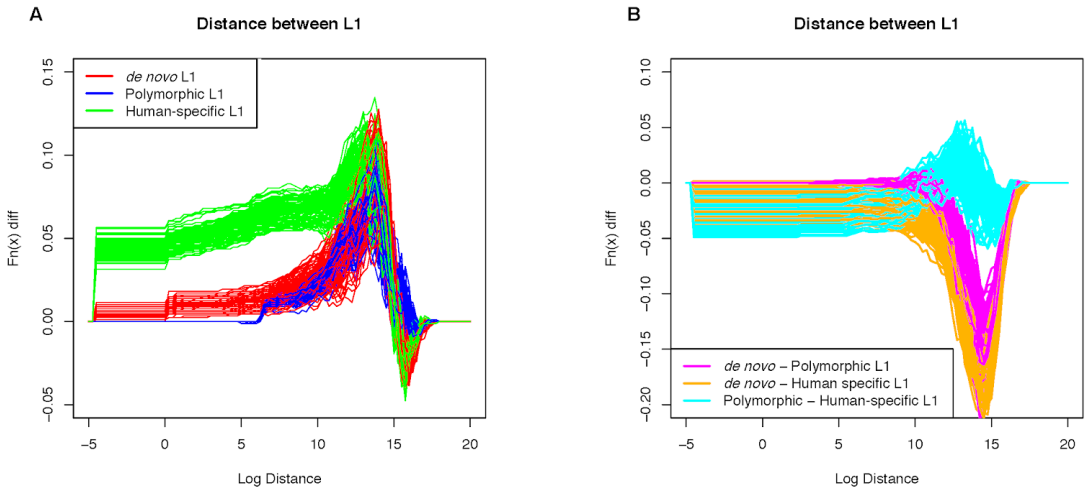


Figure 3-1. Distribution of distances between L1 elements. A. Differences between observed and expected cumulative distributions of the distances between L1 elements of the same type (*de novo*, polymorphic, or human-specific). **B.** Differences between observed and expected cumulative distributions 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 to random expectation, negative differences indicate larger distance compared to random expectation.

Characterizing local landscape with various genomic features

To understand the determinants of the (non-random) 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), UCSC Genome Browser (Karolchik et al. 2004), etc.) and results from previous studies (see Material and Methods), we collected data on 49 quantitative genomic features that may influence L1 integration and fixation dynamics (Table 1 and Table S2). 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 (hESC expression), 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, most conserved elements, etc.) were not cell-line specific, and we extracted most of the other features (e.g. Histone modifications, DNA methylation, etc.) from hESC.

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. 2-1). We excluded regions overlapping with unsequenced gaps (Kent et al. 2002) and repetitive regions with artifactual ChIP-seq or DNase-seq signals (Table S3; Materials and Methods), 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 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 (Fig. S6) and were excluded from subsequent analyses. Thus, a total of 41 high-resolution and five low-resolution features were retained.

Pairwise comparisons of high-resolution features with Functional Data Analysis

To capture multi-scale (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**) (Febrero-Bande and de la Fuente 2012)(Ramsay and Silverman 2005), using the low- and high-resolution features that were significant according to IWTomics test and the same six comparisons (Fig. 2-2). The discriminatory strength of each feature was quantified with pseudo-R^2s from these sFLRs. *Fourth*, to quantify joint effects of multiple features, many of which can interact and are correlated according to our clustering analysis (Fig. S6), we built multiple FLRs (**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 2-1; S2). 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 vs. controls (Figs. 3-2A and S7; Table 1) and of human-specific vs. *de novo* L1 flanks (Figs. 3-2B and S7; 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 Appendix (Figs. S8-S9 and Table S2).

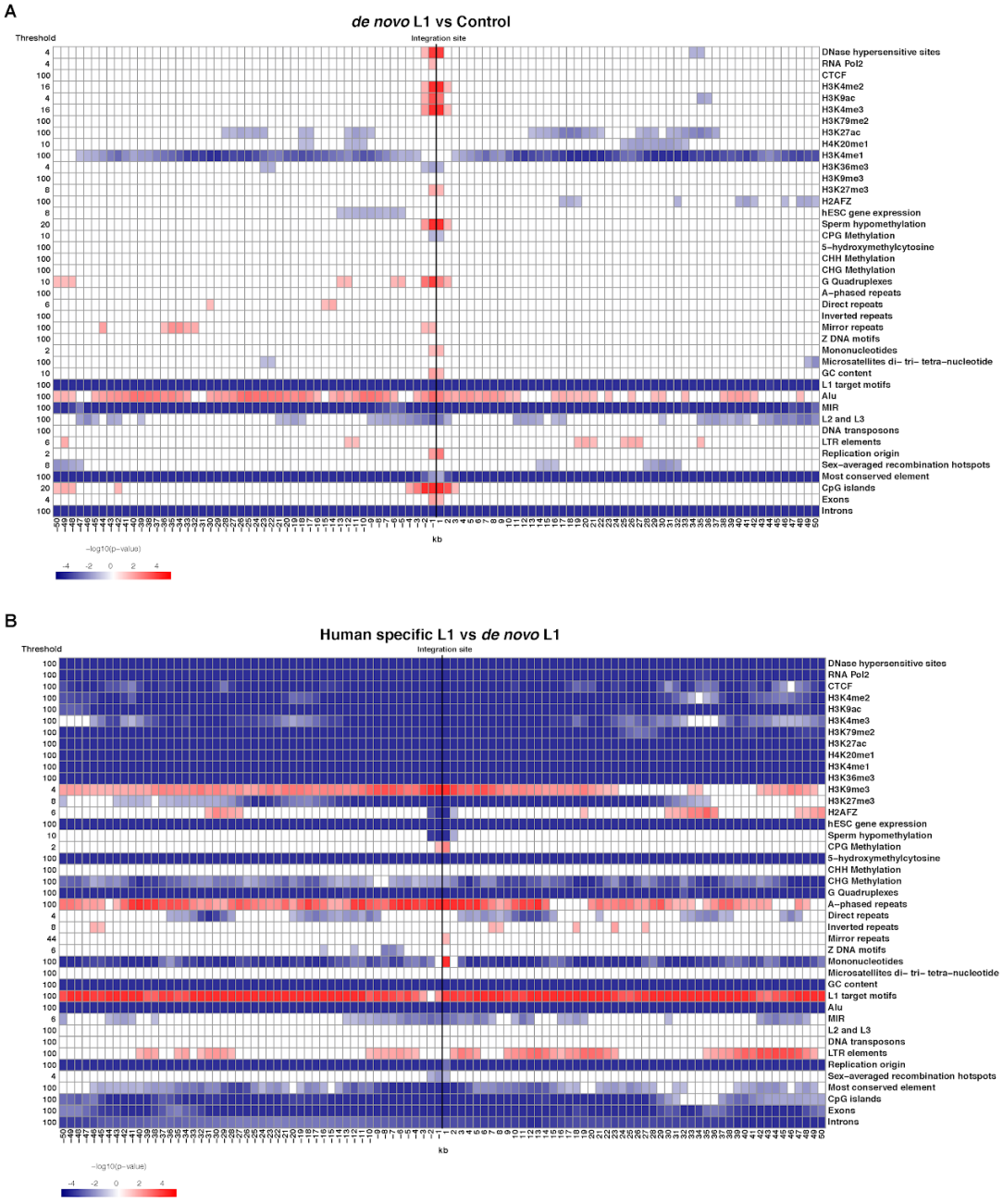


Figure 3-2. Summary of IWTomics results for individual high-resolution features. A. *De novo* L1 flanking regions vs. control regions. **B**. Human-specific L1 vs. *de novo* L1 flanking regions. The X-axis represents the position analyzed within the 100-kb flanking regions of L1 elements (or 100-kb control 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 log10 scale. White: nonsignificant difference (*p*-value>0.05). Red: significant difference, with over-representation of the feature. Blue: significant difference, with under-representation of the feature. The selected scale thresholds corresponding to the adjusted *p*-value curves are noted on the left (column 'Threshold'). The control regions in this study contain less than 7% coverage by all annotated L1 elements.

de novo L1 insertion landscape

To investigate insertion preferences, we compared genomic features in the flanks of *de novo* L1s vs. control regions. The univariate IWTomics analysis (Fig. 3-3) contrasting low-resolution features suggested that *de novo* L1 insertions are significantly and positively associated with early replication timing (p=0.0001; Fig. 3-3C), and significantly and negatively associated with telomere hexamers (p=0.0001; Fig. 3-3E).

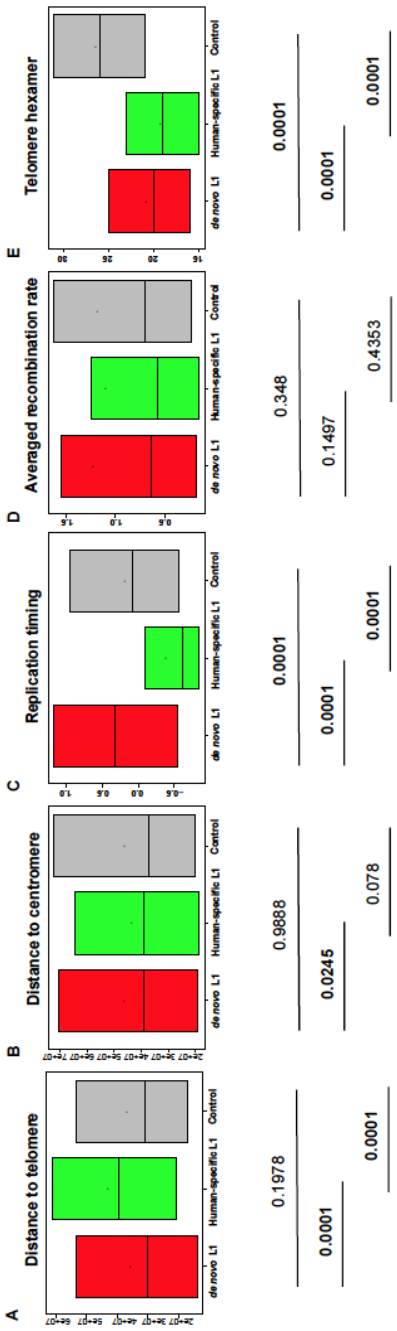


Figure 3-3. Summary of IWTomics results for individual low-resolution features. A. Distance to the telomere. **B.** Distance to the centromere. **C.** Replication timing. **D.** Sex-averaged recombination rate. **E.** Count of telomere hexamers. Each panel presents the boxplots of the feature in the 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 Fig. S9. The control regions in this study contain less than 7% coverage by all annotated L1 elements.

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 (Figs. 3-2A and S7A-B). 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, 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. 3-2A). Additionally, IWTomics identified 12 high-resolution features with underrepresented signals at *de novo* L1 flanks, suggestive of their negative influence on L1 insertion preferences (Fig. 3-2A). 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. 3-2A).

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 vs. *de novo* L1s. The univariate IWTomics analysis contrasting low-resolution features (Fig. 3-3) 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. 3-2B) identified six high-resolution features that were significantly overrepresented at human-specific L1 flanks vs. 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 vs. those of *de novo* L1s (Fig. 3-2B), suggesting that the regions might undergo selection against L1 fixation, and thus lack fixed L1 elements. While 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 microsatelites were enriched close to the integration site but underrepresented along the remainder of the flanks (Fig. 3-2B), suggesting distinct associations of this feature at different scales.

Also, 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, *Alu*s, 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 Pol2.

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 *Alu*s (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 features 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. 3-2A). 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. 3-2B). 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 (Fig. S8A; Table S2). This is consistent with

our results showing that, in the genome, polymorphic L1s are located closer to human-specific than *de novo* L1s (Fig. 3-1). 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) DNase I hypersensitive sites (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 lI sites were enriched at integration sites of *de novo* L1s (Figs 3-2A and S7A), 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 vs. *de novo* elements (Fig. 3-2B). 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 human embryonic stem cells (hESC 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; Anon)—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. 3-2A; S7B). The associations of these features with L1 integration were confirmed by their significance in both single and multiple FLR 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. 3-2B). 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 (Figs. 3-2B and S7C). 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 vs. *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 non-heterochromatic 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 human embryonic stem cell line H1), 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 (Oliver and Greene 2009) (Rodriguez et al. 2008); (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); (4-5) 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, while

sperm hypomethylation was enriched (Fig. 3-2A); 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. 3-2B); 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. 3-2B); 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. 2006; 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. 3-2A); however, only G-quadruplexes were selected by the mFLR (Table 1). In the comparison of human-specific vs. *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. 3-2B). 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 (Yi et al. 2010) 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. 3-2A) 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 to *de novo* L1s (Fig. 3-2B); 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 (Fig. S18) as well as comparing their enrichment against different genomic features, including GC-content. L1 target motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, 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* vs. 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. 3-2A), 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. Other 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 vs. 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 vs. controls for each level of GC content (Fig. S11). 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 (Fig. S11B-D), 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 to *de novo* L1s, suggesting that telomeric regions are less tolerant of L1 fixation. However, telomere hexamers were significantly underrepresented in *de novo* L1 flanks vs. controls (strong effect in sFLR, selected in mFLR), and in the flanks of fixed vs. *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 (TPE), 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 (Plohl 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. 3-3). While 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 transposable elements—*Alu*s, MIRs, L2/L3 elements, DNA transposon, and LTR elements—revealed important

associations between some such elements and L1 transposition dynamics. Specifically, *Alu*s were overrepresented, while MIRs and L2s/L3s were underrepresented, over 100 kb analyzed for *de novo* L1 flanks vs. controls (the underrepresentation of L2s/L3s was 'patchy'; Fig. 3-2A). All three effects were strong in sFLRs, and *Alu*s 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. 3-2A); 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. 2017). 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 *Alu*s 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. 3-2A). 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 vs. *de novo* L1 flanks comparison, *Alu*s were globally underrepresented, and MIRs and LTRs were under- and overrepresented, respectively, but in a more 'patchy' fashion. *Alu*s had a very strong effect in sFLR, and were selected by the mFLR. Higher coverage of LTR elements in the flanks of human-specific vs. *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 *Alu*s 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. 3-2B).

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 might also suggest 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. 3-3), 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 vs. *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. 3-2B). 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. 3-2 and 3-3, Table 1) and developed two integrative biological models relating the local genomic landscape with L1 insertion and fixation preferences (Fig. 3-4). In these models, the scale and the direction (enrichment vs. depletion) of the signal originate from IWTomics results (Fig.3-2) and are depicted by the width and positive vs. 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. 3-4).

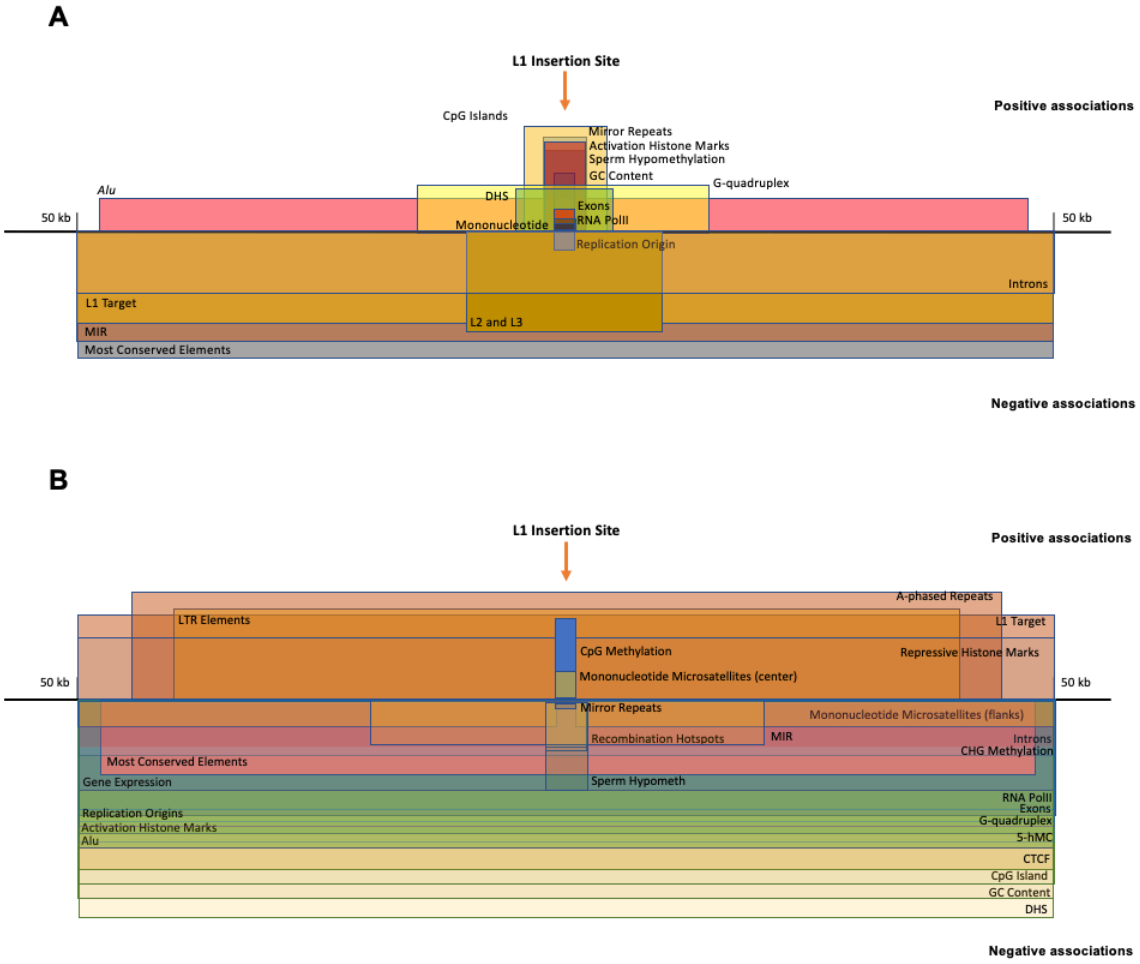


Figure 3-4. Integrative models of L1 transposition dynamics based on IWTomics and sFLR results. **A.** A model for insertion preferences. **B.** A model for fixation preferences. The horizontal black line represents the linear genome structure, with boundaries marking the 100-kb flanking region centered 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).

A model of L1 insertion (Fig. 3-4A). 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). On the other hand, regions enriched with old inactive TEs (ancient L2/L3 and MIR elements) are usually GC-poor (Matassi et al. 1998; Medstrand et al. 2002), and most conserved elements as a rule are present in non-genic (i.e. AT-rich) regions, explaining why they appear to be negative predictors over large regions in Fig. 3-4A.

A model of L1 fixation (Fig. 3-4B). 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.

Impact 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. 3-4B). 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. 3-4B). 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.

To sum up, here 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. Moreover, our results suggest that L1 transpositional activities, in turn, re-shape the genomic landscape over the course of evolution. The findings significantly extend our understanding of L1 transposition dynamics and provides insights into how TEs shape the structure and evolution of the human genome.

References

1. Anon. EpiGenie Epigenetics Background, Tools and Database. Available from: https://epigenie.com/epigenie-learning-center/
2. Asthana S, Roytberg M, Stamatoyannopoulos JA, Sunyaev S. 2005. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.* preprint:e254.
3. 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* 129:823–837.
4. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.
5. Belancio VP, Deininger PL, Roy-Engel AM. 2009. LINE dancing in the human genome: transposable elements and disease. *Genome Med.* 1:97.
6. Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* 34:1512–1521.
7. 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.* 19:837–844.
8. Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: Stability and function of G-quadruplex structures. *Nat. Rev. Genet.* 13:770–780.
9. Boeke JD. 1997. LINEs and Alus—the polyA connection. *Nat. Genet.* 16:6.
10. Boissinot S. 2004. The Insertional History of an Active Family of L1 Retrotransposons in Humans. *Genome Research* 14:1221–1231.

11. 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.* 103:9590–9594.

12. Boissinot S, Entezam A, Furano AV. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* 18:926–935.

13. Bourgeois Y, Boissinot S. 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes* 10:419.

14. Branco MR, Ficz G, Reik W. 2012. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat. Rev. Genet.* 13:7–13.

15. Calado RT, Dumitriu B. 2013. Telomere dynamics in mice and humans. *Semin. Hematol.* 50:165–174.

16. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky 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.

17. Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18:71–86.

18. Cohen NM, Kenigsberg E, Tanay A. 2011. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* 145:773–786.

19. 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* 37:18081–18093.

20. 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* 34:2289–2291.

21. Cremona MA, Xu H, Makova KD, Reimherr M, Chiaromonte F, Madrigal P. 2019. Functional data analysis for computational biology. *Bioinformatics* 35:3211–3213.

22. Croll D, Lendenmann MH, Stewart E, McDonald BA. 2015. The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics* 201:1213–1228.

23. Deininger P. 2011. Alu elements: know the SINEs. *Genome Biol.* 12:236.

24. Deininger PL, Batzer MA. 2002. Mammalian retroelements. *Genome Res.* 12:1455–1465.

25. Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4:e1000071.

26. Elbarbary RA, Lucas BA, Maquat LE. 2016. Retrotransposons as regulators of gene expression. *Science* 351:aac7247.

27. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

28. Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat. Rev. Genet.* 2:549–555.

29. Febrero-Bande M, de la Fuente MO. 2012. Statistical Computing in Functional Data Analysis: TheRPackagefda.usc. *Journal of Statistical Software*. Available from: http://dx.doi.org/10.18637/jss.v051.i04

30. Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87:905–916.

31. 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* 177:837–851.e28.

32. Fu H, Baris A, Aladjem MI. 2018. Replication timing and nuclear structure. *Curr. Opin. Cell Biol.* 52:43–50.

33. 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.* 292:14456–14472.

34. Ghirlando R, Felsenfeld G. 2016. CTCF: making the right connections. *Genes Dev.* 30:881–891.

35. 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.* 25:1135–1146.

36. Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky 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.* 28:1767–1778.

37. 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* 259:81–88.

38. Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429:268–274.

39. He Y, Ecker JR. 2015. Non-CG Methylation in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* 16:55–77.

40. 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.* 9:2911–2920.

41. 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* 14:894–911.

42. 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.

43. 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. *Mob. DNA* 5:14.

44. Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* 94:1872–1877.

45. Jurka J. 2004. Evolutionary impact of human Alu repetitive elements. *Curr. Opin. Genet. Dev.* 14:603–608.

46. 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.* 101:1268–1272.

47. 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:D493–D496.

48. Kejnovský E, Michalovova M, Steflova 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:e45519.

49. 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.* 21:2038–2048.

50. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.

51. Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* 16:78–87.

52. 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* 128:1231–1245.

53. Kines KJ, Belancio VP. 2012. Expressing genes do not forget their LINEs: transposable elements and gene expression. *Front. Biosci.* 17:1329–1344.

54. 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* 467:1099–1103.

55. Koren A, Polak P, Nemesh 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.* 91:1033–1040.

56. Kvikstad EM, Makova KD. 2010. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res.* 20:600–613.

57. 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* 409:860–921.

58. 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.* 42:1733–1746.

59. Lees-Murdock DJ, De Felici M, Walsh CP. 2003. Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* 82:230–237.

60. Lexa M, Steflova P, Martinek T, Vorlickova M, Vyskot B, Kejnovsky E. 2014. Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* 15:1–12.

61. Lindič N, Budič M, Petan T, Knisbacher BA, Levanon EY, Lovšin N. 2013. Differential inhibition of LINE1 and LINE2 retrotransposition by vertebrate AID/APOBEC proteins. *Retrovirology* 10:156.

62. 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* 462:315–322.

63. Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11:2453–2465.

64. 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.* 104:8005–8010.

65. Mahtani MM, Willard HF. 1998. Physical and genetic mapping of the human X chromosome centromere: Repression of recombination. *Genome Res.* 8:100–110.

66. Matassi G, Labuda D, Bernardi G. 1998. Distribution of the mammalian-wide interspersed repeats (MIRs) in the isochores of the human genome. *FEBS Lett.* 439:63–65.

67. Matsui H. 2014. Variable and boundary selection for functional data via multiclass logistic regression modeling. *Comput. Stat. Data Anal.* 78:176–185.

68. 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:e1004531.

69. 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.* 12:1483–1495.

70. Meier L, Van De Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *J. R. Stat. Soc. Series B Stat. Methodol.* 70:53–71.

71. 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.

72. 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*. Available from: http://dx.doi.org/10.7554/eLife.30058

73. 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* 146:1029–1041.

74. 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.* 34:852–856.

75. 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* 310:321–324.

76. 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.* 40:1124–1129.

77. Nishida H, Suzuki T, Ookawa H, Tomaru Y, Hayashizaki Y. 2005. Comparative analysis of expression of histone H2a genes in mouse. *BMC Genomics* 6:108.

78. 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:e1008291.

79. Oliver KR, Greene WK. 2009. Transposable elements: powerful facilitators of evolution. *Bioessays* 31:703–714.

80. Ong C-T, Corces VG. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* 15:234–246.

81. 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. *Mol. Cell. Biol.* 26:1865–1878.

82. Petri R, Brattås PL, Sharma Y, Jönsson ME, Pircs K, Bengzon J, Jakobsson J. 2017. LINE-2 transposable elements are a source for functional human microRNAs and target sites. Available from: http://dx.doi.org/10.1101/218842

83. Petri R, Brattås PL, Sharma Y, Jonsson ME, Pircs K, Bengzon J, Jakobsson J. 2019. LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLoS Genet.* Available from: http://dx.doi.org/10.1371/journal.pgen.1008036

84. 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.

85. Plohl M, Prats E, Martinez-Lage A, Gonzalez-Tizon A, Mendez 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.* 277:19839–19846.

86. Ramsay J, Silverman BW. 2005. Functional Data Analysis. Springer Science & Business Media

87. Rangasamy D, Greaves I, Tremethick DJ. 2004. RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. *Nat. Struct. Mol. Biol.* 11:650–655.

88. Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* 46:21–42.

89. 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.* 25:1091–1103.

90. 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.* 36:770–784.

91. Rosser JM, An W. 2012. L1 expression and regulation in humans and rodents. *Front. Biosci.* 4:2203–2225.

92. 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.* 20:761–770.

93. 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.* 24:243–247.

94. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves 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* 148:335–348.

95. 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.* 82:1–18.

96. Sinden RR. 2012. DNA Structure and Function. Elsevier

97. Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9:657–663.

98. Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* 390:206–213.

99. 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.* 80:1816–1820.

100. 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.* 16:564–571.

101. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger 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* 74:555–570.e7.

102. 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.

103. 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:e1002154.

104. 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.* 14:1607–1616.

105. Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13:36–46.

106. Tsompana M, Buck MJ. 2014. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 7:33.

107. Venkatesan S, Khaw AK, Hande MP. 2017. Telomere Biology-Insights into an Intriguing Phenomenon. *Cells*. Available from: http://dx.doi.org/10.3390/cells6020015

108. Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31:1838–1844.

109. 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:e1002842.

110. 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.* 30:88–99.

111. Wallrath LL, Lu Q, Granok H, Elgin SCR. 1994. Architectural variations of inducible eukaryotic promoters: Preset and remodeling chromatin structures. *Bioessays* 16:165–170.

112. Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337:1675–1678.

113. 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.* 39:457–466.

114. Wimmer K, Callens T, Wernstedt A, Messiaen L. 2011. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet.* 7:e1002371.

115. Xie Y, Mates L, Ivics Z, Izsvák Z, Martin SL, An W. 2013. Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. *Mob. DNA* 4:10.

116. Yang L, Brunsfeld J, Scott L, Wichman H. 2014. Reviving the dead: history and reactivation of an extinct l1. *PLoS Genet.* 10:e1004395.

117. 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.* 46:8299–8310.

118. Yi M, Volfovsky N, Luke BT, Bacolla A. 2010. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic acids*. Available from: https://academic.oup.com/nar/article-abstract/39/suppl_1/D383/2507465

119. Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 13:335–340.

120. Zaratiegui M. 2017. Cross-Regulation between Transposable Elements and Host DNA Replication. *Viruses*. Available from: http://dx.doi.org/10.3390/v9030057

121. Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919.

122. 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:e1008043.

123. Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences* 67:43–62.

124. Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12:7–18.

Chapter 4

Reproducibility and generalizability of the study

Part of the data and analysis in this chapter are published as a research article by Chen, D., Cremona, M.A., Qi, Z., Mitra, R.D., Chiaromonte, F. and Makova, K.D., 2020. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. *Molecular Biology and Evolution.* Reuse of content from the publication for thesis is in compliance of the journal policies.

Summary of Analysis

Reproducibility is critical in scientific research, and as the field of Life Sciences becomes increasingly data-driven, it is now possible to efficiently reproduce most of the analysis by leveraging the publicly available data from different studies. Here we examined the robustness of the experimental design and computational methods employed in this study. We collected multiple datasets representing different sources of L1 integrations and genomic features. We then applied our analytical framework on those sets to test the reproducibility and generalizability of the findings. Finally, we discussed the limitation of the current study and several implications for future work.

Examining the robustness of experimental design

Analysis of *MspI* and *TaqI* site enrichment with different genomic features

We employed the restriction enzymes *MspI* and *TaqI* in the *de novo* L1 insertion experiment (Fig. 2-2), which corresponds to the sites CCGG and TCGA respectively. In order to exclude the possibility that our observation of GC-enrichment near *de novo* L1s is driven by the high GC-

content of the restriction enzyme sites, we studied the enrichment of these sites against different genomic features involved in the FDA analysis, including GC-content. In particular, we extracted the counts of each restriction enzyme recognition site within all the 100-kb regions (L1 flanking regions and control regions) in the current study, and examined their correlation with other high-resolution genomic features (Fig. 4-1). None of the correlations (Spearman) between restriction sites counts and genomic features exceeded 0.8 in absolute value. In fact, *MspI* and *TaqI* showed the highest correlation between themselves (absolute value = 0.45) and the next highest with Exons (0.30 between *MspI* and Exons; 0.21 between *TaqI* and Exons) and Most conserved elements (0.27 between *MspI* and Most conserved elements; 0.34 between *TaqI* and Most conserved elements). This suggests that our results are not driven by potential experimental biases originating from the composition of the two restriction enzyme recognition sites. In addition, we examined the genome-wide distribution of the two restriction enzyme sites, which might introduce potential bias regarding the detection of *de novo* L1s in the sequencing experiment. Specifically, we analyzed the pairwise distances among all the *MspI* and *TaqI* restriction sites in the human genome (Fig. S18). Our results suggested that the two types of restriction sites in the genome are located close to each other, while more than 87% of the pairwise distances are less than the read length of Illumina MiSeq sequencing (150 bp). Therefore, the *MspI* and *TaqI* restriction sites in the genome are adequately enriched (close enough to each other) and are not likely to impact the detection of *de novo* L1s.

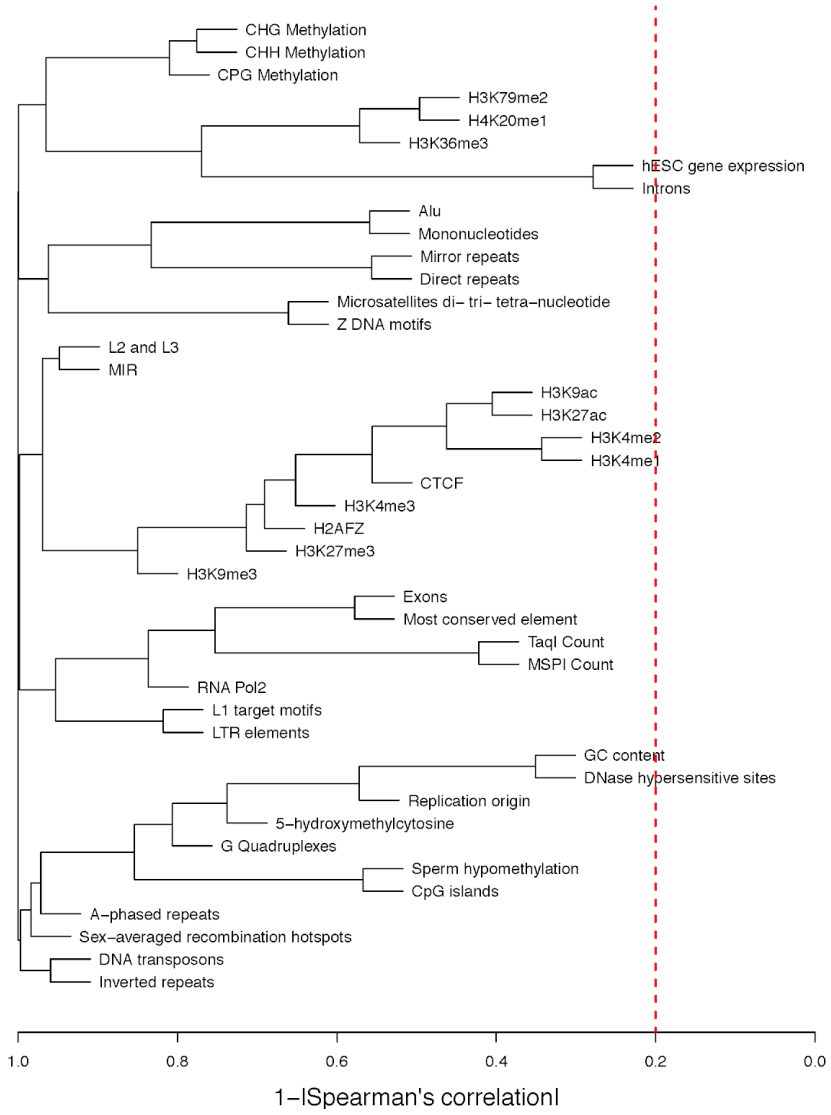


Figure 4-1. Correlation among *MspI* and *TaqI* restriction sites and other genomic features using all windows from the current study. Hierarchical clustering (complete linkage function) based on Spearman's correlation and including *MspI* and *TaqI* restriction sites counts and all features analyzed in the study. The red dashed line marks an absolute value of 0.8 for Spearman's correlation.

Genome-wide analysis of poly(A/T) sequences

Regarding the detection of *de novo* L1 insertions, it is possible that some poly(A/T)s might be trimmed (and the corresponding reads kept) even if they are present in the genome and not encoded by L1s, which might lead to a shift of the integration points. Therefore, we performed a further examination of the genome-wide distribution of poly(A/T) sequences with respect to the *de novo* L1s from the current study. Specifically, we first extracted the genomic coordinates of all poly(A/T) sequences with at least 15 bp in length. Next, we examined the number of (trimmed) reads corresponding to our *de novo* L1 elements and their 1-kb flanks that overlap with the genomic poly(A/T) sequences (Table S8). We found that only a small portion of *de novo* L1s and of their 1-kb flanking regions (<1.3% in both cases) overlap with the polyA/T encoded in the genome (See table below), which is unlikely to compromise the robustness of the *de novo* L1 dataset in our study.

Table 4-1 Number of L1 reads or of their 1-kb flanking regions overlapping genomic poly(A/T) tracts. All L1 *de novo* insertions were used (no filtering).

Dataset	Number of elements/regions	Number of regions overlapping with genomic polyA/T
de novo L1 element reads	17,037	7
1-kb flanks of *de novo* L1 reads	17,037	207

Revisit filtering strategy for *de novo* L1 insertions

Here we revisited the characterization of *de novo* L1 insertions and addressed potential false positive due to PCR-artefacts or random plasmid integration events. In order to ensure the identification of unique *de novo* L1 insertions, we trimmed the poly-Ts at the 3'-end of the reads (read 2) that reached the poly-A tail, and then used the barcode and at least one of the restriction sites as signatures of successful insertion (Fig. S1). Subsequently, 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 bps, given that it is very unlikely to obtain two very close insertions with the same barcode.

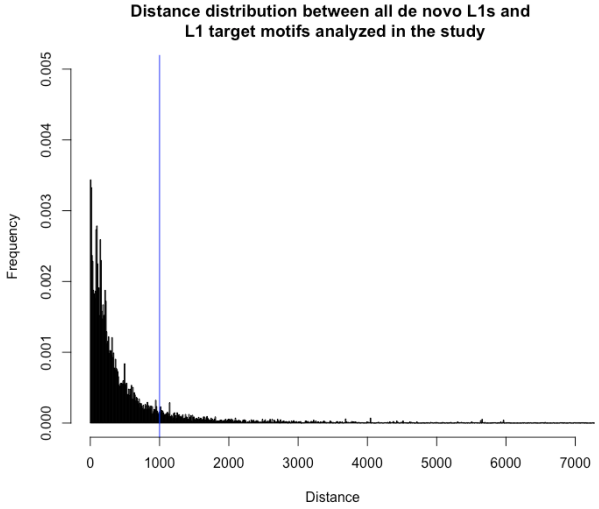
Meanwhile, to exclude the potential bias from PCR-artefacts or random plasmid integration events in the sequencing experiment, we further filtered our *de novo* L1 dataset using a more stringent criterion—only keeping 2,091 insertions supported by a poly-A/T tail longer than 15 bp. This filtering procedure enables us to keep only a fraction of the reads that does reach the junction between the flanking sequence and L1, which yielded the subset of L1 insertions with the exact insertion sites (compared to other approximated insertion loci when the reads do not reach the polyA/T regions).

The stringently filtered subset allowed us to further compare the genomic features such as L1 target site signals with the original *de novo* L1 set and examine the robustness of our current filtering strategy. For instance, the distribution of distances between L1 insertions and the closest consensus L1 target site motifs is shown below (Fig. 4-2). We observed that the closest consensus L1 target site motif was at a distance less than 1 kb for 85.5% of L1 insertions from the first dataset and 88.4% of L1 insertions from the second dataset. While endonuclease nicking at the target DNA occurs at the L1 insertion site, distances between the detected L1 insertion and consensus L1 target site motifs have been reported between a few bases and several hundred bases (Feng et al. 1996; Jurka

1997; Cost et al. 2002; Flasch et al. 2019; Sultana et al. 2019; Zhao et al. 2019) -- depending on the position of the nicking site of target DNA and the length of the polyA within the transposed RNA (Jurka 1997; Cost et al. 2002; Szak et al. 2002). The variation shown by the distances in our study may be due to the following: (a) endonuclease nicking is not entirely specific, and thus there might have been cases when the actual target motif corresponding to the insertion was different from the seven consensus motifs considered in our analyses (Jurka 1997; Cost and Boeke 1998; Boissinot 2004; Zhao et al. 2019); and (b) some of the insertion sites were approximated from the sequencing reads in the original, unfiltered *de novo* L1 dataset.

Therefore, the *de novo* L1 insertions captured in this study are not likely to be false positives, instead, the detection accuracy is supported by both the robust filtering strategy and nearby L1 target motifs.

A



B

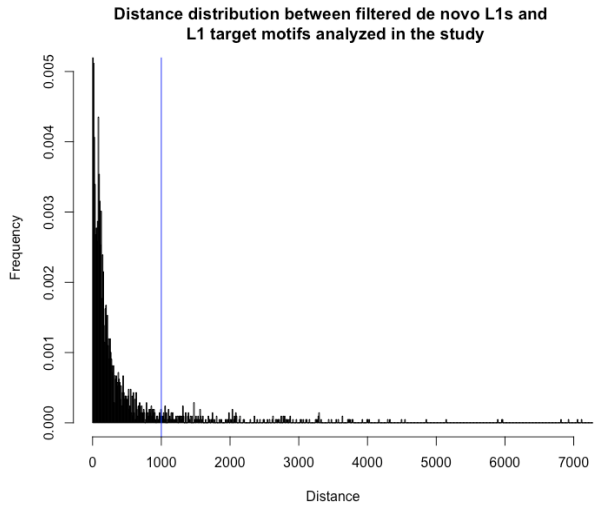


Figure 4-2. Frequency distribution of distances between *de novo* L1 insertions and consensus L1 target site motifs. (A) All *de novo* L1s from the current study. **(B)** Stringently filtered *de novo* L1s containing polyA/T (polyA/T >15 bp). The X-axis indicates the distance between *de novo* L1s and L1 target motifs. The Y-axis indicates the frequency of the corresponding distance value. The blue vertical line marks distance of 1,000 bp.

Reproduce L1 target motif analysis with a further filtered *de novo* L1 set

We analyzed the distribution of distances between L1 elements and consensus L1 target site motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, and TCAAAA) (Jurka 1997; Zhao et al. 2019). Our results suggested an underrepresentation of L1 target sites near *de novo* L1s compared with the (L1-depleted) control, which is counter-intuitive. To further analyze the presence of L1 target site motifs near the *de novo* L1s within the FDA framework, we also contrasted L1 target motifs between *de novo* L1 and control regions, using two *de novo* L1 datasets. The first dataset included all 17,037 *de novo* L1 insertions analyzed in our manuscript. The second was a more stringently filtered subset of the first, consisting of 2,091 *de novo* L1 insertions supported by a poly-A/T tail longer than 15 bp (as described in the previous section). As a result, the L1 target site signals were still underrepresented in the *de novo* L1 vs. control comparison (Fig. 4-3B, see below), which is consistent with our initial findings (Figs. 3-2 and 4-3A). Furthermore, the comparison revealed that in both cases, the mean motif counts (per 1 kb window) in the *de novo* L1 regions were between two and three (Fig. 4-4), which does not indicate a complete depletion of L1 target site motifs. In addition, the results further supported that the detection of L1 insertions with our current filtering strategy is robust.

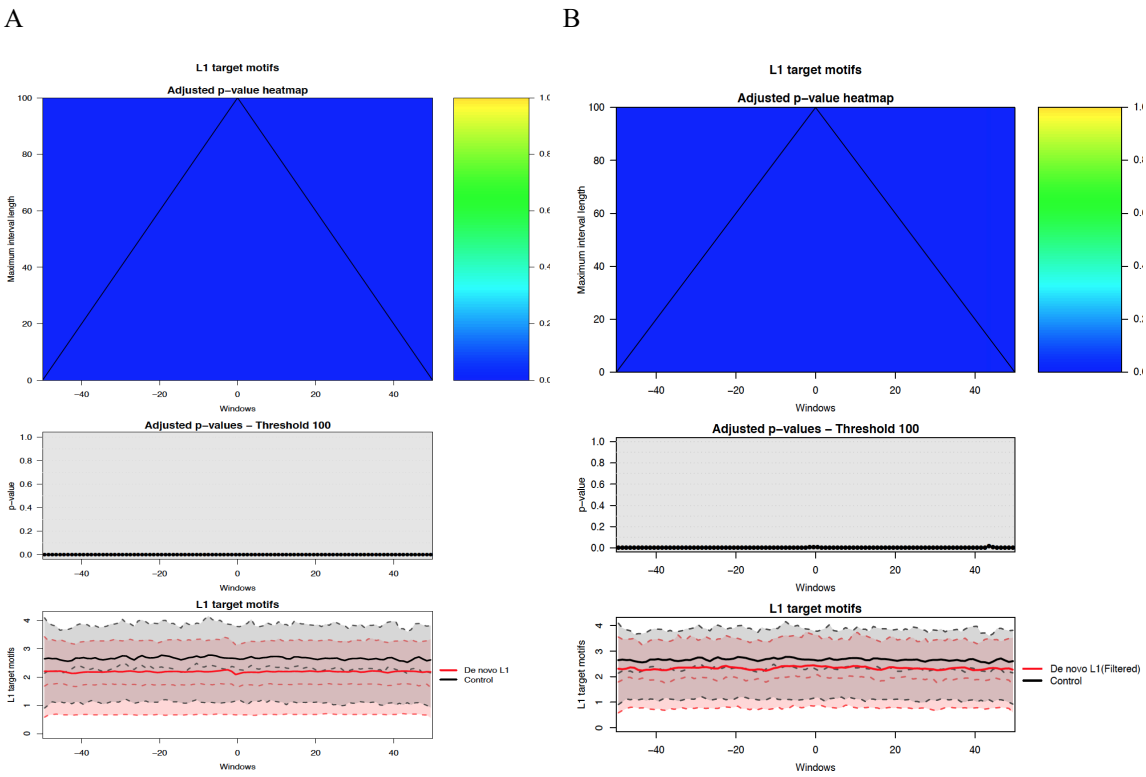


Figure 4-3. IWTomics results for L1 target site motifs in *de novo* L1s and controls. Comparisons of L1 target site signals between **(A)** All *de novo* L1s from the current study vs. controls, and **(B)** Stringently filtered *de novo* L1s containing polyA/T (polyA/T >15 bp) vs. controls. The top heat map shows the IWTomics-adjusted *p*-value curves at all possible scales—the X-axis reports the position in the flanking region, from -50 kb to +50 kb; the Y-axis reports the scale used for *p*-value correction, from 1 kb (no correction) to 100 kb (considering the whole region). The central panel shows the adjusted *p*-value curve at the selected scale, with a gray area indicating significant *p*-values (≤0.05). The bottom panel reports pointwise boxplots of the feature values (dotted curves: 25%, 50%, and 75% quantiles, solid curves: averages). The control regions used have less than 7% coverage by all annotated L1 elements.

Analysis of aneuploid hotspots in HEK-293T cell lines

Here we studied integration preferences for *de novo* L1s using engineered L1 sequences from kidney stem cells (HEK-293T). This represents a useful model system to capture *de novo* L1 insertions. 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 (Fig. S1; Materials and Methods). 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 (Fig. S3), 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 (Fig. S14), 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 (Fig. S13), suggesting that multiple insertion events were actually driven by local genomic landscape features instead of by amplified regions in the genome of HEK-293T cells.

Testing reproducibility of the study using publicly available datasets

de novo L1 datasets from different studies

The *de novo* L1 insertions included in our study were harvested from a cell line experiment and might not reflect 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). Therefore, we compared our findings with those of two recent *de novo* L1 integration datasets generated in hESC (Flasch et al. 2019) and HeLa (Sultana et al. 2019) cells (Table S6). 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 (Table 4-2). For instance, active histone marks and early replicating domains contributed positively to L1 integration (though with different strengths) across all three studies. This suggested that our biological model of the L1 insertion dynamics is generalizable. However, some other findings were inconsistent among the studies (e.g., for DHS and H3K27me3; Table 4-2). 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 datasets from (Flasch et al. 2019; Sultana et al. 2019) (Figs. S16 and 17) but might be explained by differences in cell lines and genomic scales used in different studies (Flasch et al. 2019; Sultana et al. 2019; Chen et al. 2020).

Table 4-2 Comparison of genomic landscape features and their contribution to L1 activities in different studies. Chromatin=chromatin structure, Transcription = transcription regulation and gene expression, Res.=Resolution. "NS" cells indicate features that showed no significant differences (in current study). "NI" cells indicate features not included (from other recent studies).

Group	Name	**Current study (HEK-293T)**	**Sultana et al 2019 (Hela)**	**Flasch et al 2019 (hESC,PA-1, K562, Hela, NPC)**
		L1 Insertion (*de novo* L1 vs control)	L1 Insertion	L1 Insertion
Chromatin	DNase hypersensitive sites	Positive	Slightly Positive	Open chromatin does not promote local L1 integration
Chromatin	RNA Pol II	Positive	NI	NI
Chromatin	CTCF	NS	NI	NI
Transcription	H3K4me2	Positive	Positive	Minimal enrichment in some enhancer states in certain cells (e.g. HeLa-JVM and hESC); Insertions slightly depleted in regions with active transcription epigenetic marks
Transcription	H3K9ac	Positive	Positive	
Transcription	H3K4me3	Positive	Positive	
Transcription	H3K79me2	NS	Positive	
Transcription	H3K27ac	Positive	Positive	
Transcription	H4K20me1	Positive	Negative	
Transcription	H3K4me1	Positive	Positive	
Transcription	H3K36me3	Positive	Positive	
Transcription	H3K9me3	NS	Negative	
Transcription	H3K27me3	Positive	Negative	
Transcription	H2AFZ	NS	Positive	
Transcription	Gene expression	NS	Slightly Positive	Level of expression not directly correlated with integration.
DNA methylation	Sperm hypometh	Positive	NI	NI
DNA methylation	CpG methylation	NS	NI	NI
DNA methylation	5-hMc	NS	NI	NI
DNA methylation	CHH methylation	NS	NI	NI
DNA methylation	CHG methylation	NS	NI	NI
Non-B DNA, microsats	G-quadruplex	Positive	NI	NI
Non-B DNA, microsats	A-phased repeats	NS		
Non-B DNA, microsats	Direct repeats	NS		

Non-B DNA, microsats	Inverted repeats	NS	Preferred integration into AT-rich low-complexity repetitive DNA	Positive (L1 insertions detected in tandem repeat sequences located on different chromosomes)
Non-B DNA, microsats	Mirror repeats	Positive		
Non-B DNA, microsats	Z DNA motifs	NS		
Non-B DNA, microsats	Mononuc. Microsats	Positive		
Non-B DNA, microsats	Di-, tri-, and tetranucl. microsats	NS		
Nucleotide composition	GC-content	Positive	Negative	Negative
L1 target motifs	L1 target motifs	Negative	NI	Positive
Other TEs	Alu	Positive	No strong evidence for preferred integration into existing transposable elements	L1 insertions found in SINEs and LINEs (Enriched; Dataset S15)
Other TEs	MIR	Negative		
Other TEs	L2 and L3	Negative		
Other TEs	DNA transposons	NS		L1 Insertions found in transposon-free regions (Not enriched; Dataset S15)
Other TEs	LTR elements	NS		
Replication	Replication origins	Negative	No evidence of enrichment at replication origins	No strong preference for L1 integration at origins or termination zones
Recombination	Recomb. hotspots	NS	NI	NI
Selection	Most conserved elements	Negative	NI	Not enriched
Selection	CpG island	Positive	NI	NI
Selection	Exons	Positive	NI	Genes not preferred by insertions, yet fewer insertions in introns than expected (Figure 3D) (Engineered L1s readily integrated into the introns of genes. However, genes were not preferential L1 integration targets.)
Selection	Introns	Negative	NI	
Chromosomal location	Distance to centromere	NS	NI	Positive (Centromeric or
Chromosomal location	Distance to telomere	NS	NI	

Chromosomal location	Telomere hexamer	Negative	NI	telomeric regions were found to contain highly non-random clusters of insertions.)
Replication	Replication timing	Positive (Early replicating domains)	Positive (Early replicating domains)	Positive (Early replicating domains)
Recombination	Recombination rate	NS	NI	NI

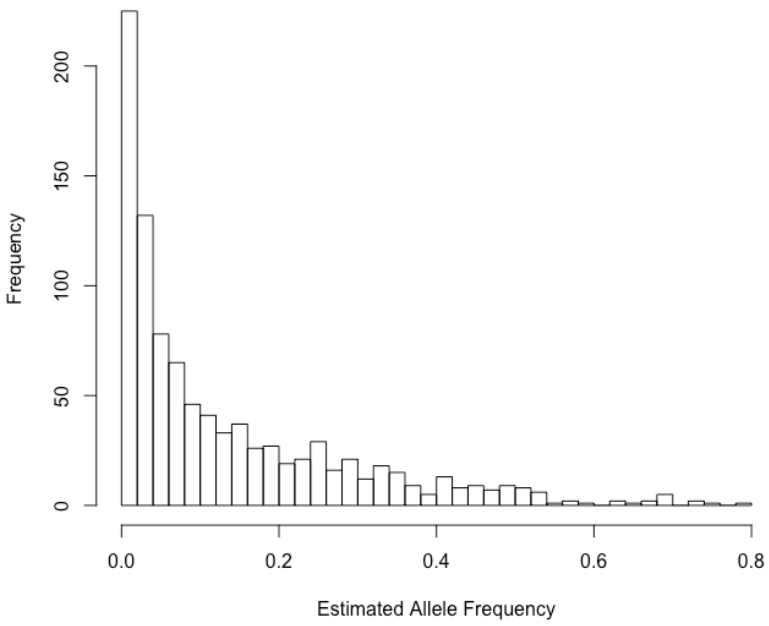
Revisit polymorphic L1s based on allele frequency information

We are aware that the polymorphic L1 dataset is relatively small compared to our *de novo* L1 dataset, though its size is comparable to that of our L1HS dataset. Therefore, we examined several other data sets of human polymorphic L1s (Stewart et al. 2011; Consortium and The 1000 Genomes Project Consortium 2015; Yu et al. 2017), including two studies based on the 1000 Genomes/Human Genome Structural Variation Consortium (Stewart et al. 2011)(Yu et al. 2017). The following data table (Table S5) summarizes sample information and insertion allele frequency spectra based on the relevant datasets provided by these studies. In general, the main findings from these additional L1 polymorphism studies are comparable to those reported by (Ewing and Kazazian 2011)—in terms of both estimated insertion rates and allele frequency spectrum. The datasets from the additional studies, too, could be analyzed with our approaches and statistical pipelines. However, we opted not to include them in our current study because (i) they have substantial overlap with the data we already analyzed (over 40% of the polymorphic L1 elements from (Stewart et al. 2011) are represented in our referenced dataset), or (ii) they are limited to a single population (Yu et al. 2017) (Table S5). In comparison, the polymorphic L1 dataset 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 (Swewart et al. 2011; Yu et al. 2017) (Table S5).

Next, to further address the evolutionary stages of the polymorphic L1s and their potential impact on our results, especially the pairwise comparisons involving the polymorphic L1 dataset, we took into account the allele frequency information of the elements. We first calculated the estimated allele frequencies for each element as described in (Ewing and Kazazian 2011) and divided the polymorphic L1s into four subsets based on the allele frequency (0-25%, 25-50%, 50-75%, and 75-100% quantiles)(Fig. 4-4).

A



B

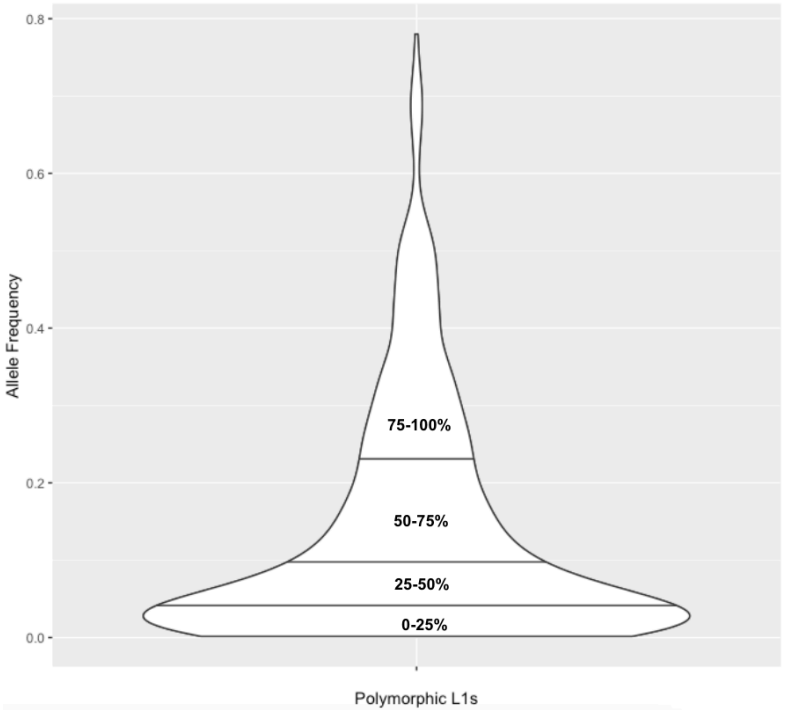
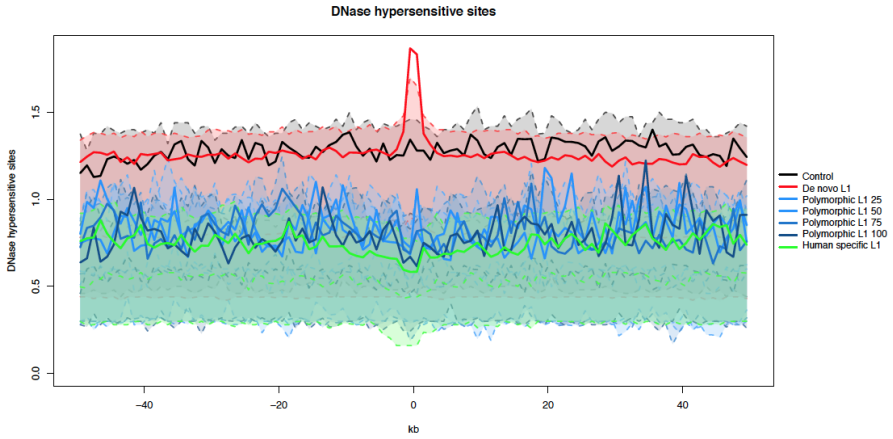


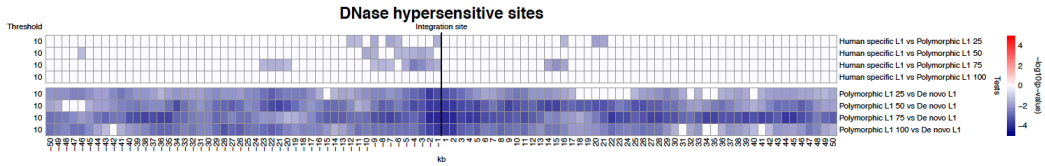
Figure 4-4. Reanalyze the polymorphic L1s based on allele frequency. A. Histogram showing the estimated of allele frequencies for polymorphic L1 insertion sites based on binary present/absent insertion site genotypes in the sample of 310 unrelated individuals (Ewing and Kazazian 2011). **B.** Violin plot showing the four polymorphic L1 subsets based on allele frequency quantiles (0-25%, 25-50%, 50%-75%, 75%-100%).

We then reproduced the pairwise comparison and IWTomics analysis using the four polymorphic L1 subsets (Fig. 4-5). Our results suggested that genomic feature signals (e.g., DHS and H3K27ac) are consistent across the four polymorphic L1 subsets (Fig. 4-5A; C). Meanwhile, the subset with highest allele frequencies (75-100% quantile) has signals closest to the human-specific L1s, indicated by no significant difference in the 100-kb flanks based on the *IWTomics* analysis (Fig. 4-5B; D). The results suggested that polymorphic L1s with higher allele frequency are likely to be closer to the fixation stage than insertion stage, which might also help explain our observation that polymorphic L1s are located closer to human-specific L1s compared to *de novo* L1s based on the distance distribution analysis (Fig. 3-1). This also complement our previous findings and models based on the main comparisons, which only involved *de novo* L1s and human-specific L1s (Figs 3-2 and 3-4). In addition, our results also indicated that polymorphic L1s may be an important resource for future studies. For instance, one can utilize the allele frequency of polymorphic L1s to parse out different selection stages and further characterize the transition from insertion to fixation of the L1 elements.

A



B



C

D

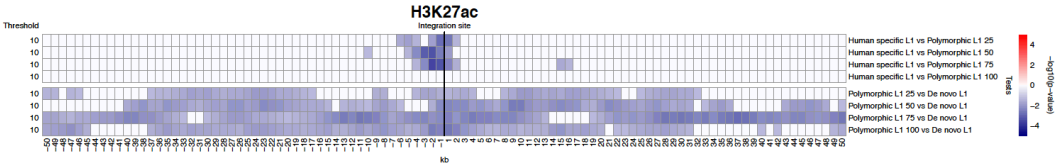


Figure 4-5. IWTomics analysis for selected genomic features using polymorphic L1 subsets based on estimated allele frequencies. Comparisons of genomic features between flanking regions of different polymorphic L1s and other L1 regions, where the former are parsed based on allele frequencies; specifically, presenting 0-25%, 25-50%, 50-75%, and 75-100% quantiles. A. DNase Hypersensitive Sites (DHS) B. H3K27ac C. H3K27ac. In each figure, the top panel reports pointwise boxplots of the feature values (dotted curves - 25%, 50%, and 75% quantiles) with averages (solid curves). Different L1 flanking regions are represented by colored curves and control regions are shown in black curves. The bottom heat map shows the IWT adjusted p-value curves at all possible scales - the X-axis reports the position in the flanking region, from -50 kb to +50 kb; the Y-axis reports the scale used for p-value correction, from 1 kb (no correction) to 100 kb (considering the whole region). The control regions contain less than 7% coverage by all annotated L1 elements.

A random control set without considering genomic L1 sequences

Given the strategy of constructing the control set in the pairwise comparisons, we were aware that the 100-kb L1-poor regions chosen as control may influence our results. For instance, the control regions are unlikely to be neutral in terms of genomic features, and may introduce bias to the overrepresented and underrepresented signals. However, we think that such regions, although not perfect, are appropriate for our study design and can generate robust signals based on the following reasoning. First, we designed our study in such a way that the conclusions and models were drawn from six pairwise comparisons, three of which do not involve this control set (polymorphic L1 vs. *de novo* L1, human-specific L1 vs. *de novo* L1, and polymorphic L1 vs. human-specific L1). Our observations were consistent across different pairwise comparisons, thus our findings were not driven by the choice of controls. Second, we repeated the analysis using a completely random control set, created without imposing the absence of other LINE elements. In this alternative control set the genomic feature signals shifted towards the ones of the human-specific L1 set, yet the vast majority of our main observations and conclusions remained the same (Fig. 4-6). In fact, the shift of control signals towards L1HS (and away from *de novo* L1s) is likely due to older, fixed L1 elements (e.g. L1PAs) which are abundant in random control regions (L1s constitute 17% of the human genome).

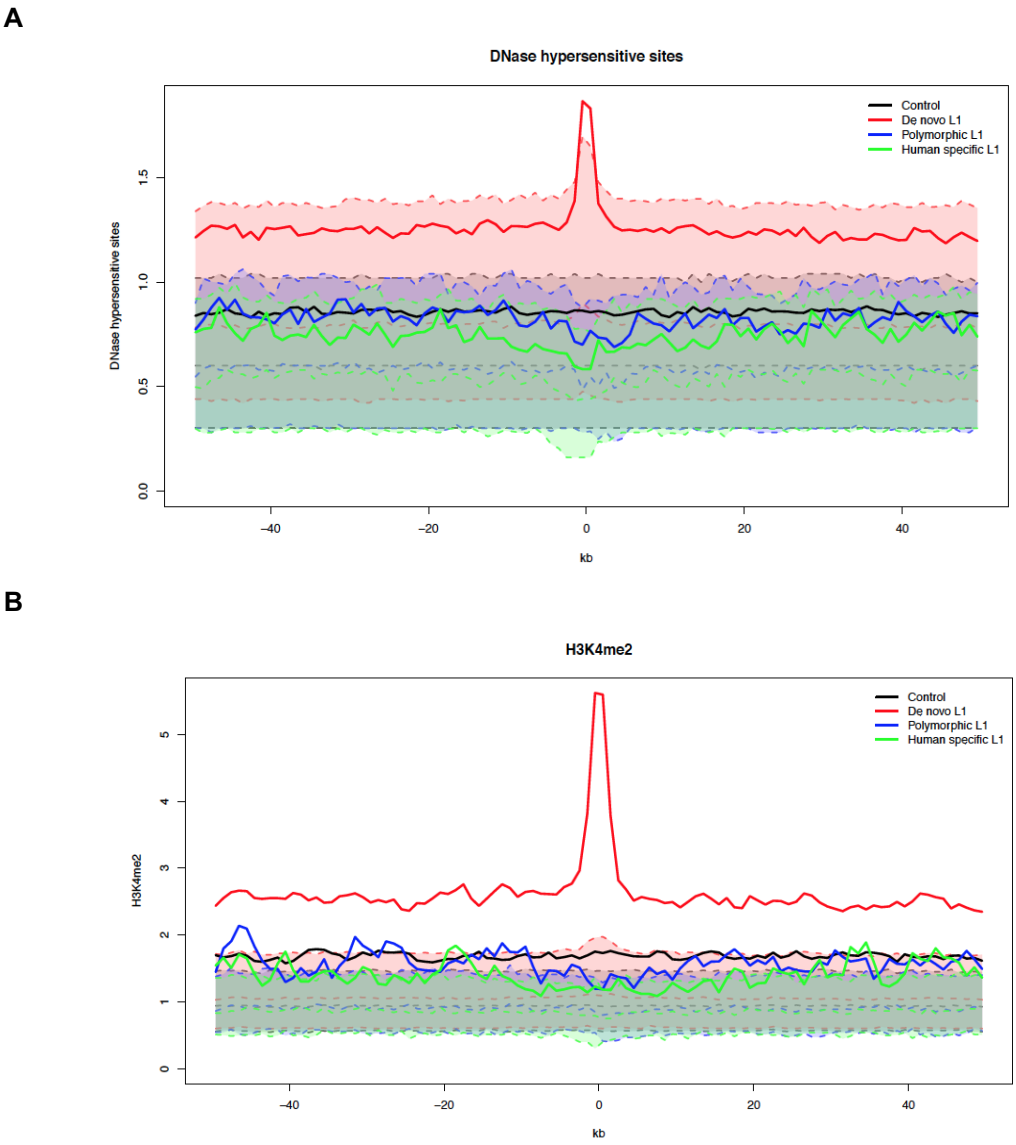


Figure 4-6. DHS and H3K4me3 signals for a random control set. Pointwise boxplots of (A) DHS and (B) H3K4me2 signal values in L1 flanking regions vs. controls using a completely random control set created without imposing the absence of other LINE elements. Boxplots of the feature values are shown in different quantiles (dotted curves: 25%, 50%, and 75% quantiles, solid curves: averages). The same color scheme is used in boxplots and scatterplots (Black: control; Red: *de novo* L1; Blue: polymorphic L1; Green: human-specific L1).

Genomic features produced in different cell lines

Given that our *de novo* L1 dataset was generated in HEK-293T cells, while a substantial set of genomic features were generated from the hESC cells, we were aware of the potential biases from such differences, especially for the cell line-specific features such as histone modification signals. Therefore, 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 to the same features generated in hESC cells. The results indicated substantial genome-wide correlation between the features from HEK-293T (or HEK-293) and hESC (Table S4; Fig. S15). Therefore, the genomic feature datasets employed in our study are generally representative of the genomic landscape of HEK-293T cells.

Limitations of the current study

In this study, we collected the *de novo* L1s from an integration experiment performed in HEK-293T. Although this is a stem cell line and we excluded the potential bias from aneuploidy hotspots (Lin et al. 2014; Binz et al. 2019)(Fig. S14), we are aware that it does not fully represent the karyotype and genome landscape of germline cells. The issue can be further reflected by the inconsistencies between similar studies in different cell lines (Flasch et al. 2019; Sultana et al. 2019; Chen et al. 2020). Therefore, future work should ideally utilize *de novo* L1 insertions collected from large trio sequencing experiments, alternatively, corrections of cell-specific structural variations and rearrangements can be made by whole-genome sequencing of the cell lines (Sultana et al. 2019). In addition, our mFLR models explained as much as ~30% of the variability in L1 insertion and fixation behavior (Table 2-1), which suggested a strong explanatory power and allowed us to gain important insights. However, we also realized that a substantial share of variability was not explained, and the reason can be two-fold. First, 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. Seconds, some genomic features affecting L1 integration and fixation dynamics might still be missing from this study. Additional features, once information on them becomes available, should be incorporated in future investigation.

References

1. 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* 479:534–537.

2. Bedrosian TA, Quayle C, Novaresi N, Gage FH. 2018. Early life experience drives structural variation of neural genomes in mice. *Science* 359:1395–1399.

3. 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. *Scientific Reports*. Available from: http://dx.doi.org/10.1038/s41598-019-47002-0

4. Boissinot S. 2004. The Insertional History of an Active Family of L1 Retrotransposons in Humans. *Genome Research* 14:1221–1231.

5. Consortium T 1000 GP, The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68–74. Available from: http://dx.doi.org/10.1038/nature15393

6. 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* 37:18081–18093.

7. Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* 21:5899–5910.

8. Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* 21:985–990.

9. Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87:905–916.

10. 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* 177:837–851.e28.

11. 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.* 25:1135–1146.

12. Greven S, Scheipl F. 2017. A general framework for functional regression modelling. *Stat. Modelling* 17:1–35.

13. Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* 94:1872–1877.

14. 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.* 23:1303–1312.

15. Lees-Murdock DJ, De Felici M, Walsh CP. 2003. Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* 82:230–237.

16. 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.

17. 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.

18. 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.* 102:18644–18648.

19. Rosser JM, An W. 2012. L1 expression and regulation in humans and rodents. *Front. Biosci.* 4:2203–2225.

20. 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. 20:761–770.

21. Scott EC, Devine SE. 2017. The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses. Available from: http://dx.doi.org/10.3390/v9060131

22. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger 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* 74:555–570.e7.

23. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol.* 3:0052.

24. Usset J, Staicu A-M, Maity A. 2016. Interaction models for functional regression. *Computational Statistics & Data Analysis* 94:317–329.

25. Walker JA, Konkel MK, Adrian M, Stewart C, Kural D, Stro MP, Urban AE, Grubert F, Lam HYK, Lee W-P, et al. 2011. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. Available from: http://dx.doi.org/10.1371/journal.pgen.1002236

26. 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* 6:1–11.

27. 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:e1008043.

Chapter 5

Conclusions

Summary

The main focus of this dissertation is to characterize the transposition dynamics of TEs and understand their contribution to the architecture and evolution of the human genome. Specifically, we addressed the objectives by testing the working hypotheses: *Different local genomic landscape features contribute in various ways to the L1 insertion and fixation preferences in the human genome.* First, we presented a study design to investigate the genome-wide distribution of L1s at different evolutionary distances and the correlation between local genomic features and L1 insertion and fixation preferences. Second, we applied the study design to three large L1 datasets and an extensive list of high-resolution genomic landscape features to test the above working hypothesis. We built an integrative model of L1 transposition dynamics based on our findings (Chen et al. 2020). Finally, we tested the robustness of the study design and reproducibility of our findings. We conducted multiple computational experiments to test the robustness of the study design. We also validated the reproducibility of our findings using recently published L1 datasets and genomic feature signals from different sources.

In general, we found that the genomic distribution of *de novo* L1s, polymorphic L1s and fixed, human-specific L1s are not random, while they reveal distinct patterns by different evolutionary times. We also developed an integrative view of the L1 transposition dynamics (Fig. 5-1), suggesting that L1s preferentially integrate into active, open-chromatin regions enriched in non-B DNA motifs and high transcriptional activities (Fig. 5-1A and B). However, such insertions are mostly removed from the gene pool, possibly due to strong selection pressure on those regions (Fig.

5-1C). In contrast, the L1 elements are more likely to reach high frequencies and to become fixed in non-active genomic regions that are largely free of purifying selection—depleted of genes and most non-coding conserved elements, although such regions usually have closed chromatin structures and are less prone to L1 insertion events (Fig. 5-1). Furthermore, our results also suggest that L1 insertions can even potentially modify local genomic landscape by extending CpG methylation and increasing mononucleotide microsatellite density (Fig. 3-2).

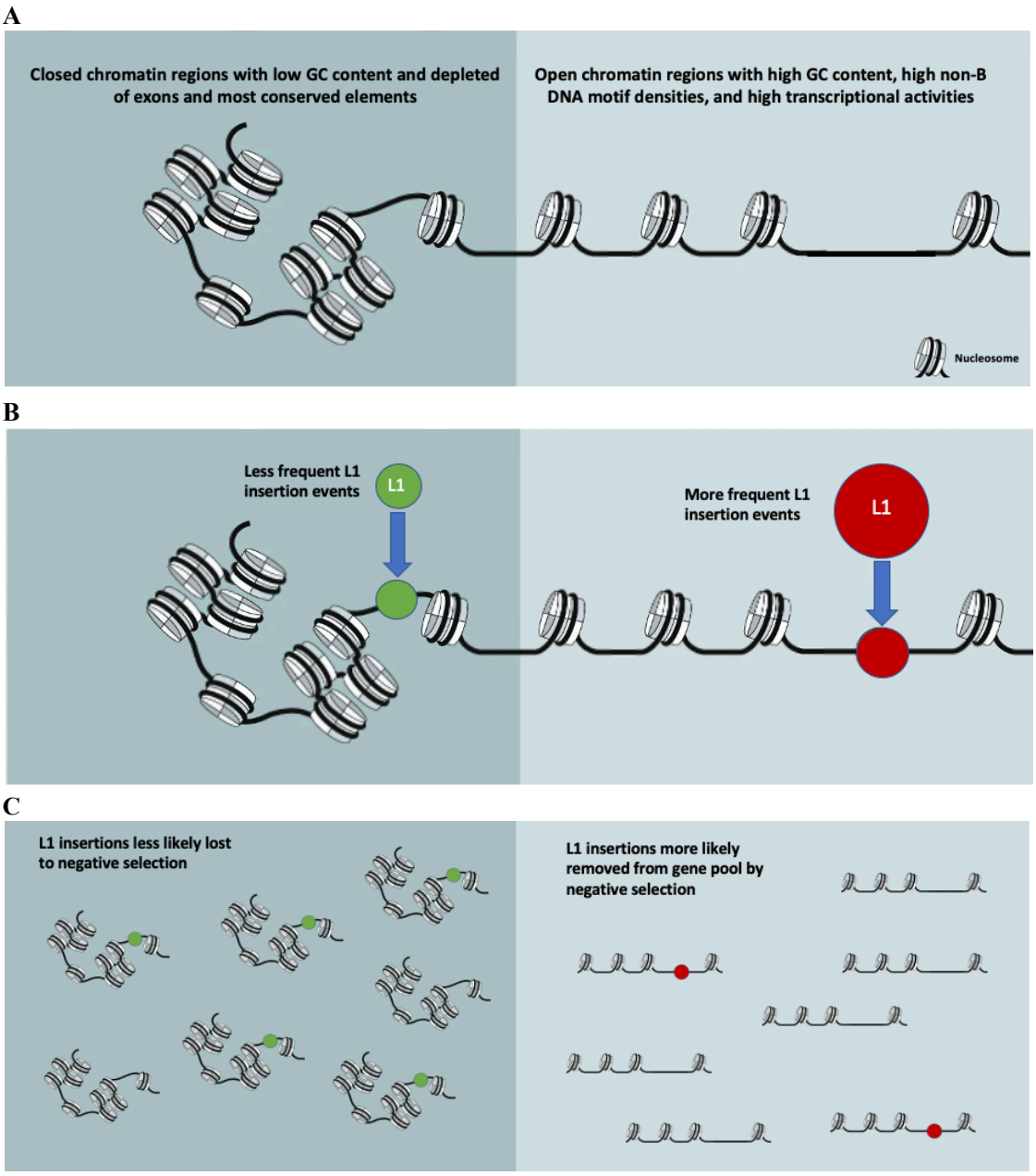


Figure 5-1. An overview of L1 transposition dynamics in the genome. An overview of the L1 transposition dynamics in closed chromatin regions (Left) compared to open chromatin regions (Right) showing **A.** Different nucleosome distribution and chromatin accessibility; **B.** Different frequencies of L1 insertion events; **C.** Different number of L1s retained in the gene pool after negative selection. The visualization of nucleosomes was inspired by (Klemm et al., 2019).

Significance and future directions

In this dissertation, we performed the first genome-wide analysis of L1 transposition dynamics in an evolutionary framework and leveraged an extensive list of genomic landscape features at high resolution. We demonstrated that the genomic distribution of human L1s is driven by the local genomic landscape, and our analysis revealed the potential mechanisms through which regional genomic characteristics influence new element insertions and their abilities to fix in the genome. Our findings provided a dynamic view of the human L1 transposition, while we connected the integration, selection, and fixation processes of L1s with local genomic features. This is critical for understanding the impact of L1s in the genome functions and evolution. The work also provides significant clinical implications in the study of different genetic disorders and cancers, given the presence of L1 activities in both germline and somatic cells. In addition, we presented a successful application of FDA framework, which allowed us to effectively address scale and location of the features' effects on specific genomic intervals, and can be widely applied in the field of genomics. Altogether, the findings in this dissertation substantially facilitate understanding of TE integration and fixation preferences, pave the way for uncovering their role in human health and diseases, and inform their use as mutagenesis tools in genetic studies.

Future studies can be conducted from the following aspects:

1. We studied integration preferences for *de novo* L1 insertions in kidney stem cells (HEK-293T). While this is a useful model system to study L1 activities, we are aware that cell lines might not fully represent the same karyotype and genomic landscape for TE activities as in germline cells. Therefore, our findings can be further validated in future large-scale trio resequencing studies, using a similar analytical framework.
2. We reproduced our analysis on recently published *de novo* L1 datasets and found generally similar patterns (Flasch et al. 2019; Sultana et al. 2019; Chen et al. 2020). However, some

findings were still inconsistent among the studies, possibly due to different cell lines used, different genomic scales analyzed, and other methodological differences. Future studies applying the same framework and genomic landscape features across different cell lines can help explain the cell and tissue-specific differences with more confidence. More importantly, since L1 insertions are often found in different somatic tissues, addressing such differences can further reflect the potential role of L1 transposition in human development and provide important clinical implications (Muotri et al. 2005; Kano et al. 2009; Baillie et al. 2011; Bedrosian et al. 2018).

3. It is possible that some genomic features affecting L1 insertion and fixation preferences might still be missing from our list, since there was a substantial share of variability that our models cannot explain. Therefore, additional genomic features should be incorporated in future studies as the data become increasingly available. For instance, information on L1 mRNAs can be utilized to identify expressed donor copies and might help explain some cell-specific differences. Also, new data from chromosome conformation capture (3C) technologies can be added to indicate the high-order chromatin organization near L1s. In addition, the post-integration selection on L1s could be further characterized by additional features such as fixation index (F_{ST}), phyloP scores, and SNP densities.
4. Anticipated advances in statistical methods, particularly in the domain of functional variable selection and functional logistic regression, are likely to provide better models. For instance, 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 pre-selecting features based on individual tests. Future studies can also benefit from improved mFLR algorithms implemented with interaction terms (Usset et al. 2016; Greven and Scheipl 2017), which will help better explain the joint effect between two or more predictors.

5. Future studies with similar experimental design and analytical framework can be applied to the investigation of other TE families, such as *Alus* and ERVs, which will further shed light on how different TEs jointly contribute to the structure, function, and evolution of the human genome.

References

1. 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 479:534–537.

2. Bedrosian TA, Quayle C, Novaresi N, Gage FH. 2018. Early life experience drives structural variation of neural genomes in mice. Science 359:1395–1399.

3. Chen D, Cremona MA, Qi Z, Mitra RD, Chiaromonte F, Makova KD. 2020. Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. Mol. Biol. Evol. Available from: http://dx.doi.org/10.1093/molbev/msaa194

4. 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 177:837–851.e28.

5. Greven S, Scheipl F. 2017. A general framework for functional regression modelling. Stat. Modelling 17:1–35.

6. 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. 23:1303–1312.

7. 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. 102:18644–18648.

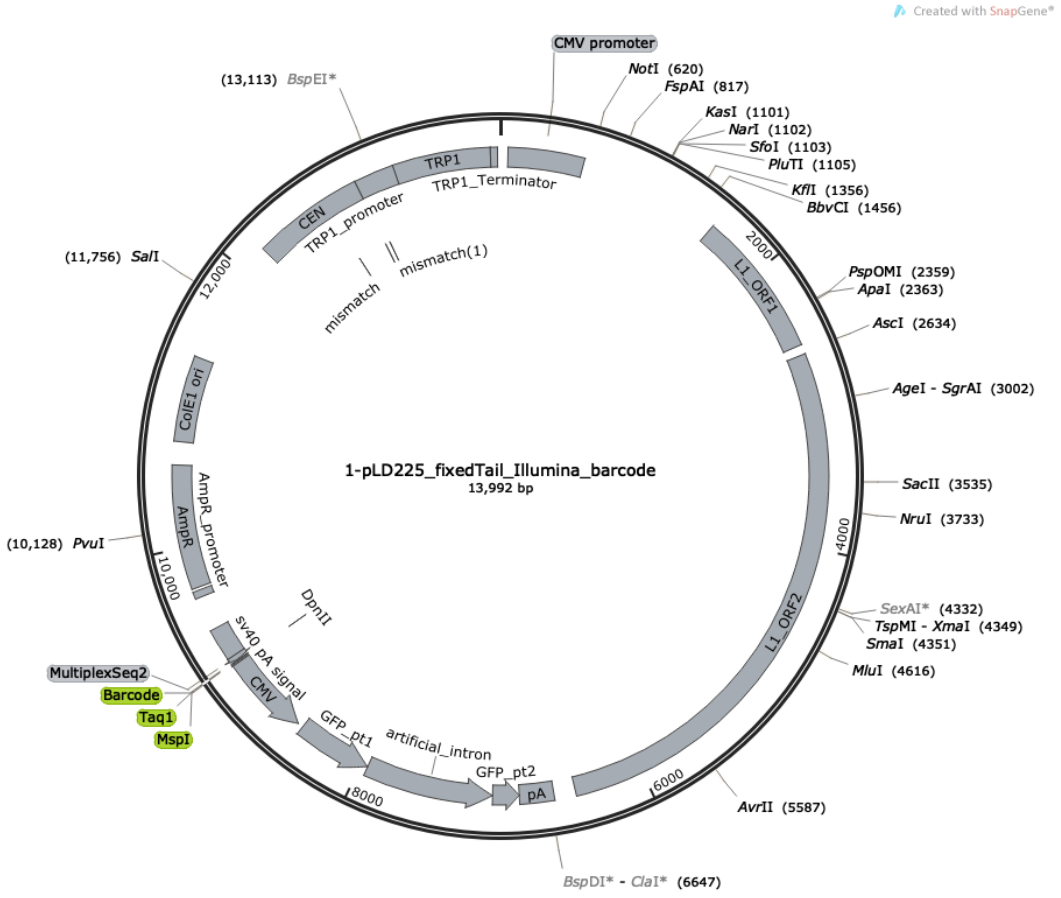
8. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, Pioger L, Nigumann P, Saccani S, Andrau J-C, et al. 2019. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. Mol. Cell 74:555–570.e7.

9. Usset J, Staicu A-M, Maity A. 2016. Interaction models for functional regression. Computational Statistics & Data Analysis 94:317–329.

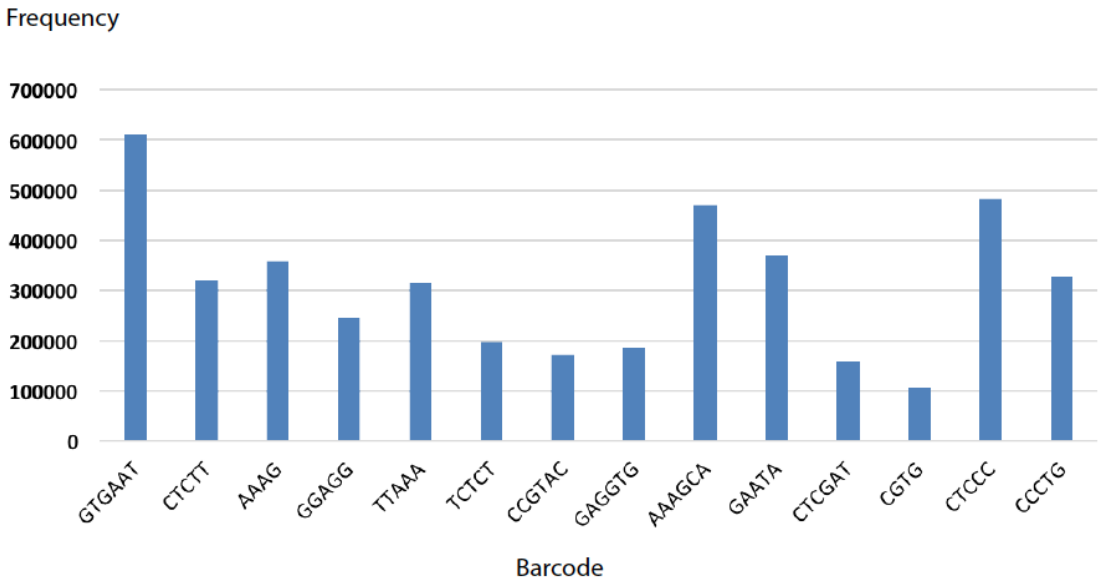
Appendix

Figure S1. Using Inverse-PCR to identify *de novo* L1 insertions *in vivo.*
A. Sequence map of plasmid pld225. Locations of barcode and two restriction enzymes are highlighted green. **B.** Frequency distribution of 14 barcodes implemented in the vector. **C.** The combinations of 14 barcodes and two restriction enzyme sites implemented in the vector, with 28 combinations in total.

A



B



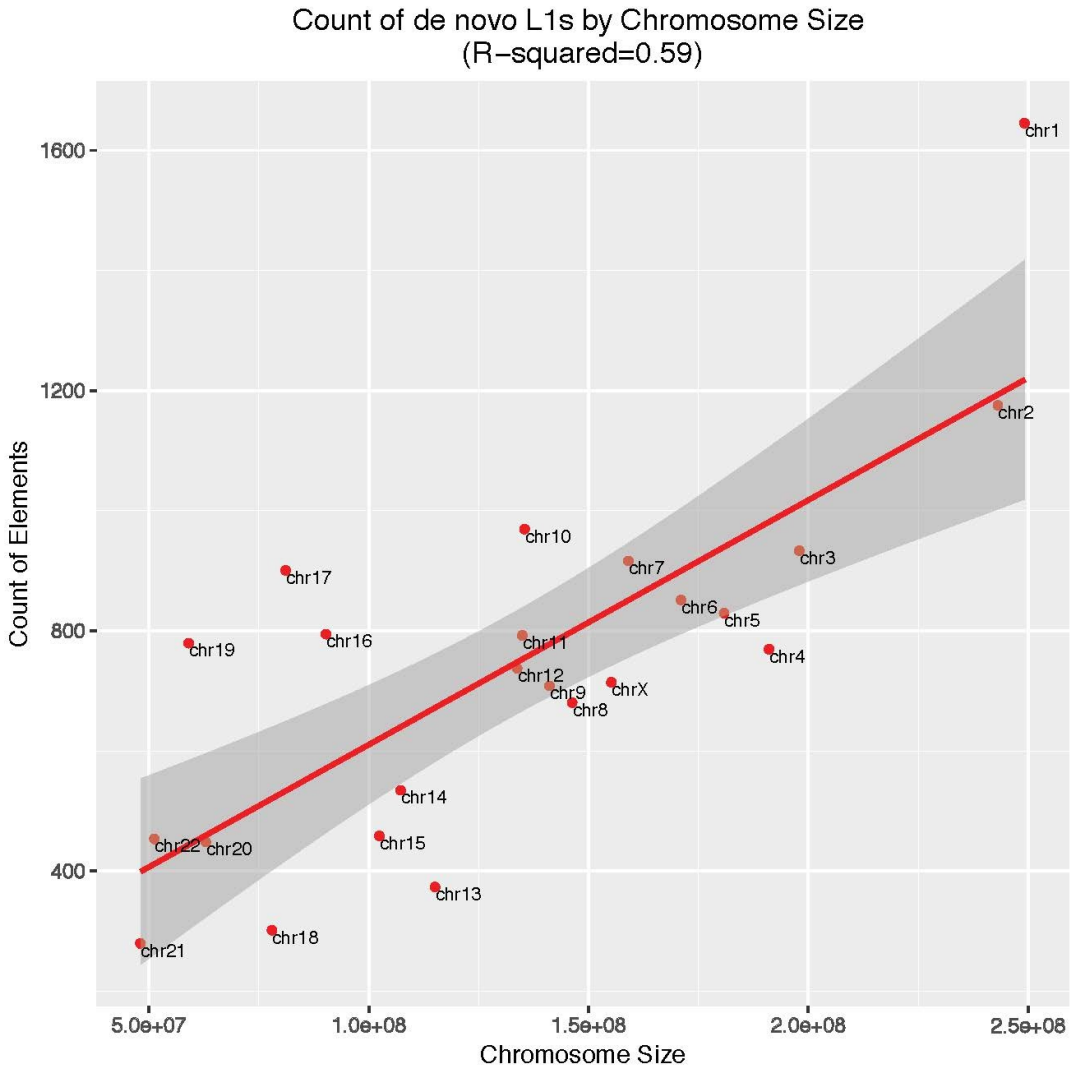
C

GTGAATTCGA	GAGGTGTCGA	GTGAATTCGACCGG	GAGGTGTCGACCGG
CTCTTTCGA	AAAGCATCGA	CTCTTTCGACCGG	AAAGCATCGACCGG
AAAGTCGA	GAATATCGA	AAAGTCGACCGG	GAATATCGACCGG
GGAGGTCGA	CTCGATTCGA	GGAGGTCGACCGG	CTCGATTCGACCGG
TTAAATCGA	CGTGTCGA	TTAAATCGACCGG	CGTGTCGACCGG
TCTCTTCGA	CTCCCTCGA	TCTCTTCGACCGG	CTCCCTCGACCGG
CCGTACTCGA	CCCTGTCGA	CCGTACTCGACCGG	CCCTGTCGACCGG

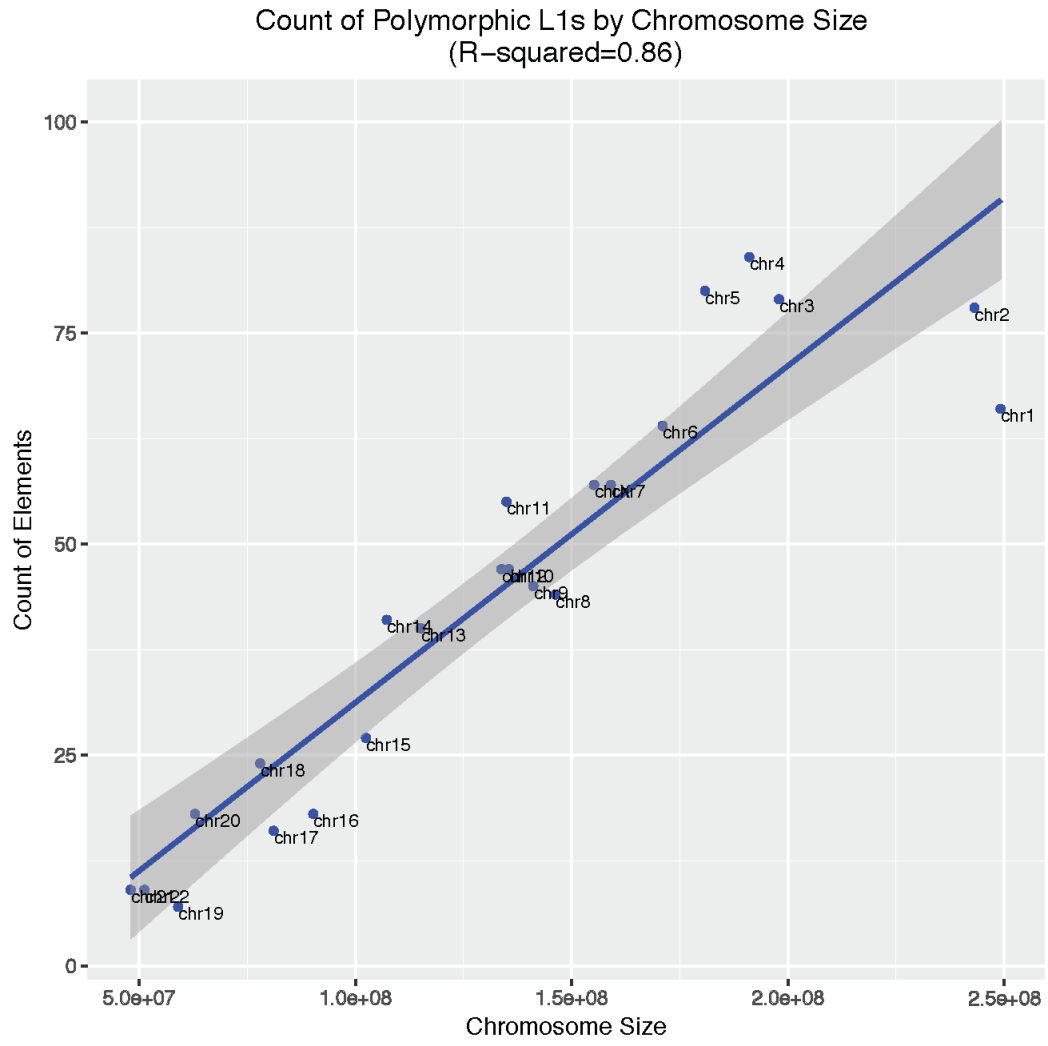
Figure S2. Karyotype plots of the chromosome-wise distribution of L1 elements.
Genomic position of *de novo*, polymorphic, and human-specific L1s. The height of the colored bars (Red: de novo L1s; Blue: polymorphic L1s; Green: human-specific L1s) above each chromosome represents the number of elements from the corresponding dataset. Cytogenetic band information was retrieved from the UCSC Genome Browser, which was used to annotate different chromosomal regions, as seen on Giemsa-stained chromosomes. Centromere regions are represented as red blocks.

Figure S3. Scatter plots of the chromosome-wise distribution of L1 elements.
Element count of **A.** *de novo* L1s; **B.** polymorphic L1s; **C.** human-specific L1s against chromosome size. Fitted regression lines and 95% confidence intervals are shown for each scatter plot. Correlation between the count of elements and chromosome sizes are indicated below the main title.

A



B


Count of Polymorphic L1s by Chromosome Size
(R−squared=0.86)
Count of Elements
Chromosome Size
100
75
50
25
0
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr4
chr5
chr3
chr2
chr1
chr6
chr11
chr7
chr8
chr14
chr13
chr15
chr18
chr20
chr16
chr17
chr19

C

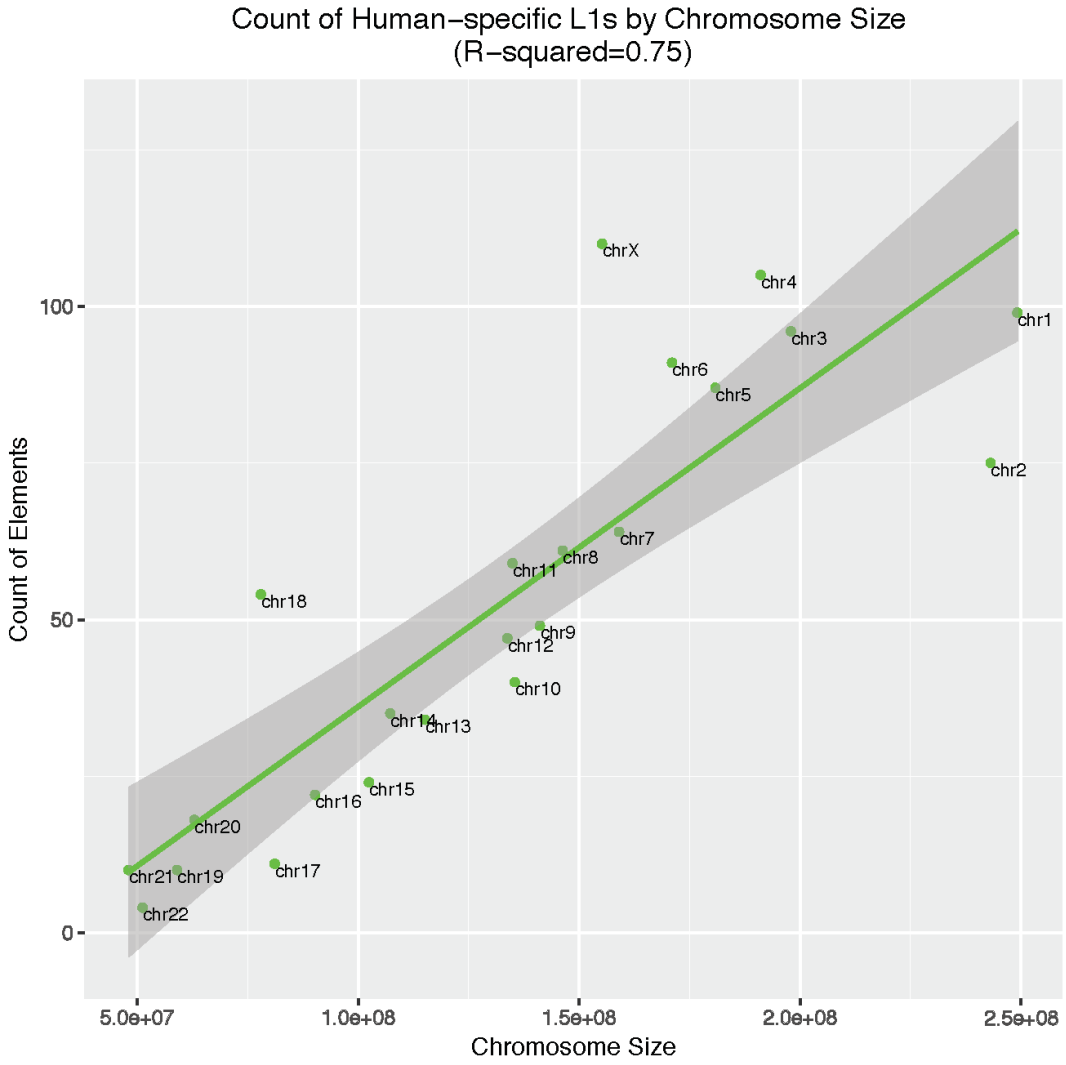
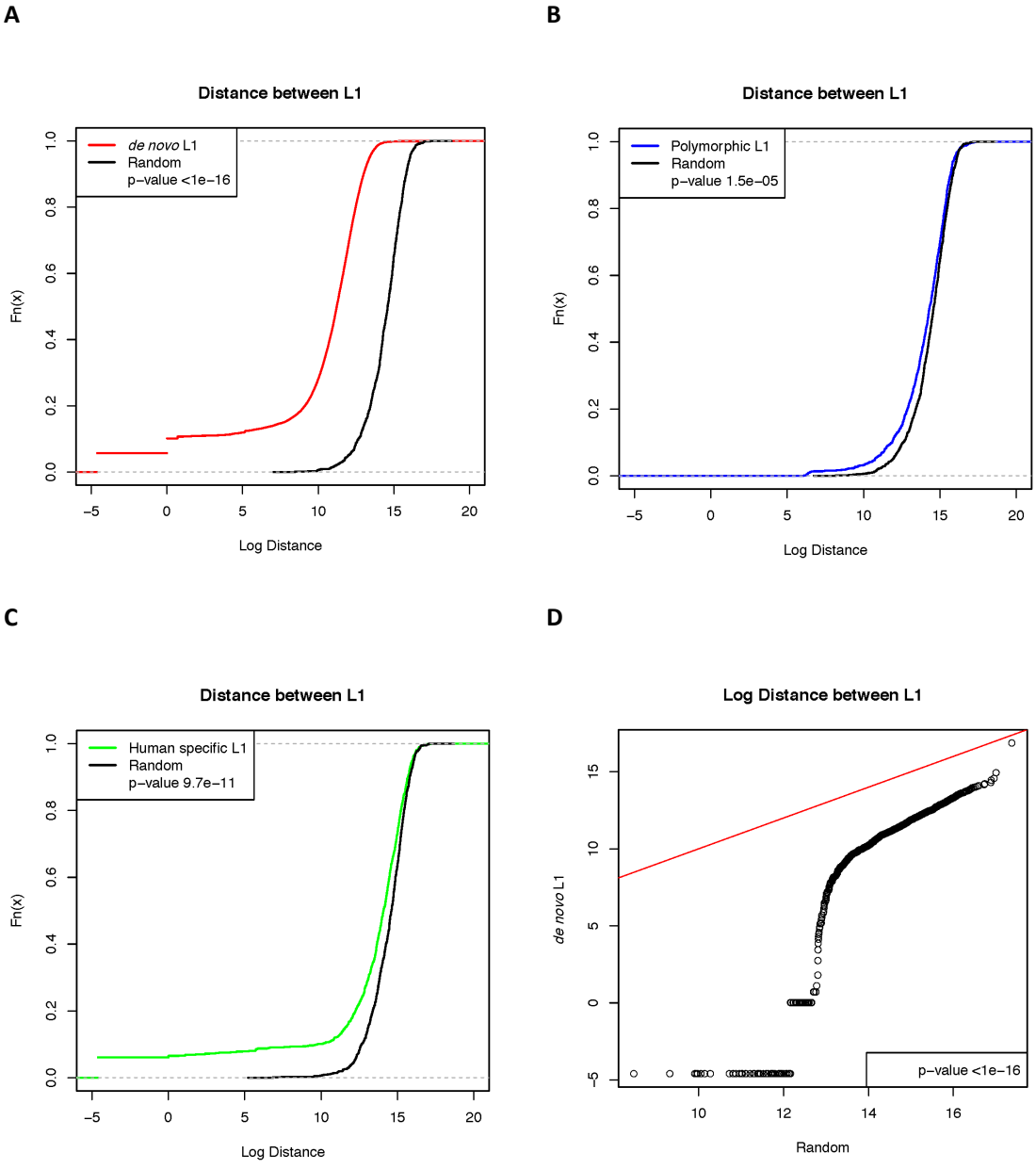
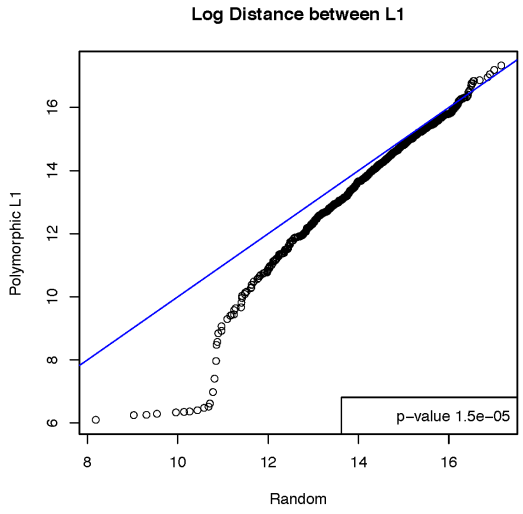

Count of Human−specific L1s by Chromosome Size
(R−squared=0.75)
Count of Elements
100
50
0
chrX
chr4
chr1
chr3
chr6
chr5
chr2
chr7
chr8
chr11
chr18
chr9
chr12
chr10
chr14
chr13
chr15
chr16
chr20
chr21
chr19
chr17
chr22
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
Chromosome Size

Figure S4. Distribution of distances between L1 elements of the same type, compared with the random distribution.
A-C: Cumulative distribution of the distance between L1 elements of the same type (red: *de novo* L1; blue: polymorphic L1; green: human-specific L1), compared with the expected cumulative distribution (black). **D-F**: Q-Q plots of the distance distribution between L1 elements of the same type (Y-axis), compared with the expected distribution (X-axis). Colored line represents the identity line. Each panel reports the *p*-value of a bootstrap Kolmogorov–Smirnov test for equality of observed and expected distributions. Distances are reported on a log scale.



E


Log Distance between L1
Polymorphic L1
Random
p−value 1.5e−05
6
8
10
12
14
16
8
10
12
14
16

F

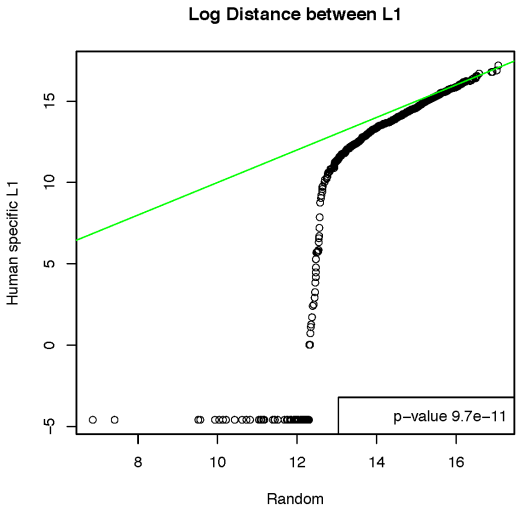
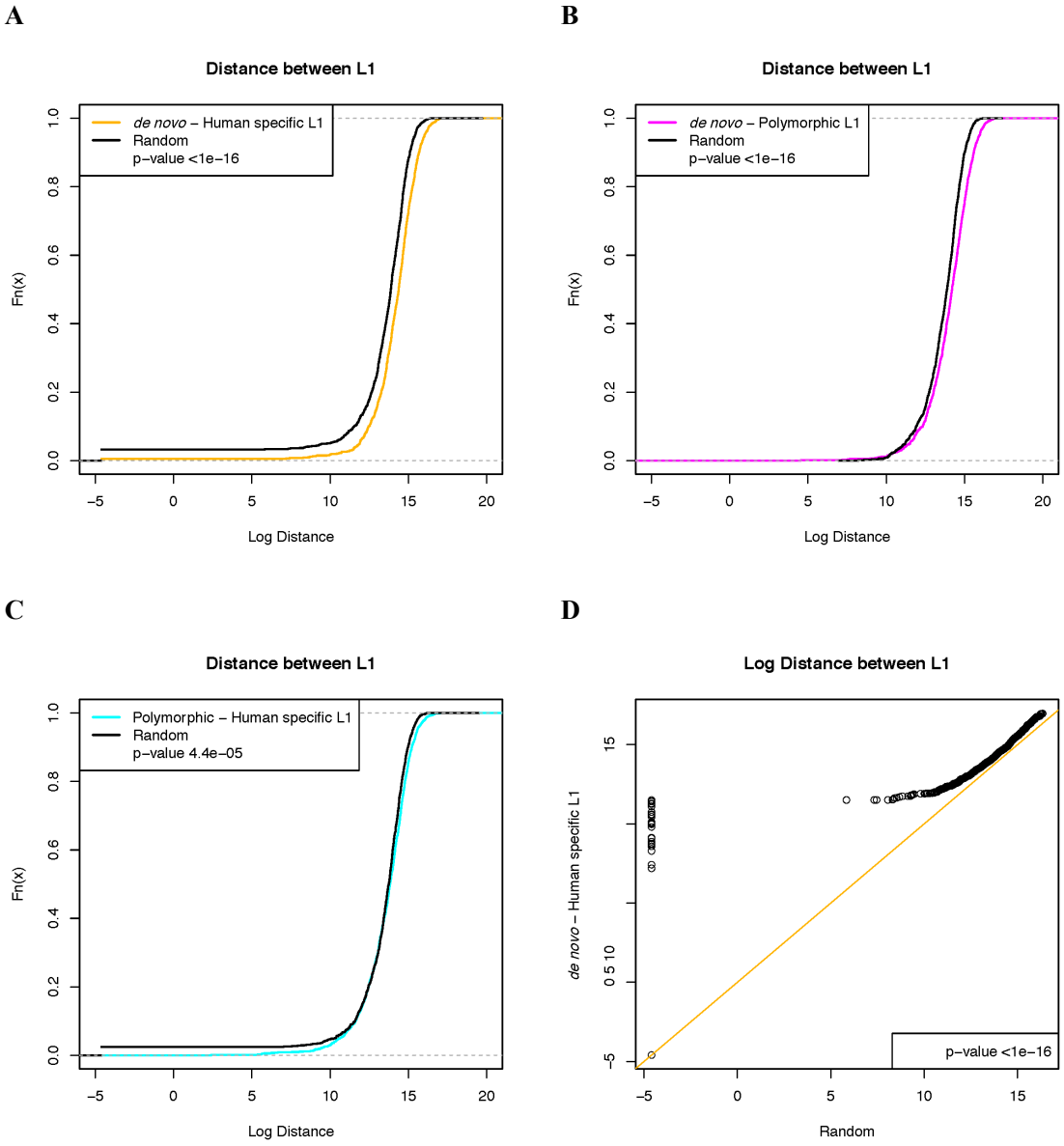
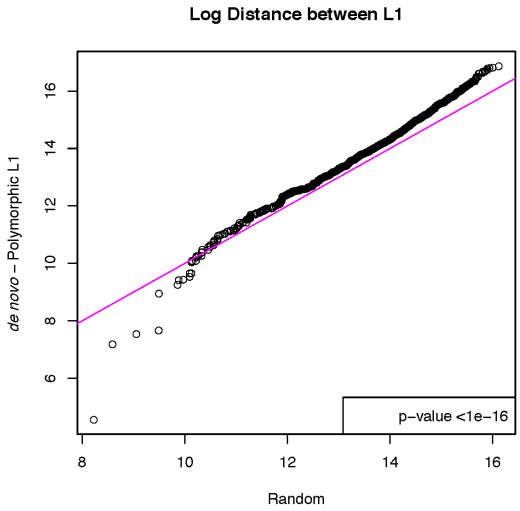

Log Distance between L1
Human specific L1
Random
p−value 9.7e−11
−5
0
5
10
15
8
10
12
14
16

Figure S5. Distribution of distances between L1 elements of different types, compared with the random distribution.

A-C. Cumulative distribution of the distance between L1 elements of different types (orange: *de novo* L1 and human-specific L1; magenta: *de novo* L1 and polymorphic L1; cyan: polymorphic L1 and human-specific L1, compared with the expected cumulative distribution (black). **D-F**. Q-Q plots of the distance distribution between L1 elements of different types (Y-axis), compared with the expected distribution (X-axis). The colored line represents the identity line. Each panel reports the *p*-value of a two-sample bootstrap Kolmogorov–Smirnov test for equality of observed and expected distributions. Distances are reported on a log scale.



E


Log Distance between L1
de novo – Polymorphic L1
Random
p–value <1e–16
6
8
10
12
14
16

F

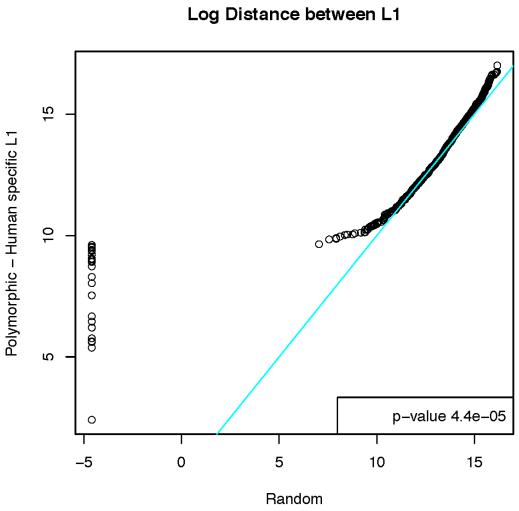

Log Distance between L1
Polymorphic – Human specific L1
Random
p–value 4.4e–05
5
10
15
–5
0
5
10
15

Figure S6. Clustering of correlations among high-resolution genomic features.
Hierarchical clustering based on Spearman's correlation among 45 high-resolution features. The red dashed line shows our threshold for high correlation between features (absolute value of Spearman's correlation above 0.8). Only one feature in each of the two clusters (shown in cyan and green) was kept for subsequent analyses.

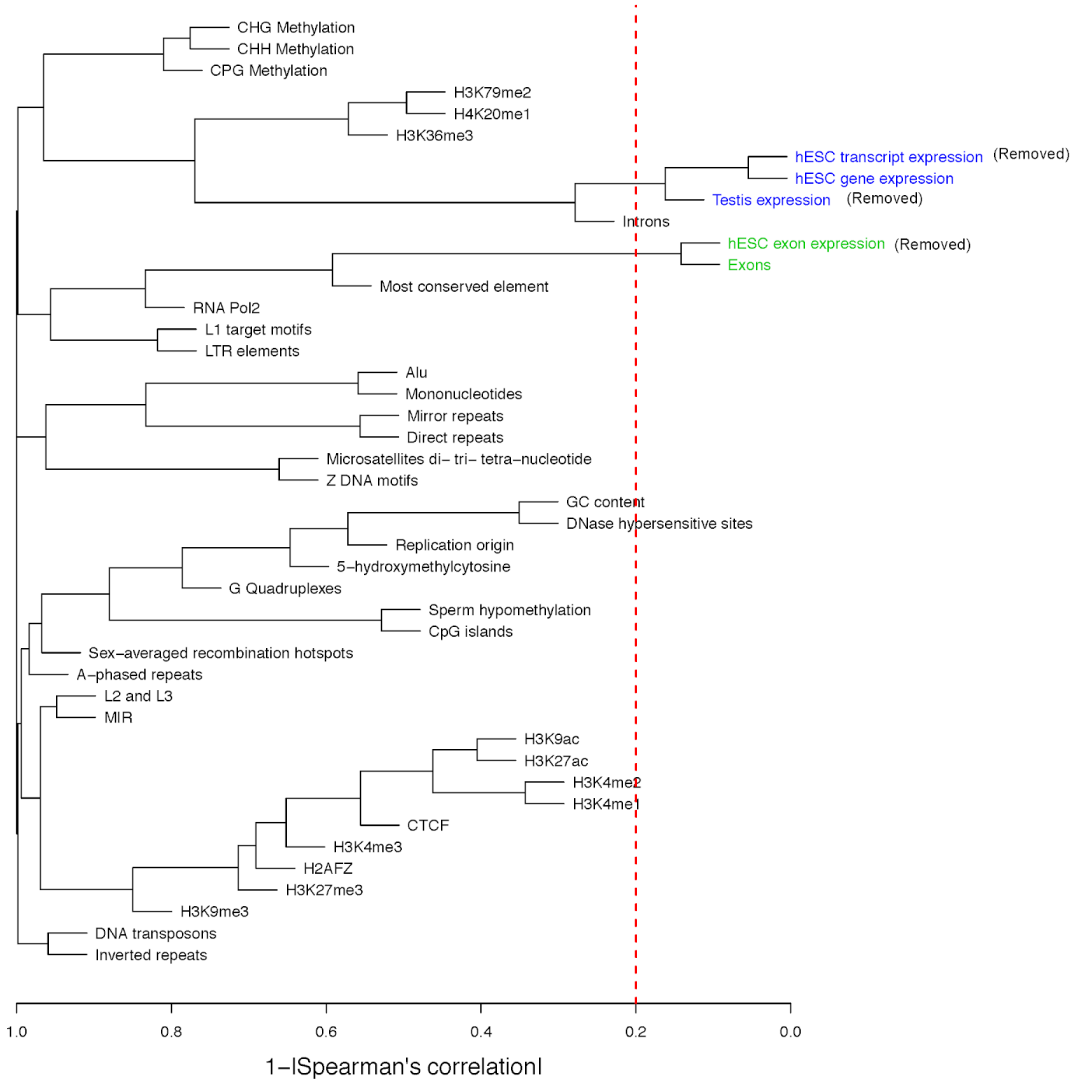
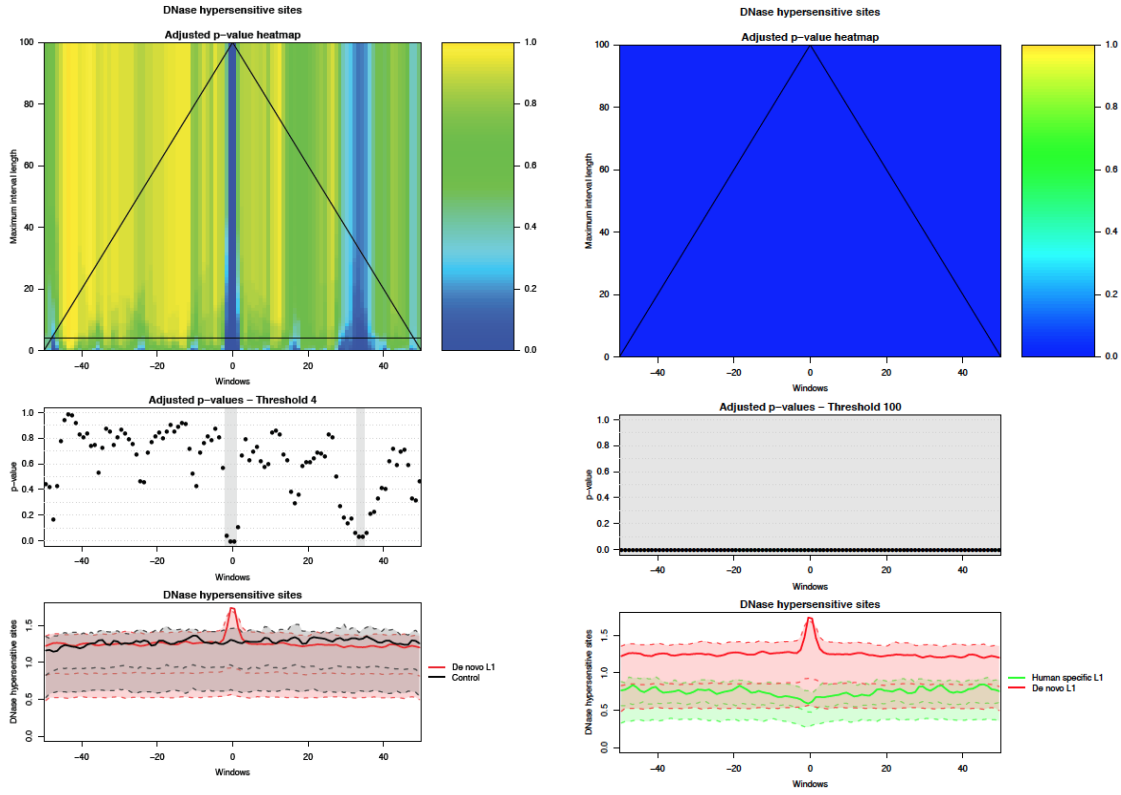
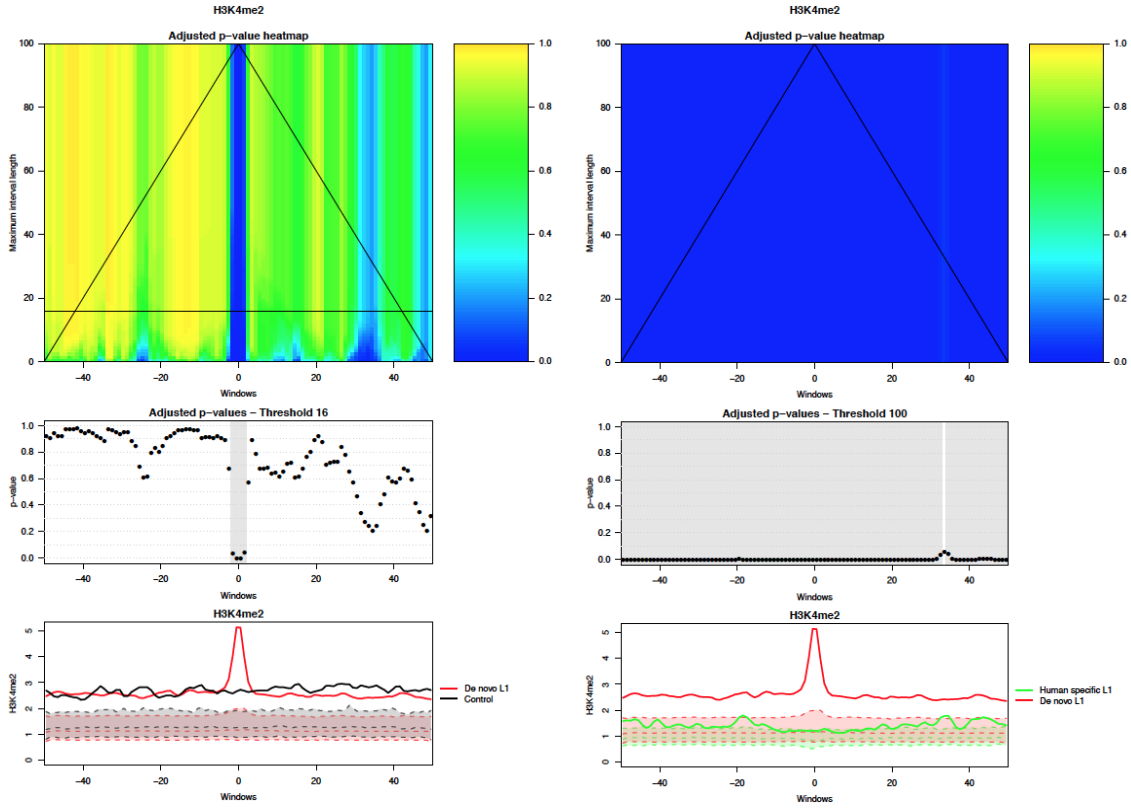


Figure S7. IWTomics results for selected genomic features.
Comparisons of genomic features between L1 flanking regions and controls (left: *de novo* L1 vs. control; right: human-specific L1 vs. *de novo* L1). **A.** DNase hypersensitive sites (DHS) **B.** H3K4me2 **C.** H3K9me3. In each figure, the top heat map shows the IWTomics adjusted *p*-value curves at all possible scales - X-axis reports the position in the flanking region, from -50 kb to +50 kb; Y-axis reports the scale used for *p*-value correction, from 1 kb (no correction) to 100 kb (considering the whole region). The central panel shows the adjusted *p*-value curve at the selected scale, with gray area indicating significant *p*-values (≤0.05). Bottom panel reports pointwise boxplots of the feature values (dotted curves: 25%, 50%, and 75% quantiles, solid curves: averages). The control regions contain less than 7% coverage by all annotated L1 elements.

A



B


H3K4me2
Adjusted p−value heatmap
Maximum interval length
Windows
H3K4me2
Adjusted p−value heatmap
Maximum interval length
Windows
Adjusted p−values − Threshold 16
p−value
Windows
Adjusted p−values − Threshold 100
p−value
Windows
H3K4me2
H3K4me2
De novo L1
Control
Windows
H3K4me2
H3K4me2
Human specific L1
De novo L1
Windows

C

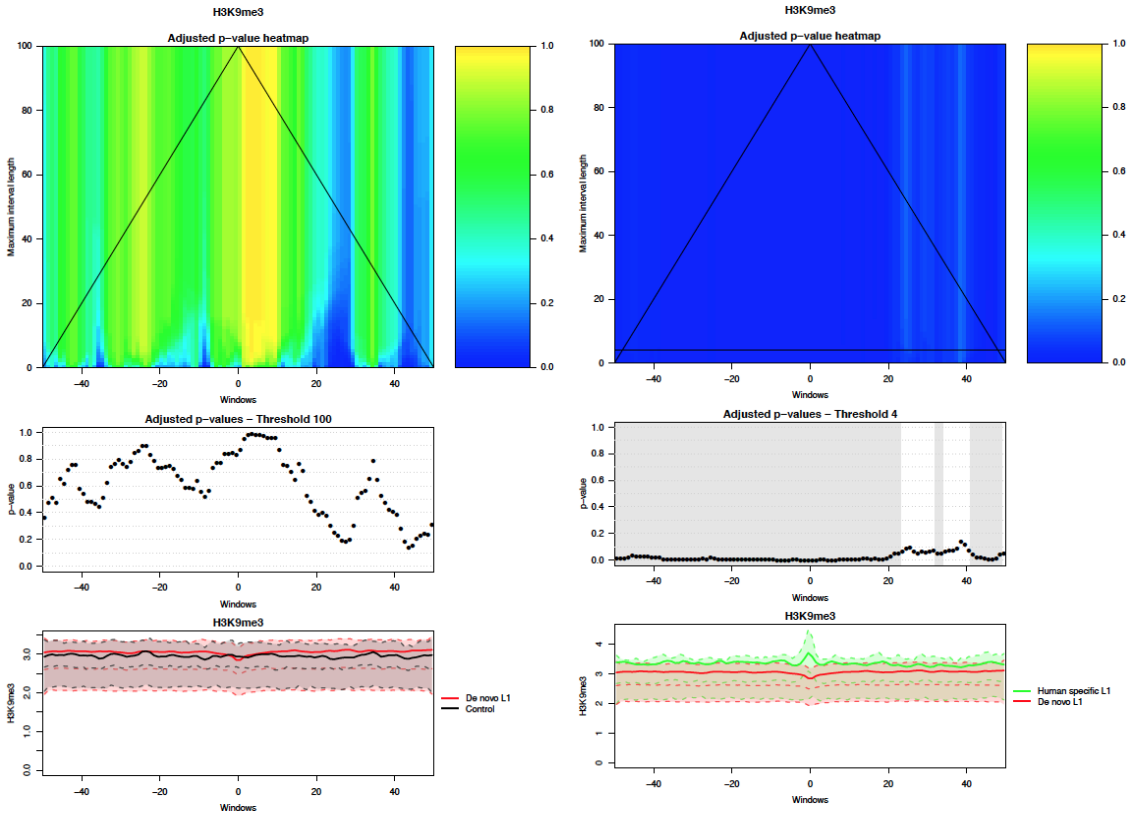
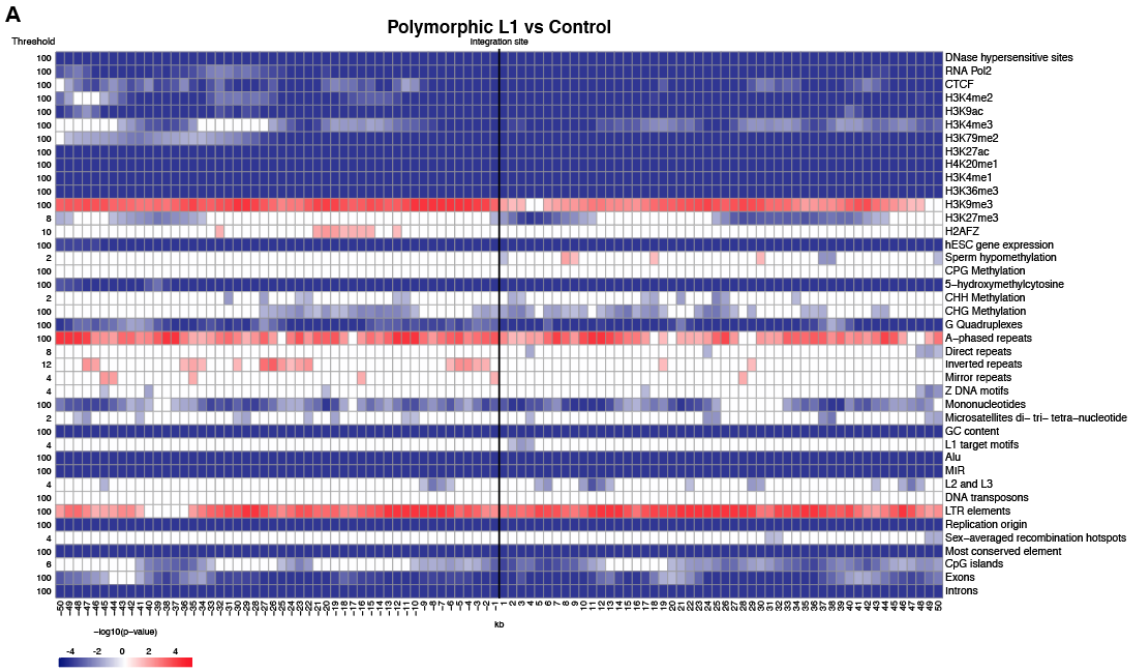
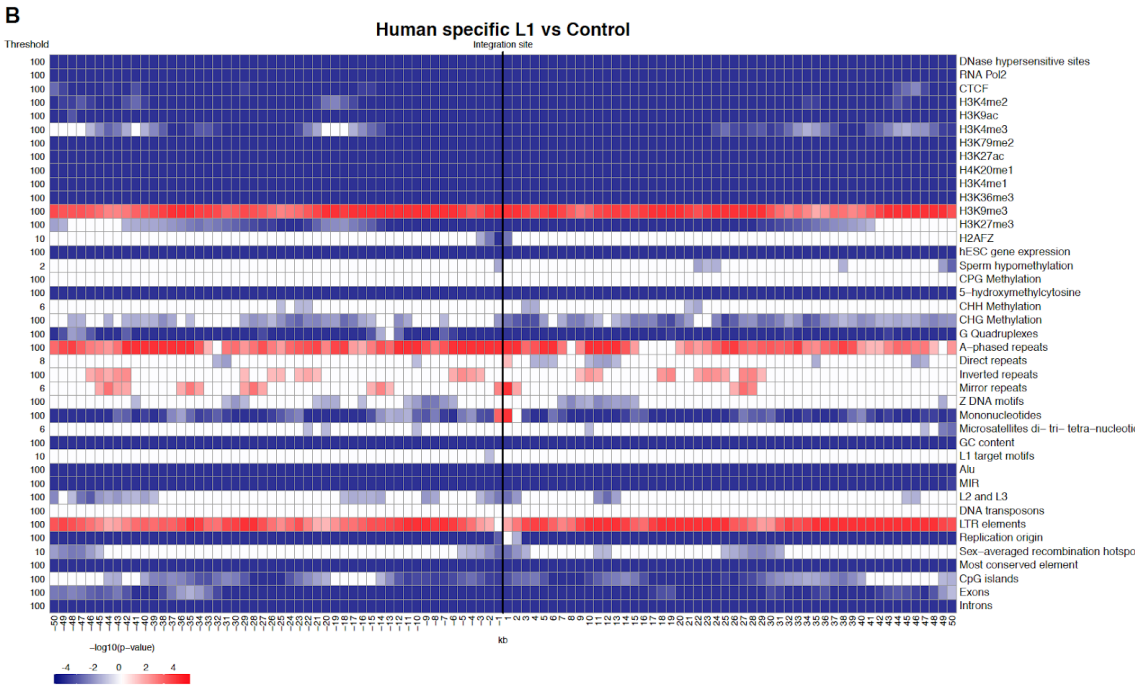
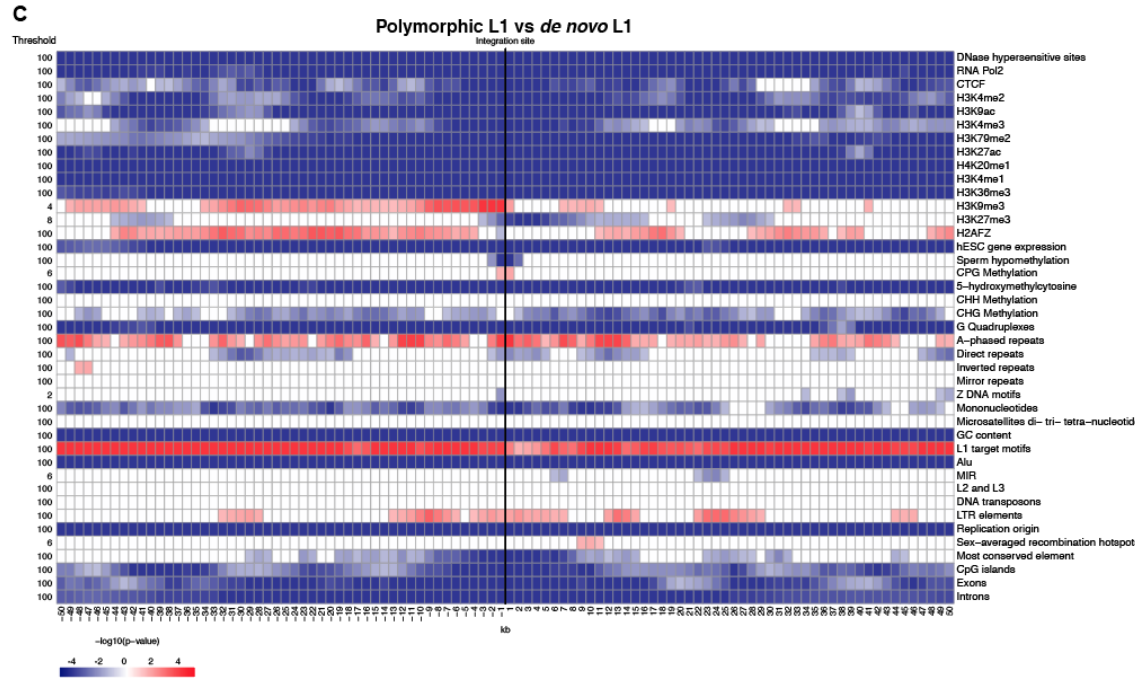

H3K9me3
Adjusted p−value heatmap
Maximum interval length
Windows
Adjusted p−values − Threshold 100
p−value
H3K9me3
H3K9me3
De novo L1
Control
H3K9me3
Adjusted p−value heatmap
Maximum interval length
Windows
Adjusted p−values − Threshold 4
p−value
H3K9me3
H3K9me3
Human specific L1
De novo L1

Figure S8. Summary of ITWomics results on high-resolution features.
A. Polymorphic L1 and control regions. **B.** Human-specific L1 and control regions. **C.** Polymorphic L1 and *de novo* L1 regions. **D.** Human-specific L1 and polymorphic L1 regions. The X-axis represents the position analyzed within the 100-kb flanking region of the L1 element, while each unit is a 1-kb window. The black vertical line across the center represents the insertion site. Each row represents one genomic feature and reports the adjusted *p*-value curve on a log10 scale. White: non-significant difference (*p*-value>0.05). Red: significant difference, with over-representation of the feature in L1 flanking regions. Blue: significant difference, with under-representation of the feature in L1 flanking regions. The selected IWTomics scale threshold is noted on the left. Note that the results of *de novo* L1 vs control are summarized by taking the median *p*-values over 10 random samples of the same size. The control regions contain less than 7% coverage by all annotated L1 elements.




B
Human specific L1 vs Control
Threshold
Integration site
DNase hypersensitive sites
RNA Pol2
CTCF
H3K4me2
H3K9ac
H3K4me3
H3K79me2
H3K27ac
H4K20me1
H3K4me1
H3K36me3
H3K9me3
H3K27me3
H2AFZ
hESC gene expression
Sperm hypomethylation
CPG Methylation
5−hydroxymethylcytosine
CHH Methylation
CHG Methylation
G Quadruplexes
A−phased repeats
Direct repeats
Inverted repeats
Mirror repeats
Z DNA motifs
Mononucleotides
Microsatellites di− tri− tetra−nucleoti
GC content
L1 target motifs
Alu
MIR
L2 and L3
DNA transposons
LTR elements
Replication origin
Sex−averaged recombination hotspo
Most conserved element
CpG islands
Exons
Introns
kb
−log10(p−value)
-4 -2 0 2 4


C
Polymorphic L1 vs de novo L1
Threshold
Integration site
DNase hypersensitive sites
RNA Pol2
CTCF
H3K4me2
H3K9ac
H3K4me3
H3K79me2
H3K27ac
H4K20me1
H3K4me1
H3K36me3
H3K9me3
H3K27me3
H2AFZ
hESC gene expression
Sperm hypomethylation
CPG Methylation
5−hydroxymethylcytosine
CHH Methylation
CHG Methylation
G Quadruplexes
A−phased repeats
Direct repeats
Inverted repeats
Mirror repeats
Z DNA motifs
Mononucleotides
Microsatellites di− tri− tetra−nucleotid
GC content
L1 target motifs
Alu
MIR
L2 and L3
DNA transposons
LTR elements
Replication origin
Sex−averaged recombination hotspot
Most conserved element
CpG islands
Exons
Introns
kb
−log10(p−value)
-4 -2 0 2 4

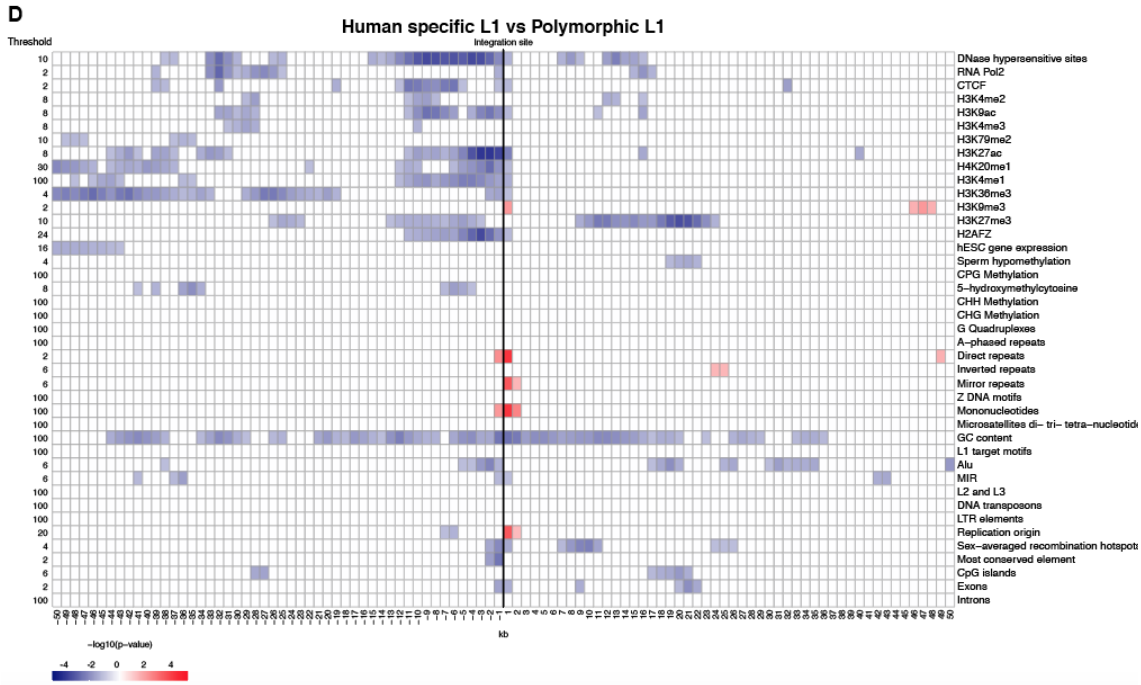

D
Human specific L1 vs Polymorphic L1
Threshold
Integration site
DNase hypersensitive sites
RNA Pol2
CTCF
H3K4me2
H3K9ac
H3K4me3
H3K79me2
H3K27ac
H4K20me1
H3K4me1
H3K36me3
H3K9me3
H3K27me3
H2AFZ
hESC gene expression
Sperm hypomethylation
CPG Methylation
5–hydroxymethylcytosine
CHH Methylation
CHG Methylation
G Quadruplexes
A–phased repeats
Direct repeats
Inverted repeats
Mirror repeats
Z DNA motifs
Mononucleotides
Microsatellites di– tri– tetra–nucleotid
GC content
L1 target motifs
Alu
MIR
L2 and L3
DNA transposons
LTR elements
Replication origin
Sex–averaged recombination hotspot
Most conserved element
CpG islands
Exons
Introns
kb
–log10(p–value)
-4 -2 0 2 4

Figure S9. IWTomics analysis on low-resolution genomic features (comparisons involving polymorphic L1s only).
A summary of the three pairwise comparisons on five low-resolution genomic features between polymorphic L1s and other datasets, using the univariate version of IWTomics. **A.** Distance to the telomere. **B.** Distance to the centromere. **C.** Replication timing. **D.** Recombination rate (sex-averaged). **E.** Count of telomere hexamers. Each panel represents the boxplot of the features in the different groups. 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 involving polymorphic L1s are noted at the bottom; significant ones (*p*-value<0.05) are in bold. The control regions contain less than 7% coverage by all annotated L1 elements.

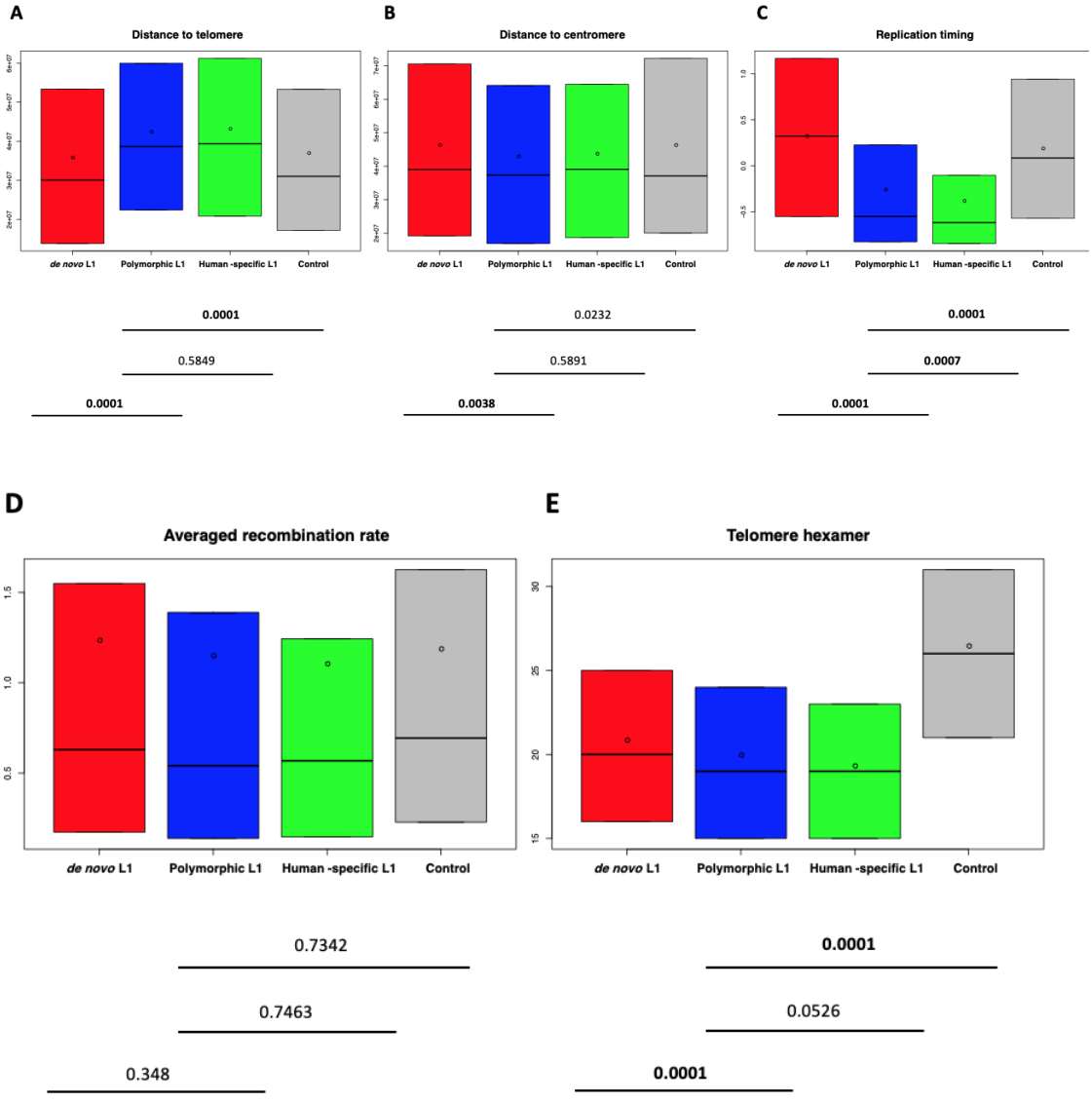
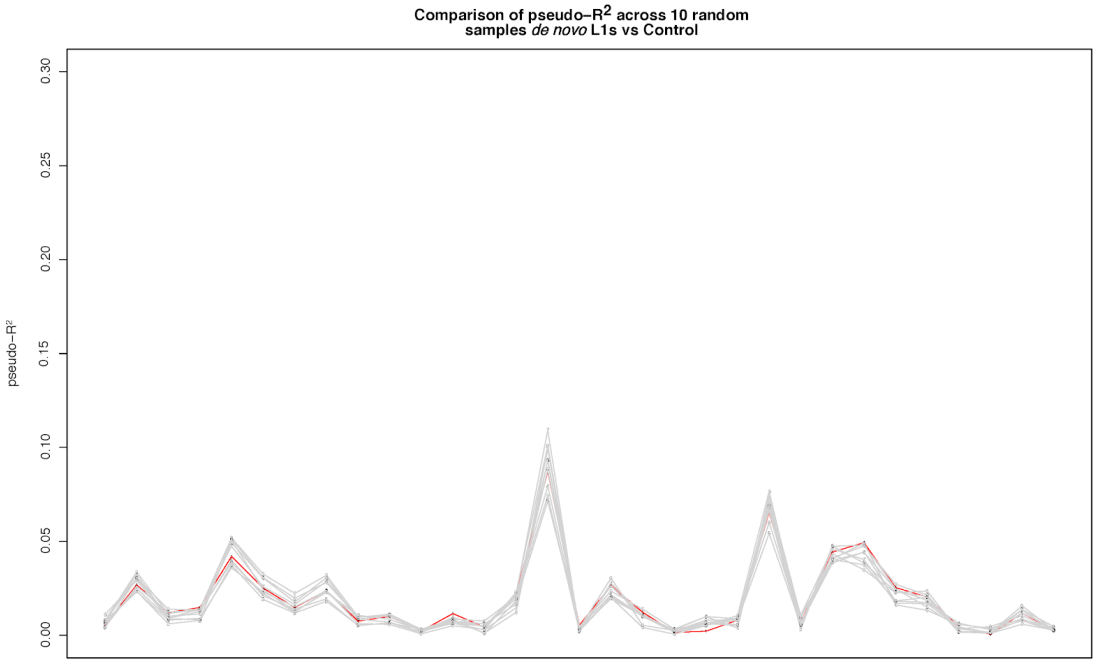
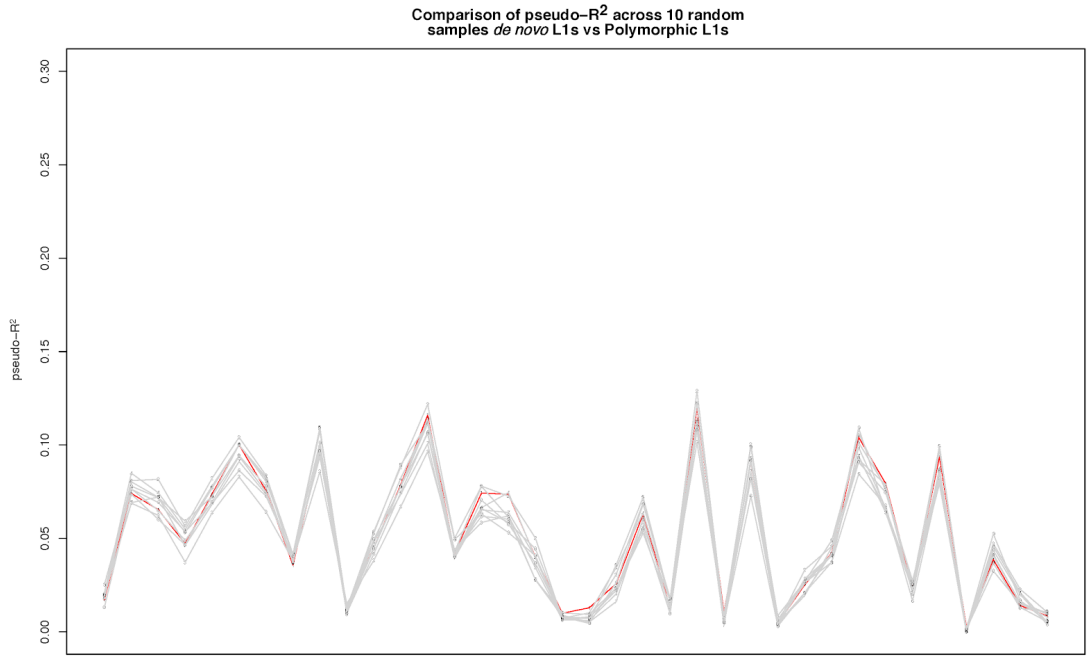


Figure S10. Selection of random subsample based on individual regression analysis.
Visualization of pseudo-R^2 for sFLR of the different genomic features across 10 random samples for the three pairwise comparisons involving the *de novo* L1 dataset (**A**. *de novo* L1s vs control; **B**. *de novo* L1s vs polymorphic L1s; **C**. *de novo* L1s vs human-specific L1s). Each point on X-axis represents a genomic feature, Y-axis is the pseudo-R^2 for the corresponding feature. Each line corresponds to a random subsample, while the red lines denote the random subsample 1 that was employed in subsequent analyses.

A



B


Comparison of pseudo−R2 across 10 random samples de novo L1s vs Polymorphic L1s
pseudo−R2
0.00
0.05
0.10
0.15
0.20
0.25
0.30

C

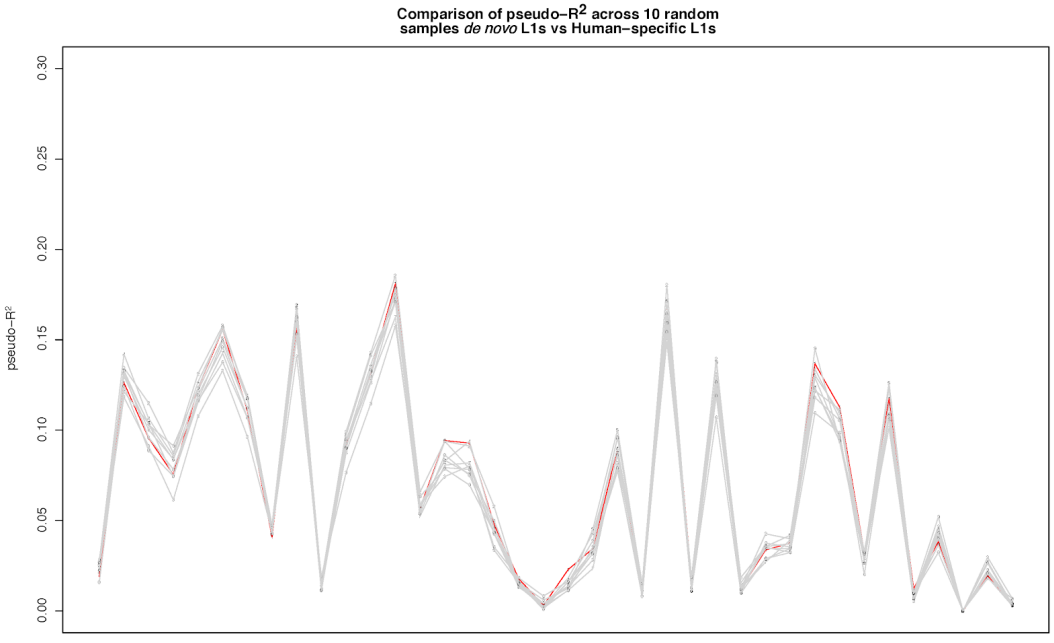
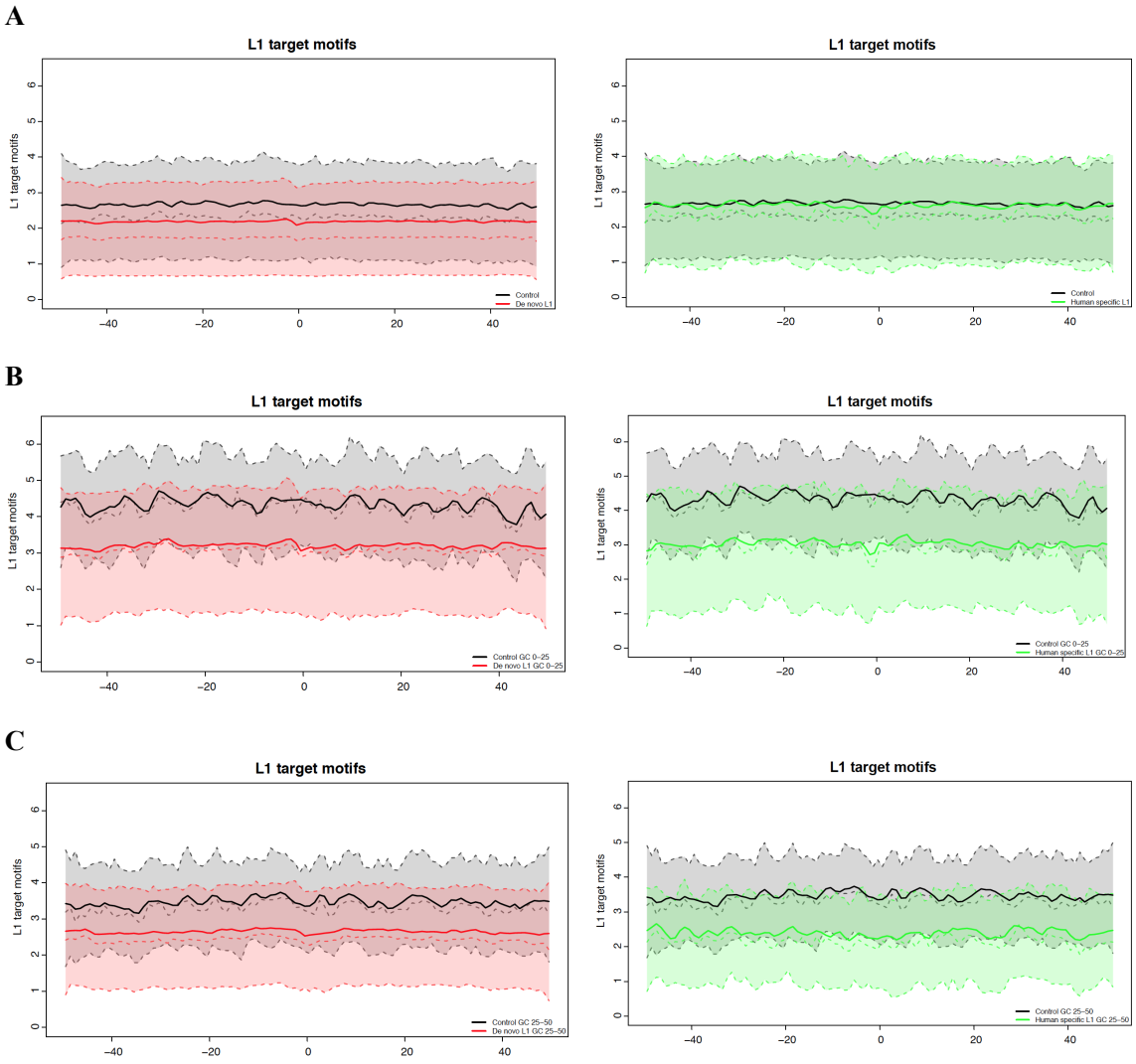

Comparison of pseudo−R2 across 10 random samples de novo L1s vs Human−specific L1s
pseudo−R2
0.00
0.05
0.10
0.15
0.20
0.25
0.30

Figure S11. Visualization of L1 target motifs after matching for GC content.
Comparisons of L1 target motifs between L1 flanking regions (left: *de novo* L1; right: human-specific L1) and controls, while matching for GC content. **A.** All regions. **B.** Regions with GC content between 0-25% quantiles. **C.** Regions with GC content between 25-50% quantiles. **D.** Regions with GC content between 50-75% quantiles. Regions with GC content between 75-100% quantiles are not reported, given the limited sample size of human-specific L1s in this category. In each pointwise boxplot, X-axis reports the position in the flanking region, from -50 kb to +50 kb; Y-axis reports the value of L1 target motif count. Each plot reports the different quantiles (dotted curves - 25%, 50%, and 75% quantiles) and averages (solid curves) of the feature values. The control regions contain less than 7% coverage by all annotated L1 elements.



D

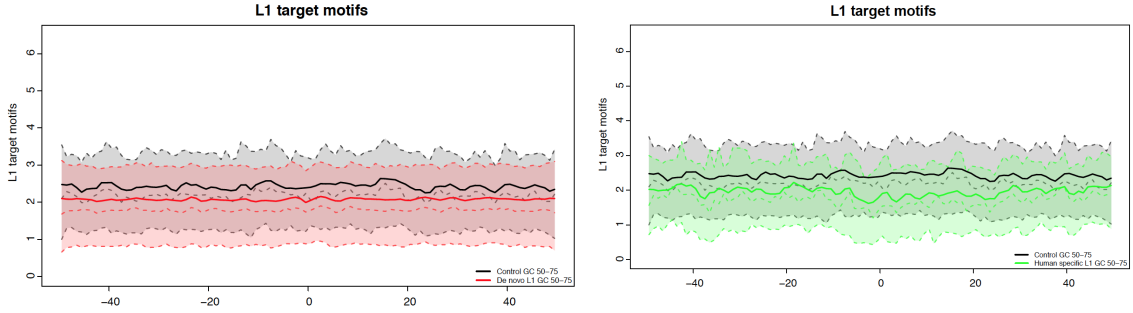
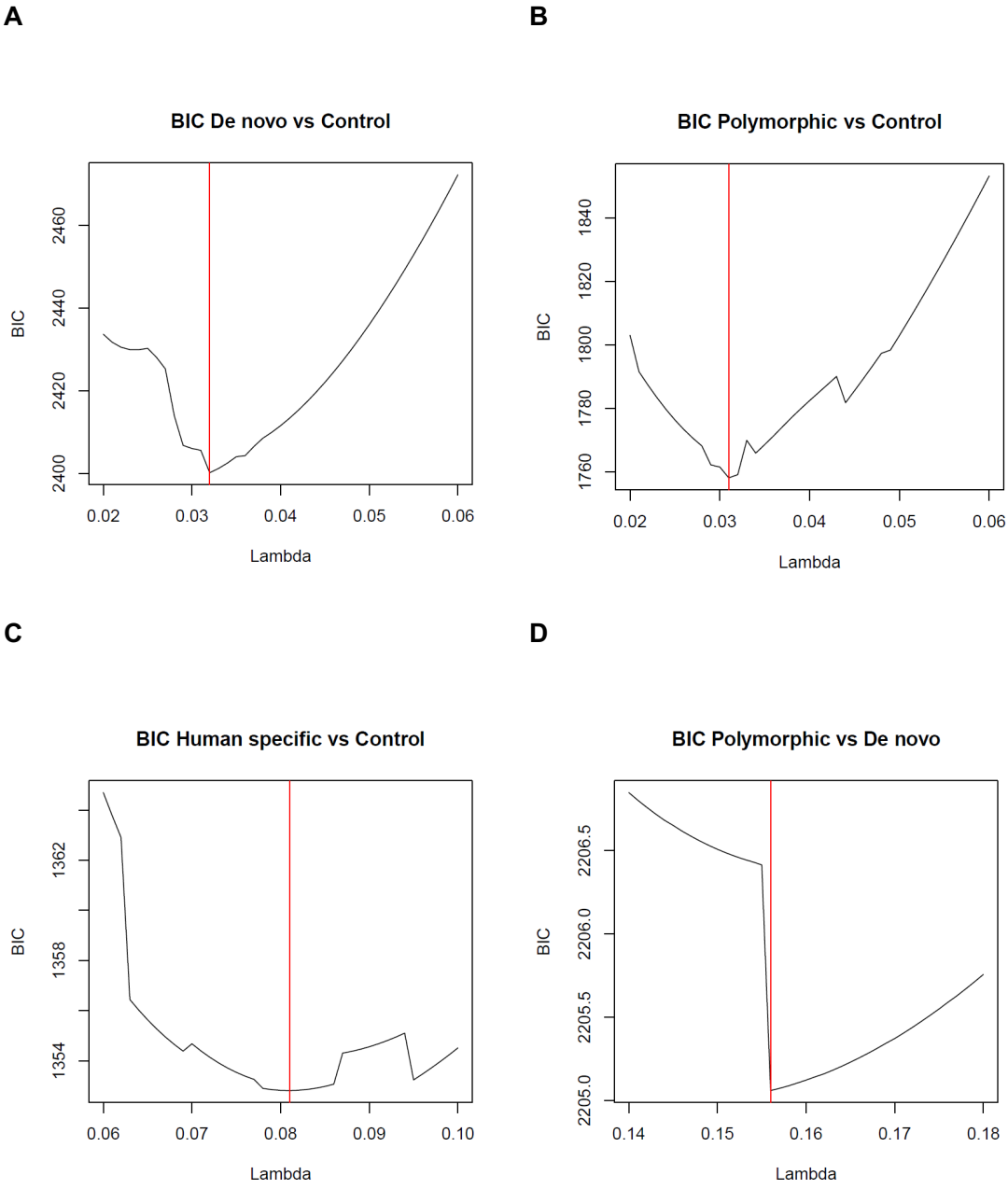
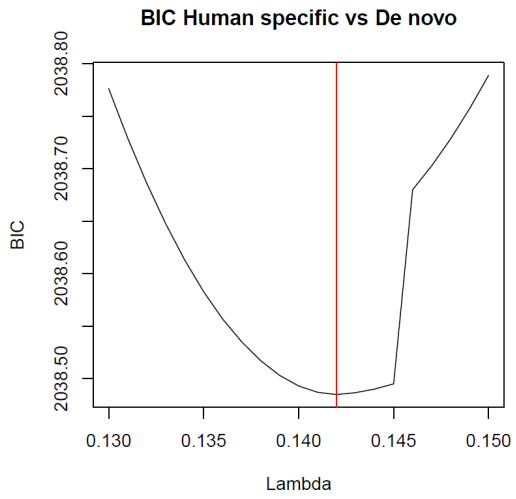

L1 target motifs
L1 target motifs
Control GC 50–75
De novo L1 GC 50–75
L1 target motifs
L1 target motifs
Control GC 50–75
Human specific L1 GC 50–75

Figure S12. Regularization parameter for variable selection in mFLR.
Plot of BIC corresponding to different values of the regularization parameter, for variable selection in mFLR models related to different comparisons. **A.** *de novo* L1 vs control regions **B.** Polymorphic L1 vs control regions. **C.** Human-specific L1 vs control regions. **D.** Polymorphic L1 vs *de novo* L1 regions. **E.** Human-specific L1 vs *de novo* L1 regions. **F.** Human-specific L1 vs polymorphic L1 regions. The red vertical line indicates the minimum BIC.



E


BIC Human specific vs De novo
BIC
2038.50
2038.60
2038.70
2038.80
0.130
0.135
0.140
0.145
0.150
Lambda

F

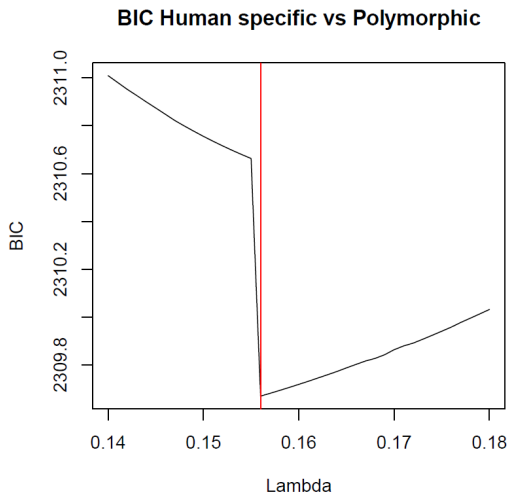
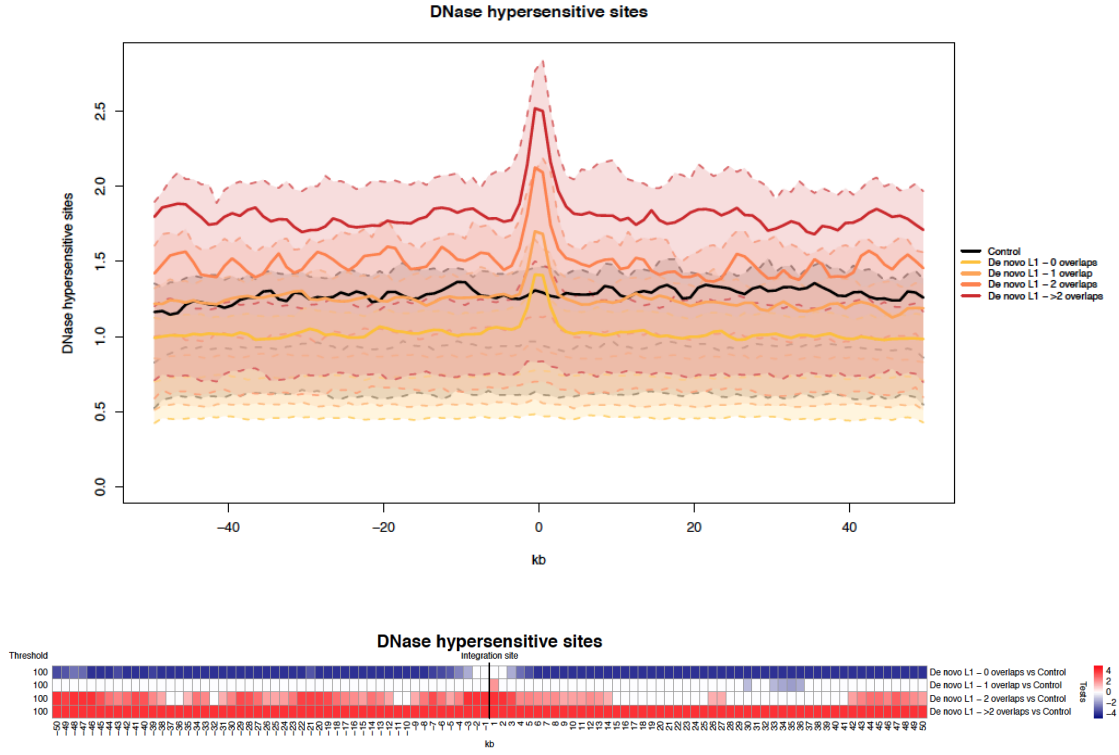
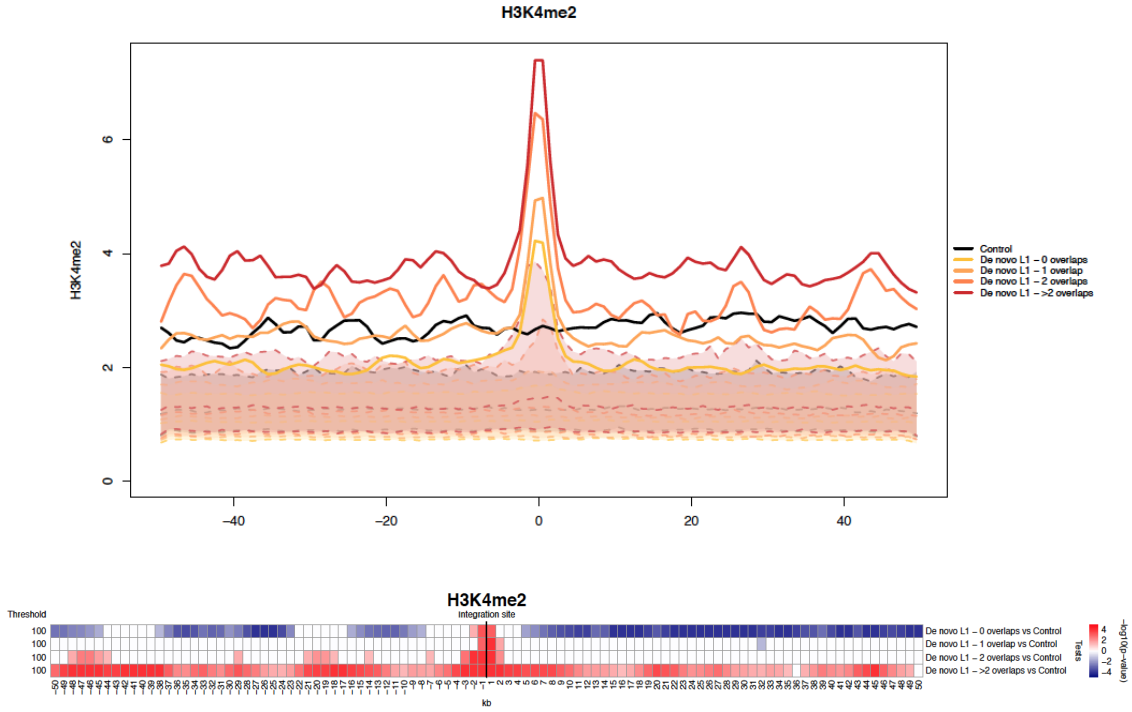

BIC Human specific vs Polymorphic
BIC
2309.8
2310.2
2310.6
2311.0
0.14
0.15
0.16
0.17
0.18
Lambda

Figure S13. Interval-Wise Testing analysis for selected genomic features at *de novo* L1 hotspots. Comparisons of genomic features between *de novo* L1 flanking regions and control regions, where the former are parsed based on whether they overlap with other *de novo* L1 flanking regions; specifically, presenting no overlap, one, two or more than two overlaps. **A.** DNase Hypersensitive Sites (DHS) **B.** H3K4me2 **C.** H3K9me3. In each figure, the top panel reports pointwise boxplots of the feature values (dotted curves - 25%, 50%, and 75% quantiles) with averages (solid curves). *de novo* L1 flanking regions are represented by colored curves and control regions are shown in black curves. The bottom heat map shows the IWT adjusted p-value curves at all possible scales - the X-axis reports the position in the flanking region, from -50 kb to +50 kb; the Y-axis reports the scale used for p-value correction, from 1 kb (no correction) to 100 kb (considering the whole region). The control regions contain less than 7% coverage by all annotated L1 elements.

A



B


H3K4me2
H3K4me2
0
2
4
6
−40
−20
0
20
40
Control
De novo L1 – 0 overlaps
De novo L1 – 1 overlap
De novo L1 – 2 overlaps
De novo L1 – >2 overlaps
H3K4me2
Threshold
Integration site
100
100
100
100
De novo L1 – 0 overlaps vs Control
De novo L1 – 1 overlap vs Control
De novo L1 – 2 overlaps vs Control
De novo L1 – >2 overlaps vs Control
kb
Tests
−log10(p−value)

C

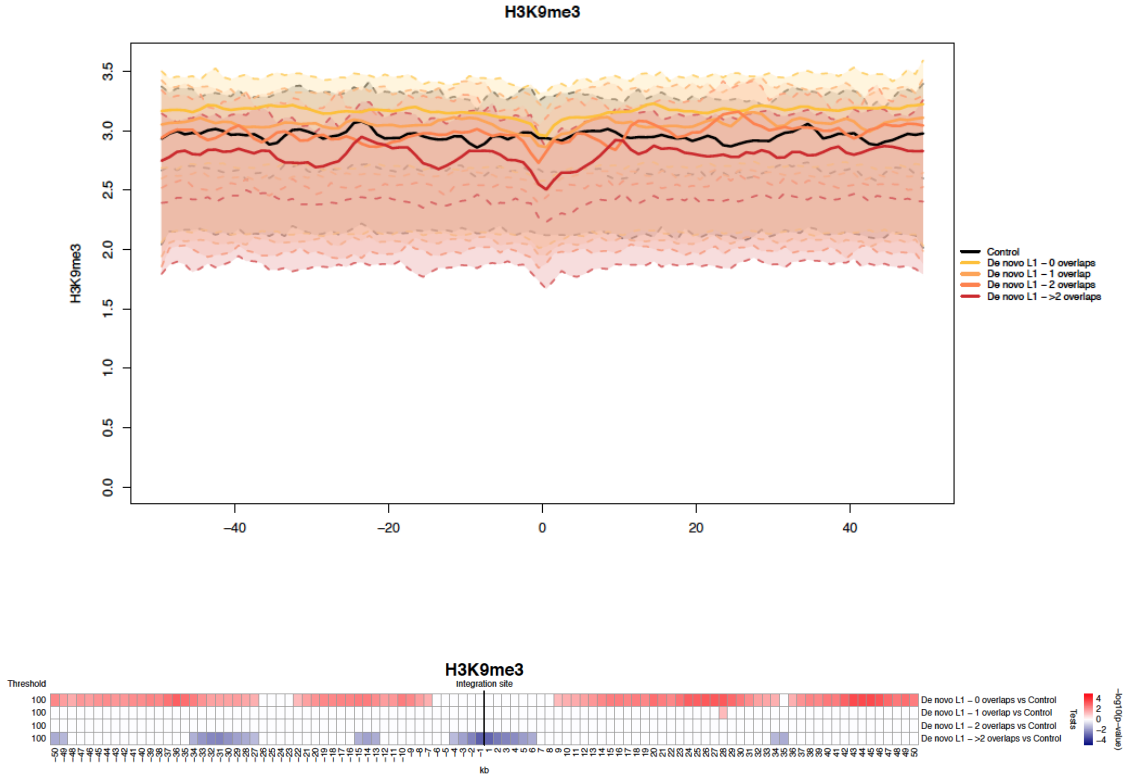

H3K9me3
H3K9me3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
−40
−20
0
20
40
Control
De novo L1 – 0 overlaps
De novo L1 – 1 overlap
De novo L1 – 2 overlaps
De novo L1 – >2 overlaps
H3K9me3
Threshold
Integration site
100
100
100
100
De novo L1 – 0 overlaps vs Control
De novo L1 – 1 overlap vs Control
De novo L1 – 2 overlaps vs Control
De novo L1 – >2 overlaps vs Control
kb
Tests
−log10(p−value)

Figure S14. Distribution of *de novo* L1s in HEK-293T duplicated sites.
Violin plots of observed *de novo* L1 insertion density (the number of insertions per base pair) in 146 Mb of HEK-293T duplicated sites (measured separately for each duplicated region) from (Lin et al. 2014) (right) and in the rest of the genome (left; measured in 100-kb windows), for comparison. The *p*-value of a t-test was 0.46, demonstrating no significant difference in mean L1 insertion density between duplicated sites and the genome background.

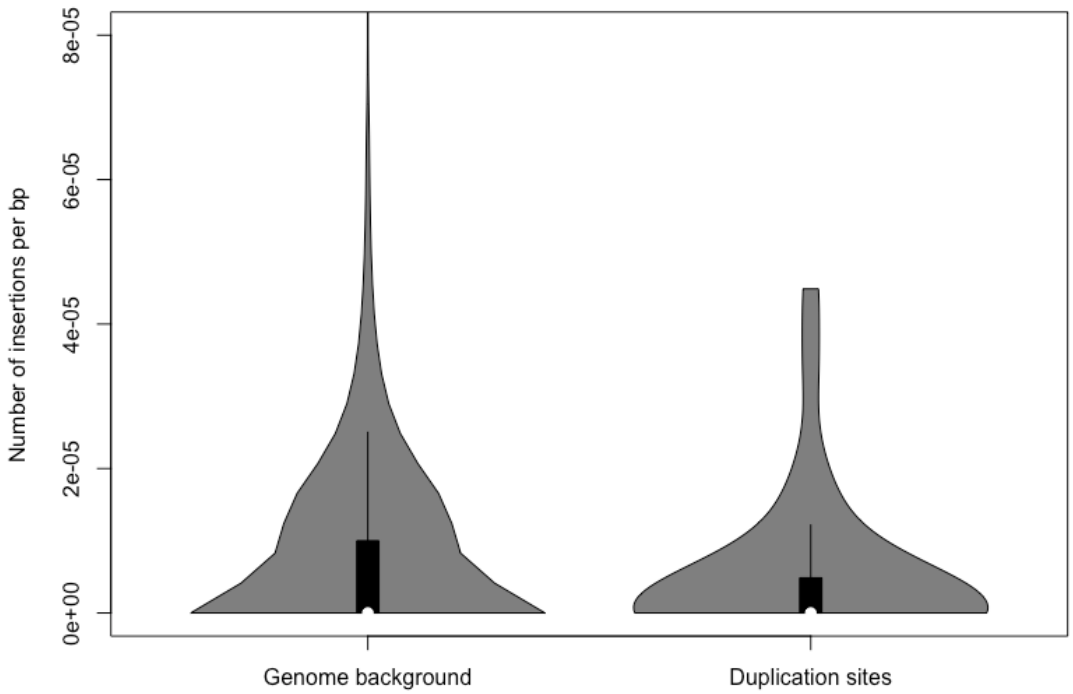
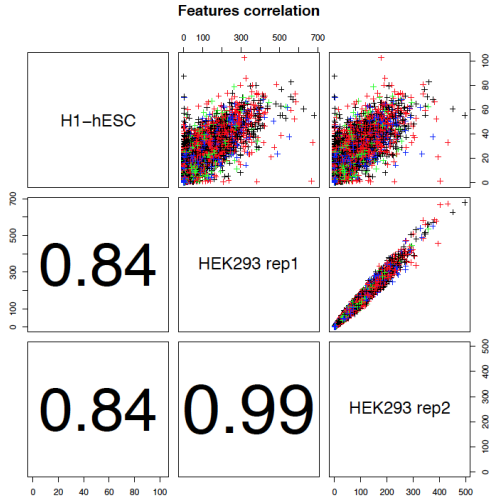
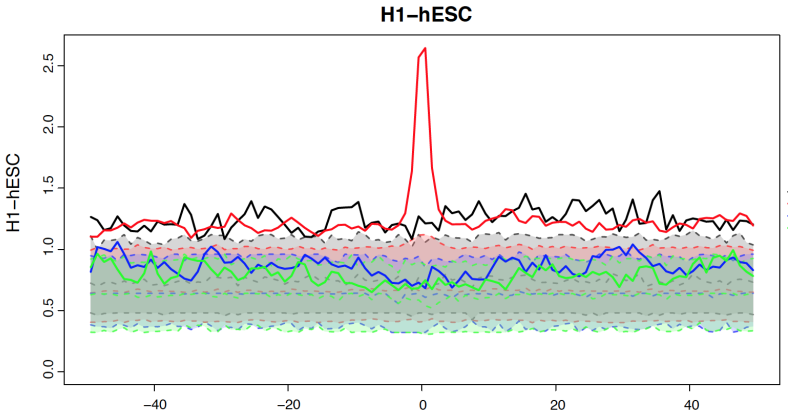


Figure S15. Comparisons of H3K4me3 signals generated from hESC versus HEK-293.
(A) A scatterplot matrix of the H3K4me3 signals in hESC and in two HEK-293 replicates, with the corresponding Peason's correlation coefficients (considering all 1-kb windows in our *de novo* L1 and control datasets). **(B-D)** Pointwise boxplots of the H3K4me3 signal values in L1 flanking regions and controls using data from **(B)** hESC; **(C)** HEK-293 replicate one; **(D)** HEK-293 replicate two. Boxplots of the feature values are shown in different quantiles (dotted curves: 25%, 50%, and 75% quantiles, solid curves: averages). The same color scheme is used in boxplots and scatterplots (Black: control; Red: *de novo* L1; Blue: polymorphic L1; Green: human-specific L1). The control regions contain less than 7% coverage by all annotated L1 elements.

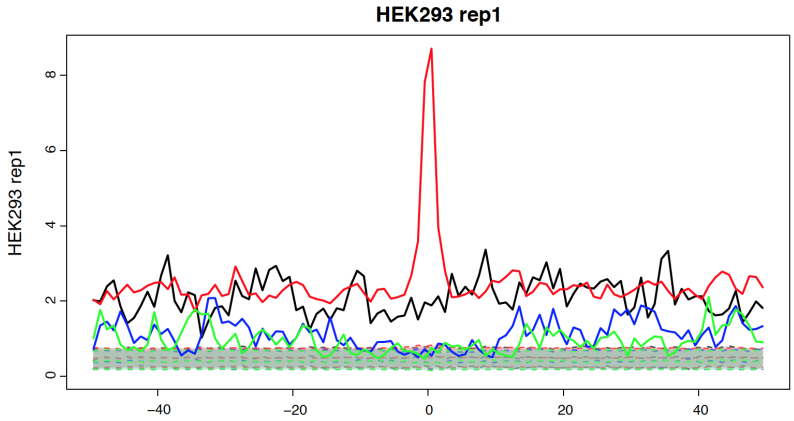
A



B



C


HEK293 rep1
HEK293 rep1
0
2
4
6
8
−40
−20
0
20
40

D

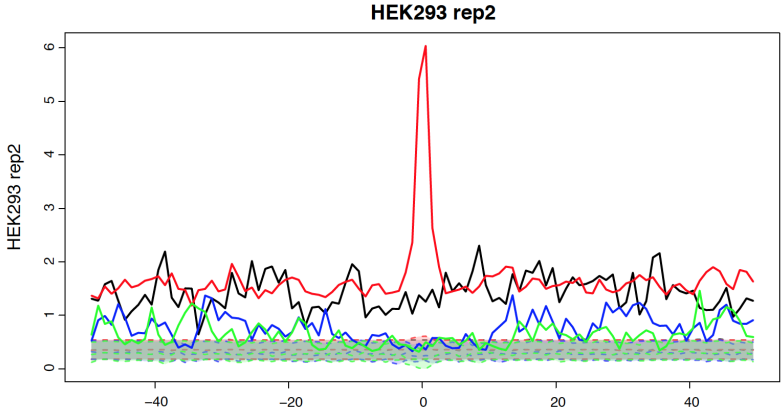
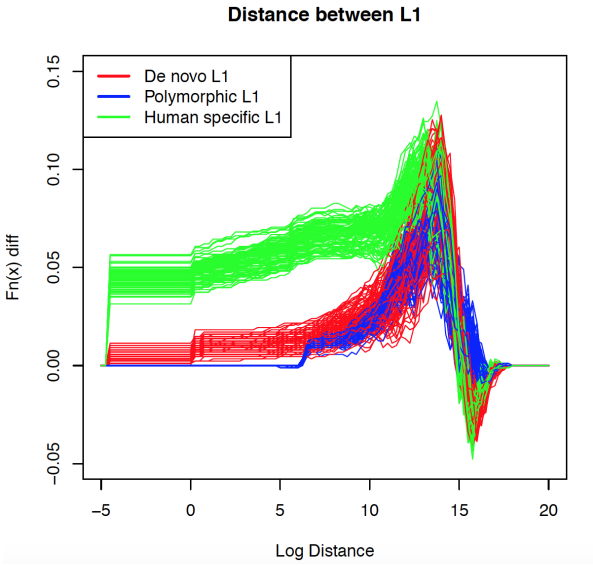
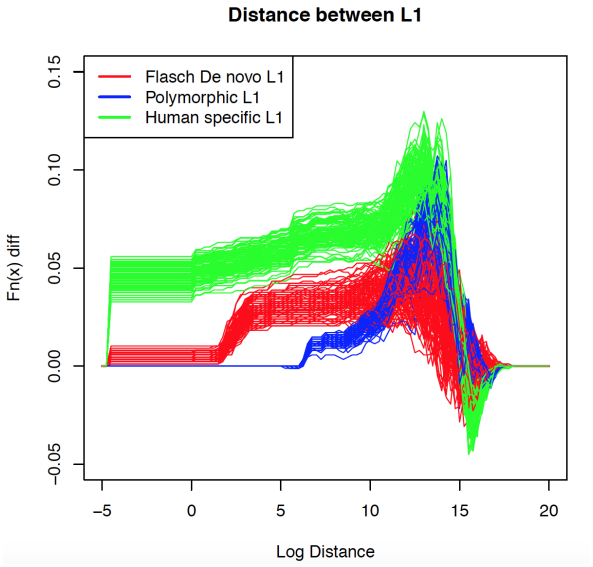

HEK293 rep2
HEK293 rep2
0
1
2
3
4
5
6
−40
−20
0
20
40

Figure S16. Comparison of distance distributions of *de novo* L1s across three studies.
A. Differences between observed and expected cumulative distributions of the distances between L1 elements of the same type (*de novo*, polymorphic, or human-specific), using **A.** *de novo* L1s obtained from HEK-293T in this study. **B.** *de novo* L1s obtained in hESC by Flasch et al. (2019). **C.** *de novo* L1s generated from Hela cells by Sultana et al. (2019).

A



B



C

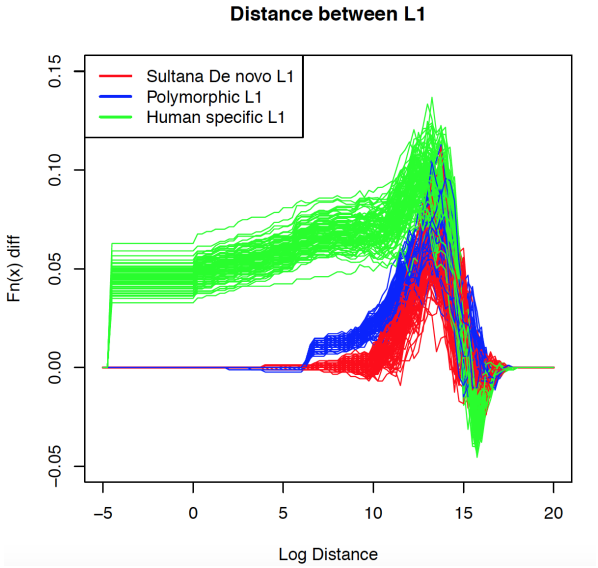
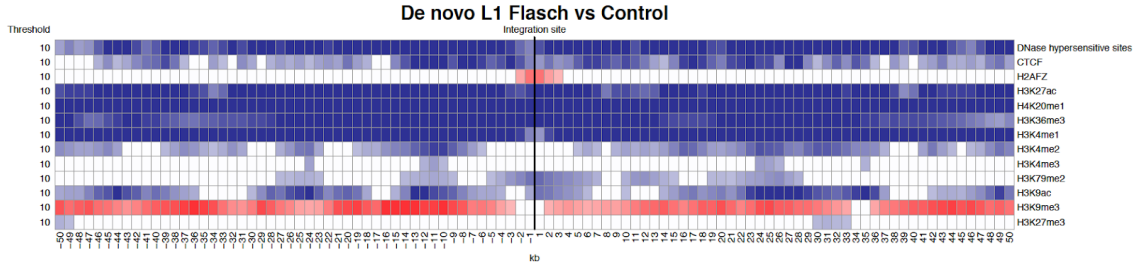

Distance between L1
Sultana De novo L1
Polymorphic L1
Human specific L1
Fn(x) diff
0.15
0.10
0.05
0.00
−0.05
−5
0
5
10
15
20
Log Distance

Figure S17. IWT analysis of selected features using *de novo* L1s from recent studies. Comparisons of *de novo* L1 vs control regions using *de novo* L1s in **A.** hESCs, from (Flasch et al. 2019), and **B.** HeLa, from (Sultana et al. 2019). The X-axis reports the position in the 100-kb flanking region (+50-kb, -50-kb) of the L1 element (the black vertical line across the center represents the insertion site). Each row represents one genomic feature and reports the adjusted *p*-value curve on a log10 scale. White: non-significant difference (*p*-value>0.05). Red: significant difference, with over-representation of the feature in L1 flanking regions. Blue: significant difference, with under-representation of the feature in L1 flanking regions. The IWTomics scale threshold is noted on the left. The control regions contain less than 7% coverage by all annotated L1 elements.

A



B

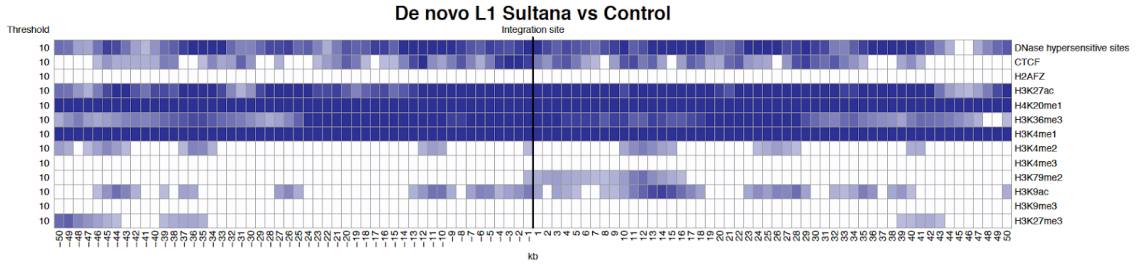


Figure S18. Distribution of distances between the two restriction sites *MspI* and *TaqI*.
Histogram showing the distribution of distances between any *MspI* or *TaqI* restriction sites in the genome. Log10 distances (with a +0.0001 shift) are plotted. The blue vertical line indicates the 150 bp threshold on this transformed scale. 87.46% of the data are below this threshold.

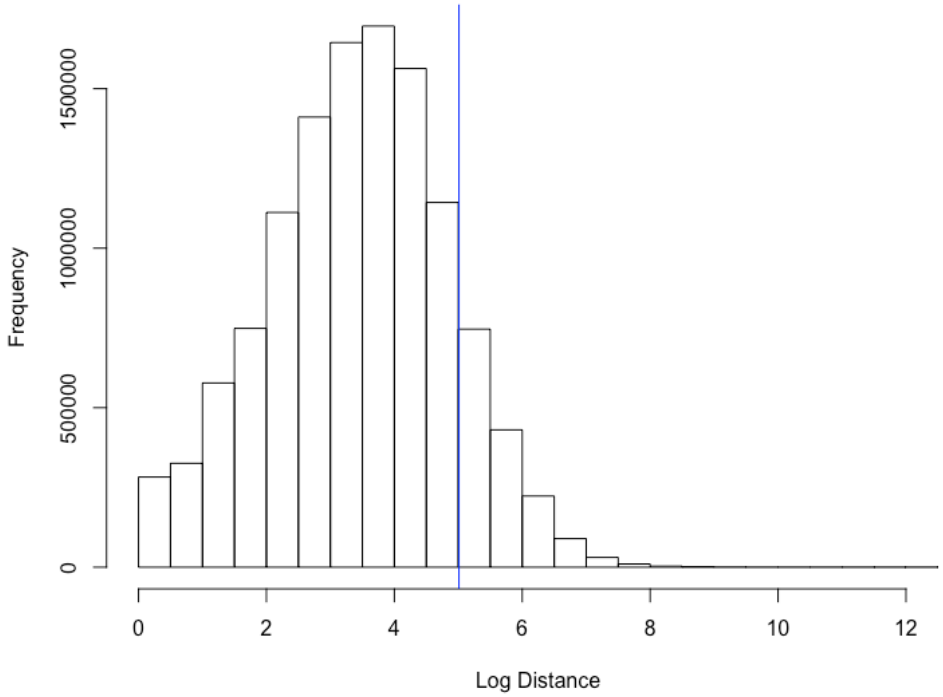


Table S1. L1 Datasets and construction of 100-kb L1 regions.

L1 Dataset	*de novo* L1	Polymorphic L1	Human-specific L1
Total number of elements analyzed	17,037	1,012	1,205
Number of autosomal elements	16,322	954	1,094
Number of 100-kb autosomal non-overlapping flanking regions after filtering gaps and blacklist	7,981	836	834
Number of 100-kb non-overlapping flanking regions on autosomes with strand information	7,981	670	725

Table S2. Genomic landscape features and their contribution to single and multiple Functional Logistic Regression for four non-major comparisons.

Chromatin=chromatin structure, Transcription = transcription regulation and gene expression, Res.=Resolution. "Not sign." cells indicate features that showed no significant differences in IWTomics tests; "Not sel." cells in mFLR columns indicate features that were not selected in the final mFLR models. Three features, including Testis Gene Expression (Brawand et al. 2011), Exon Expression (UCSC), and Transcript Expression (UCSC) were excluded in the analysis, due to their high correlation with the Gene Expression dataset (Spearman's correlation coefficient >0.8)(Fig. S4).

Group	Name	Polymorphic L1 vs control		Human-specific L1 vs control		Polymorphic L1 vs *de novo* L1		Human-specific L1 vs polymorphic L1	
		pseudo-R^2 for sFLR (%)	RCDE for mFLR (%)	pseudo-R^2 for sFLR (%)	RCDE for mFLR (%)	pseudo-R^2 for sFLR (%)	RCDE for mFLR (%)	pseudo-R^2 for sFLR (%)	RCDE for mFLR (%)
Chromatin	DNase hypersensitive sites	12.31	Not sel.	19.20	Not sel.	11.60	4.88	1.81	100
Chromatin	RNA Pol II	2.43	Not sel.	3.78	1.28	4.21	Not sel.	0.38	Not sel.
Chromatin	CTCF	17.05	Not sel.	24.53	Not sel.	7.87	Not sel.	1.58	Not sel.
Transcription	H3K4me2	20.83	Not sel.	28.09	Not sel.	10.01	Not sel.	0.60	Not sel.
Transcription	H3K9ac	22.54	Not sel.	33.05	0.36	9.78	Not sel.	1.24	Not sel.
Transcription	H3K4me3	14.40	Not sel.	19.00	Not sel.	7.56	Not sel.	0.28	Not sel.
Transcription	H3K79me2	7.54	Not sel.	8.28	Not sel.	3.57	Not sel.	0.03	Not sel.
Transcription	H3K27ac	19.31	Not sel.	28.25	Not sel.	7.41	Not sel.	1.38	Not sel.
Transcription	H4K20me1	14.54	Not sel.	20.76	Not sel.	6.54	Not sel.	0.89	Not sel.
Transcription	H3K4me1	21.36	0.29	29.20	1.04	7.38	2.48	1.12	Not sel.
Transcription	H3K36me3	11.20	Not sel.	16.41	Not sel.	4.72	Not sel.	0.62	Not sel.
Transcription	H3K9me3	0.84	Not sel.	1.16	Not sel.	1.38	Not sel.	0.50	Not sel.
Transcription	H3K27me3	9.11	1.29	14.84	Not sel.	4.53	Not sel.	0.63	Not sel.
Transcription	H2AFZ	0.09	Not sel.	1.75	Not sel.	1.71	Not sel.	1.07	Not sel.
Transcription	Gene expression	9.14	Not sel.	9.05	Not sel.	3.87	Not sel.	Not sign.	Not sign.
DNA methylation	Sperm hypometh	0.42	1.69	0.36	1.22	2.38	Not sel.	0.15	Not sel.
DNA methylation	CpG methylation	Not sign.	Not sign.	Not sign.	Not sign.	0.85	Not sel.	Not sign.	Not sign.
DNA methylation	5-hMc	11.29	Not sel.	15.65	Not sel.	7.94	Not sel.	0.39	Not sel.
DNA methylation	CHH methylation	0.23	Not sel.	0.18	Not sel.	Not sign.	Not sign.	Not sign.	Not sign.
DNA methylation	CHG methylation	0.75	Not sel.	1.17	Not sel.	1.41	Not sel.	Not sign.	Not sign.
Non-B DNA, microsats	G-quadruplex	4.76	Not sel.	6.28	Not sel.	7.43	Not sel.	Not sign.	Not sign.
Non-B DNA, microsats	A-phased repeats	8.22	0.14	10.34	Not sel.	7.36	Not sel.	Not sign.	Not sign.
Non-B DNA, microsats	Direct repeats	1.95	1.25	2.45	Not sel.	4.01	Not sel.	0.58	Not sel.

Non-B DNA, microsats	Inverted repeats	1.73	Not sel.	2.33	Not sel.	1.00	Not sel.	0.11	Not sel.
Non-B DNA, microsats	Mirror repeats	1.24	Not sel.	3.47	Not sel.	Not sign.	Not sign.	0.72	Not sel.
Non-B DNA, microsats	Z DNA motifs	2.08	0.92	3.12	Not sel.	1.29	Not sel.	Not sign.	Not sign.
Non-B DNA, microsats	Mononuc. Microsats	0.92	Not sel.	1.94	Not sel.	1.05	Not sel.	1.62	Not sel.
Non-B DNA, microsats	Di-, tri-, and tetranucl. microsats	0.51	1.16	0.73	0.74	Not sign.	Not sign.	Not sign.	Not sign.
Nucleotide composition	GC-content	12.90	14.23	18.63	15.84	11.98	3.32	0.51	Not sel.
L1 target motifs	L1 target motifs	0.92	20.27	0.64	20.12	4.22	Not sel.	Not sign.	Not sign.
Other TEs	Alu	5.31	0.91	8.73	1.74	9.07	11.19	0.53	Not sel.
Other TEs	MIR	11.04	0.78	14.03	Not sel.	0.46	Not sel.	0.86	Not sel.
Other TEs	L2 and L3	5.04	1.53	5.94	2.93	Not sign.	Not sign.	Not sign.	Not sign.
Other TEs	DNA transposons	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.
Other TEs	LTR elements	5.51	Not sel.	6.84	1.70	2.53	Not sel.	Not sign.	Not sign.
Replication	Replication origins	7.43	Not sel.	10.64	Not sel.	9.35	4.21	0.92	Not sel.
Recombination	Recomb. hotspots	0.90	0.93	2.24	Not sel.	0.27	Not sel.	0.53	Not sel.
Selection	Most conserved elements	16.99	1.15	19.97	1.20	2.58	Not sel.	0.21	Not sel.
Selection	CpG island	4.51	0.82	6.39	Not sel.	10.41	10.42	0.25	Not sel.
Selection	Exons	5.71	1.40	8.34	1.22	6.30	Not sel.	0.22	Not sel.
Selection	Introns	6.75	Not sel.	6.18	Not sel.	1.29	Not sel.	Not sign.	Not sign.
Chromosomal location	Distance to centromere	Not sign.	Not sign.	Not sign.	Not sign.	0.17	Not sel.	Not sign.	Not sign.
Chromosomal location	Distance to telomere	0.81	Not sel.	1.01	Not sel.	1.31	Not sel.	Not sign.	Not sign.
Chromosomal location	Telomere hexamer	13.14	Not sel.	16.49	Not sel.	0.33	Not sel.	Not sign.	Not sign.
Replication	Replication timing	4.89	Not sel.	8.42	Not sel.	7.56	Not sel.	0.50	Not sel.
Recombination	Recombination rate	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.	Not sign.
Total pseudo-R^2			47.64		57.61		18.66		1.81

Table S3. Construction of the comprehensive blacklist.

We considered problematic regions specific to H1-human embryonic stem cell line (H1-hESC) by adding blacklist the genomic regions that showed extreme signal in the H1-hESC ChIP-Seq control sample (ID: ENCSR000AMI) We then employed two approaches to identify regions with extreme signals.

Blacklist	**ENCODE blacklist** (Amemiya et al. 2019)	**MACS on ENCSR000AMI** (Control ChIP-seq on H1-hESC)[A]	**Customized script on ENCSR000AMI** (Control ChIP-seq on H1-hESC)[B]	**Comprehensive ENCODE blacklist**
Number of regions	292	2,094	519	861
Size of regions	10.2 Mb	0.116 Mb	4.6 Mb	11.8 Mb

[A]First, we called peaks in the control file using MACS2 with default parameters (Zhang et al. 2008; Feng et al. 2012). [B]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).

Table S4. Correlation and overlap between genome-wide epigenetic profiles in HEK-293 and hESC cell lines.

Feature	**HEK293T** (data from HEK-293 cell line were used when not available in HEK-293T)	**hESC**	**Correlation/ Overlap**
DHS (Read-depth normalized signal)	https://www.encodeproject.org/experiments/ENCSR000EJR/	https://www.encodeproject.org/experiments/ENCSR000EMU/	0.962
H3K27ac (Fold change over control)	https://www.encodeproject.org/experiments/ENCSR000FCH/	https://www.encodeproject.org/experiments/ENCSR000ANP/	0.475
H3K4me3 (Fold change over control)	https://www.encodeproject.org/experiments/ENCSR000DTU/	https://www.encodeproject.org/experiments/ENCSR814XPE/	0.804
CpG methylation (Methylated sites)	https://www.encodeproject.org/experiments/ENCSR087IEW/	(Lister et al. 2009)	Overlap=58.8%

Table S5. Comparison across recent polymorphic L1 Studies.

Study	Number of polymorphic L1s	Source	Estimated insertion rate	Figure in the publication's study showing the allele frequency spectrum
Dataset used in current study (Ewing and Kazazian 2011)	1,012	5 cross-referenced L1 polymorphism studies on 310 individuals from 13 populations (Wang et al. 2006; Beck et al. 2010; Ewing and Kazazian 2010; Huang et al. 2010; Iskow et al. 2010)	0.0037 - 0.0105 [1/95-1/270] (Ewing and Kazazian 2010; Ewing and Kazazian 2011)	Figure 1B
(Stewart et al. 2011)	792	1000 Genomes Project	0.0061 [0.0059-0.0062]	Figure 6A
(Yu et al. 2017)	2,398	1000 Genomes data from 90 Han Chinese	0.005 [1 in 200 births]	Figure 4C

Table S6. Comparison of *de novo* L1 dataset across three studies.

Cell line	HEK-293T	hESC	Hela
Total number of elements analyzed	17,037	3,582	1,565
Source	Current study	(Flasch et al. 2019)	(Sultana et al. 2019)

VITA

Di Chen

EDUCATION

The Pennsylvania State University 2014-Present
Ph.D., Intercollege Graduate Program in Genetics University Park, USA

The Pennsylvania State University 2017-2019
Master of Applied Statistics University Park, USA

Northwest A&F University 2009-2013
B.S. in Biotechnology Xi'an, China

SELECTED PUBLICATIONS

1. Guiblet, W., Cremona, M.A., Harris, R.S., **Chen, D.**, Eckert, K.A., Chiaromonte, F., Huang, YF., Makova, K.D. Non-B DNA: A major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome (In Review).

2. **Chen, D.***, Cremona, M.A.*, Qi, Z., Mitra, R.D., Chiaromonte, F., Makova, K.D. (2020). Human L1 Transposition Dynamics Unraveled with Functional Data Analysis. *Molecular Biology and Evolution* (Advance access) (*Equal Contribution).

2. Cechova, M., Vegesna, R., Tomaszkiewicz M., Harris, R.S., **Chen, D.**, Rangavittal, S., Medvedev, P., Makova, K.D. (2020). Dynamic evolution of great ape Y chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* (Advance access).

CONFERENCES AND PRESENTATIONS

2018 CSHL Transposable Elements Meeting 2018, Cold Spring Harbor, NY. Dynamic Landscape of Human L1 Transposition Revealed with Functional Data Analysis (Poster)

2018 Mobile Genetic Elements and Genome Plasticity Meeting, Santa Fe, NM. Dynamic Landscape of Human L1 Transposition Revealed with Functional Data Analysis (Poster)

2017 The Society for Molecular Biology and Evolution Annual Meeting, Austin, TX. Dynamic Landscape of Human L1 Transposition Revealed with Functional Data Analysis (Poster)

2017 Three Rivers Evolution Event Meeting, Pittsburgh, PA. Dynamic Landscape of Human L1 Transposition Revealed with Functional Data Analysis (Talk)

SELECTED AWARDS

2019 Outstanding Student Leadership Award from Huck Institutes of the Life Sciences
2018 The Huck Institutes Graduate Adviser Fellowship (1-year full tuition and stipend)
2017 Future of Science Fund Scholarship from Keystone Symposia
2014 Braddock Fellowship from Penn State University

TEACHING EXPERIENCE

Teaching Assistant, Biology: Basic Concepts & Biodiversity (3 Semesters)

Teaching Assistant, Biological Data Analysis: The Right Way (2 Summer Bootcamps)

Teaching Assistant, Molecular Evolution (1 Semester)