Nelson Chuang, M.D., M.H.S.

nchuang@som.umaryland.edu

Education

2014 – 2022	*University of Maryland Medical Center, Baltimore, MD* Ph.D. for Clinicians track, Conferral date Spring 2022 Ph.D. candidate in Molecular Medicine, Cancer Biology track
2016 – 2019	*University of Maryland Medical Center, Baltimore, MD* Gastroenterology Fellowship
2011 – 2014	*Montefiore Medical Center, Bronx, NY* Internal Medicine Residency
2007 – 2011	*Robert Wood Johnson Medical School, Piscataway, NJ* Degree: M.D.
2005 – 2006	*Johns Hopkins Bloomberg School of Public Health, Baltimore, MD* Degree: M.H.S. in the Department of Mental Health
2001 – 2005	*Johns Hopkins University, Baltimore, MD* Degree: B.S. in Biomedical Engineering concentration in Computer Science Minor: Computer Science

Honors

2022	Best IGS Trainee-Led -Omics Paper
2019	GI Division Teaching Award
2014	NIDDK T32 Fellowship
2013	Presidential Poster Award at American College of Gastroenterology Conference
2012	Montefiore Intensified Research Track

Publications and Presentations

Publications

- **Chuang NT,** Gardner EJ, Terry DM, Rivell GL, Crabtree J, Perry JA, Mahurkar AA, Hong CC, Devine SE. Mutagenesis of human genomes by endogenous mobile elements on a population scale. *Genome Research* 2021 Nov 12; 31(12):2225-35.
- Human Genome Structural Variation Consortium (including **Chuang NT**). Characterization of structural variation from de novo assembly of 64 haplotype-resolved human genomes of diverse ancestry. *Science* 2021 April 2; 372(6537):eabf7117.

- Shanta O, Noor A, Human Genome Structural Variation Group (including **Chuang NT**), Sebat J. The Effects of Common Structural Variants on 3D Chromatin Structure. *BMC Genomics* 2020 Jan 30;21(1):95.
- **Chuang NT**, Agarwal A, Kim R. Successful Hemostasis with Endoscopic Suturing of Recurrent Duodenal Ulcer Bleeding after Failed Initial Endoscopic Therapy and Embolization. *ACG Case Rep J*. 2020 Feb 20;7(2): e00275.
- Sulovari A, Li R, Audano PA, Porubsky D, Vollger MR, Logsdon GA; Human Genome Structural Variation Consortium (including **Chuang NT**), Warren WC, Pollen AA, Chaisson MJP, Eichler EE. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci U S A*. 2019 Nov 12;116(46):23243-23253.
- Human Genome Structural Variation Consortium (including **Chuang NT**). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun.* 2019 Apr 16;10(1):1784.
- Gardner E, Lam V, Harris D, **Chuang NT**, Mills R, 1000 Genomes Project Consortium, Devine SE. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Research* 2017 Nov;27(11):1916-1929.
- Scott E*, Gardner E*, Massood A, **Chuang NT**, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Research* 2016 Jun;26(6):745-55.
- **Chuang N**, Gross R, Odin J. Update on the Epidemiology of Primary Biliary Cirrhosis. *Expert Review of Gastroenterology and Hepatology*. 2011 Oct;5(5):583-90.
- **Chuang N**, Mori S, Yamamoto A, Jiang H, Ye X, Xu X, Richard LJ, Nathans J, Miller MI, Toga A, Sidman RL, Zhang J. An MRI-based Atlas and Database of the Developing Mouse Brain. *NeuroImage* 2011 Jan 1;54(1):80-9.

Abstracts

- **Chuang N**, Uradomo L, Kim R. Successful hemostasis with endoscopic suturing of persistent duodenal ulcer bleeding after failed endoscopic therapy and transarterial embolization. *Digestive Disease Weekly* 2018. Washington, D.C.
- **Chuang NT**, Gardner E, Scott E, Devine SE. Comprehensive long-range sequencing of full length long interspersed element-1 in diverse human populations. *Cold Spring Harbor Biology of Genomes* Meeting 2017. Cold Spring Harbor, NY.
- Gardner E, Lam V, Harris D, **Chuang NT**, 1000 Genomes Project Consortium, Mills, RE, Devine, SE. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Cold*

Spring Harbor Biology of Genomes Meeting 2016. Cold Spring Harbor, NY

- Scott E*, Gardner E*, Massood A, **Chuang NT**, Devine SE. An oncogenic L1 source element evades somatic repression and initiates human colorectal cancer. Abstract Accepted. *Cold Spring Harbor Biology of Genomes* Meeting 2016. Cold Spring Harbor, NY
- Gardner E, Lam V, Harris D, **Chuang NT**, 1000 Genomes Project Consortium, R. Mills, Devine SE. The Mobile Element Locator Tool (MELT). *Cold Spring Harbor Biology of Genomes* Meeting 2015. Cold Spring Harbor, NY
- **Chuang N**, Kim D, Michael H. Adult-onset Autoimmune Enteropathy treated with Infliximab and Azathioprine. *American College of Gastroenterology* 2013. San Diego, CA
- **Chuang N**, Peregina K, Stanley P, Augenlicht L. mTOR signaling in the intestinal mucosa is modulated by both diet and inhibition of Notch signaling. *Digestive Disease Weekly* 2013. Orlando, FL
- Lenza C, Ben-Menachem T, Javidian P, **Chuang N**, Michael H. Inflammatory fibroid polyps of the stomach and duodenum have characteristic endoscopic ultrasound findings: Review of cases and correlation with histology. *Digestive Disease Weekly* 2011. Chicago, IL
- **Chuang N**, Li G. Correlating per capita alcohol consumption and alcohol dependence prevalence. *American Public Health Association* 2007. Washington, D.C.
- Aggarwal M, Zhang J, **Chuang N**, Mori S. Three-Dimensional stereotactic atlas of developing mouse brains using diffusion tensor magnetic resonance imaging and micro-computed tomography. *Society of Neuroscience* 2007. San Diego, CA
- **Chuang N**, Zhang J, Li J, Xu B, Senft B, Sidman RL, Mori S. Three-Dimensional atlas of developing mouse brains based on diffusion tensor imaging. *International Society of Magnetic Resonance in Medicine* 2007. Berlin, Germany

Abstract

Title of Dissertation: Impact of mobile elements on human traits and diseases

Nelson Ta-Ching Chuang, M.D., Doctor of Philosophy, 2022

Dissertation Directed by: Scott E. Devine Ph.D., Associate Professor, Department of Medicine, Institute for Genome Sciences; Greenebaum Comprehensive Cancer Center, University of Maryland School of Medicine

Several large-scale Illumina whole genome sequencing (WGS) and whole exome sequencing (WES) projects have emerged recently that have provided exceptional opportunities to discover mobile element insertions (MEIs) and study the impact of these MEIs on human genomes. However, these projects also have presented major challenges with respect to the scalability and computational costs associated with performing MEI discovery on tens or even hundreds of thousands of samples. To meet these challenges, we have developed a more efficient and scalable version of our Mobile Element Locator Tool (MELT) called CloudMELT. We then used MELT and CloudMELT to perform MEI discovery in 57,919 human genomes and exomes, leading to the discovery of 104,350 non-redundant MEIs. We leveraged this collection: 1) to examine the population distributions and subfamilies of these MEIs, 2) to examine the mutagenesis of GENCODE genes, ENCODE-annotated features, and disease genes by these MEIs, and 3) to examine the potentially active L1 source elements that drive mobilization of new *Alu*, L1, and SVA MEIs in humans. Our study provides new insights on the L1 source elements that drive MEI mutagenesis and brings forth a better understanding of how this mutagenesis impacts human genomes.

Impact of mobile elements on human traits and diseases

by

Nelson T. Chuang

Dissertation submitted to the faculty of the Graduate School of the
University of Maryland, Baltimore in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Dedication

To my incredible wife Nicole for supporting me through this journey.

To my parents Che-nan and Wen-Ring for their love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
1kGP	1000 Genomes Project
AFCS	Amish Family Calcification Study
AFR	African
AP	Apurinic/Apyrimidinic
ASP	Antisense Promoter
AWS	Amazon Web Services
CCGD	Candidate Cancer Gene Database
cCRE	cis-Candidate Regulatory Elements
CDS	Coding exons
COSMIC	Catalogue of Somatic Mutations in Cancer
CVD	cardiovascular disease
dbGaP	Database of Genotypes and Phenotypes
dbRIP	Database of Retrotransposon Insertion Polymorphisms
DDD	Deciphering Developmental Disorders
DRP	Discordant read pairs
EC2	Elastic Compute Cloud
EN	Endonuclease
ENCODE	ENCyclopedia Of DNA Elements
ERV	Endogenous retrovirus
euL1db	European database of L1Hs retrotransposon insertions in humans

FDR	False discovery rate
FL	Full-length
FL-L1	Full-length L1
FL-L1Hs	Full-length L1 Homo sapiens
GAD	Genome Association Database
GB	Great Britain
GBD	Genome Browser Database
GENCODE	(Not an acronym) Encyclopedia of genes and gene variants
GRCh37	Genome Reference Consortium Human build 37
GRCh38	Genome Reference Consortium Human build 38
GTEx	Genotype-Tissue Expression project
GWAS	Genome wide association study
GWD	Gambian in Western Division, The Gambia - Mandinka
HERV	Human endogenous retrovirus
HERVK	Human endogenous retrovirus-K
HGSVC	Human Genome Structural Variation Consortium
HIV	Human immunodeficiency virus
HML-2	Human MMTV Like
HPT	Homopolymer tract
IDR	Intrinsic disordered region
IGV	Integrative Genome Viewer
INDEL	Insertion-Deletion
JHS	Jackson Heart Study

KAP1	KRAB-associated protein 1
LINE	Long interspersed element
LTR	Long terminal repeat
LWK	Luhya in Webuye, Kenya
MAF	Minor allele frequency
ME	Mobile element
MEI	Mobile element insertion
MELT	Mobile Element Locator Tool
MMTV	Mouse mammary tumor virus
NIST	National Institute of Science and Technology
OASIS	Omics Analysis Search and Information System
OMIM	Online Mendelian Inheritance in Man
ORF	Open reading frame
PABP	Poly(A)-binding proteins
PALMER	Pre-mAsking Long reads for Mobile Element inseRtion
PB	Petabytes
PCNA	Polymerase-delta-associated sliding DNA clamp
PIWI	PIWI-interacting RNA
REF	Reference
RISC	RNA-induced silencing complex
RNP	Ribonucleoprotein
RRM	RNA recognition motif
RT	reverse transcriptase

SINE	Short interspersed element
SpIRE	Spliced integrated retrotransposed element
STR	Short tandem repeats
SV	Structural variant
SVA	SINE variable-number tandem-repeat Alu
TOPMed	Trans-Omics in Precision Medicine
TPRT	Target primed reverse transcription
TSD	Target site duplication
UCSC	University of California Santa Clara
UK	United Kingdom
UKBB	United Kingdom Biobank
URI	Uniform resource identifiers
USA	United States of America
UTR	Untranslated region
vCPU	virtual Centralized Processing Units
VNTR	Variable number of tandem repeats
WES	Whole Exome Sequence
WGS	Whole Genome Sequence
YY1	Yin Yang 1

Chapter 1. Mobile elements and their impact on human diversity and disease

1.1 Introduction to Mobile Elements

Mobile elements (MEs), or transposable elements, refer to DNA sequences that have the ability to be mobilized and integrate into new places in an organism's genome. They are found in the human genome but also are found in almost every organism, including mammals, plants, insects, and microbes. MEs were originally described as "selfish DNA" by Francis Crick as they appeared to have no contribution to their host's fitness.[1] Their original discovery was made by Barbara McClintock in 1948 when she noticed that the kernel color patterns of maize were attributable to the presence of two genetic loci she called the *Dissociator* and *Activator* loci.[2] These "mutable loci" were able to change positions, or transpose, in the maize genome and affect the mosaic colors of maize.[3] It would take decades before the significance of McClintock's work would be recognized in 1983 when she received the Nobel Prize in Physiology or Medicine for her discovery of mobile elements or "jumping genes".

There are different types of mobile elements found in the human genome and they insert into different regions of the genome. The majority of these insertions are found in intergenic or intronic regions without functional impact on the cell.[4] However, insertions can occur in regions of the genome that disrupt gene function or alter gene expression, thereby leading to disease. The first examples of disease-associated mobile element insertions (MEIs) were discovered in 1988 by Kazazian and colleagues when they found MEIs that disrupted exon 14 of the *Factor VIII* gene in two patients with Hemophilia A.[5] Since the discovery of mobile elements and their potential for causing disease, there have been at least 130 reported cases of disease linked to a MEI.[6–9] This chapter will introduce these MEIs and discuss the pivotal role they have played in generating our genetic diversity, their mechanisms of altering gene regulation, and their complex interplay with the host defense system.

1.2 Autonomous and non-autonomous mobile elements

Mobile elements have been traditionally broken down into two classes based on their mechanisms of transposition. Class I transposons, also known as retrotransposons, are defined by mobile elements that mobilize through a “copy-and-paste” mechanism using reverse transcription of an RNA intermediate followed by insertion of the cDNA copy at a new site of the genome (Figure 1.2.1).[10] In contrast, Class II transposons, also known as DNA transposons, move around the genome using “cut-and-paste” or replicative mechanisms. They comprise 3% of the human genome, and unlike retrotransposons, there are no known active DNA transposons in humans (or mammals, with the exception of bats).[11–13]

Retrotransposons are further divided into those that have long terminal repeats (LTRs) and those that lack LTRs (non-LTR retrotransposons). Endogenous retroviruses (ERVs) are LTR retrotransposons that are found in humans and are unable to propagate themselves because no functional copies remain in human genomes, which will be discussed further below. In contrast, three types of non-LTR retrotransposons remain active in humans today including long and short interspersed elements (LINE-1 and SINE), and SINE variable-number tandem-repeat *Alu* (SVA) elements.

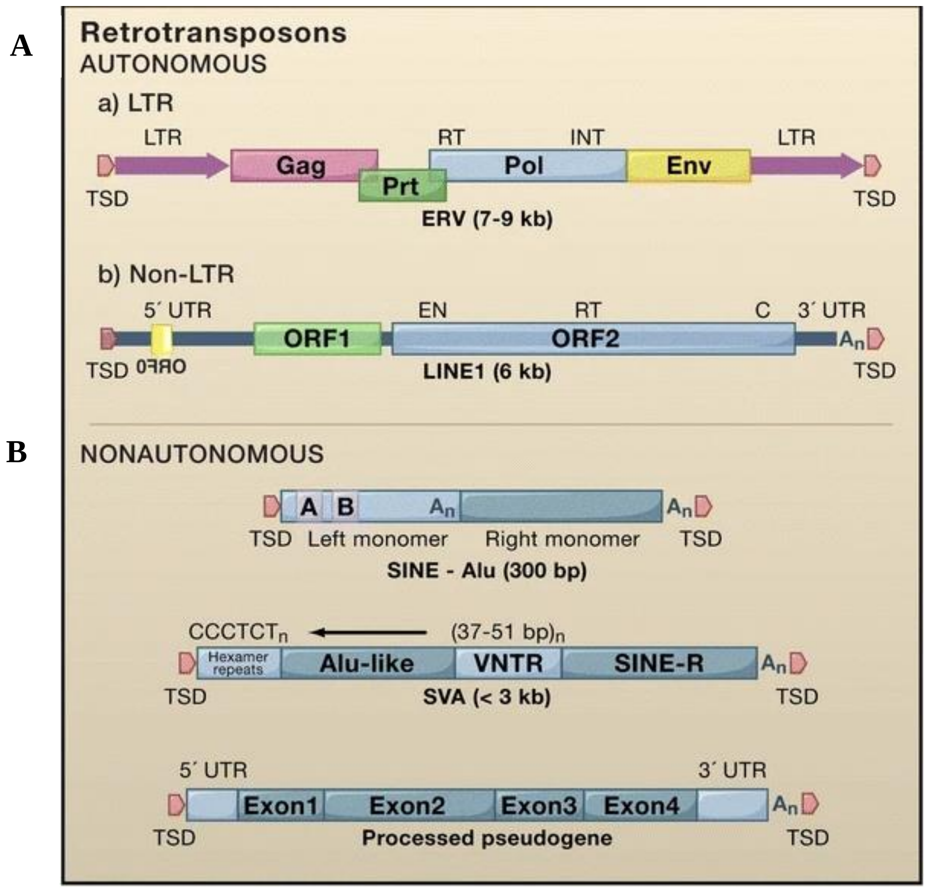


Figure 1.2.1 Different types of retrotransposons

A. Autonomous retrotransposons are able to facilitate their own “copy and paste”. They include two types: LTR and non-LTR autonomous retrotransposons. **B**. Nonautonomous retrotransposons rely on LINE1’s machinery to facilitate their mobilization in the genome. They include SINE, SVA, and processed pseudogenes. (Figure adapted from Goodier et al. 2008)[11]

1.2.1 HERV-K

Human ERVs, or HERVs, are comprised of thousands of provirus remnants that have been integrated into the human germline for the past 35 million years with more than 8% of the human genome derived from these elements.[14] Unlike exogenous viruses that undergo horizontal transmission from cell to cell using host machinery for replication, a retrovirus integrates its DNA into a host cell genome to achieve its replication. However, if a retrovirus integrates into the genome of a germline cell it can lead to vertical transmission from the parent to offspring resulting in the virus becoming "endogenized". When retroviruses are endogenized they can re-integrate their DNA elsewhere in the genome and increase their copy number further. There are about 31 HERV families, with the HERV-K family being the only group that has been active within the past few million years.[15–17] The majority of HERV elements are inactive due to mutations and deletions.[15]

HERV-K elements are ~9.5kb sequences classified by the lysine (K) tRNA that is required to prime cDNA synthesis from its RNA template.[18] Since HERV-K bears similarity to the mouse mammary tumor virus (MMTV), it is also known as HML-2 (Human MMTV Like) [19] As their classification suggests, they have identical LTR sequences of ~1kb in length on their 5' and 3' flanks. These LTR sequences are susceptible to non-homologous recombination and deletion of the open reading frames resulting in the majority of HERV-Ks to be "solo" LTRs.[20]

Although HERV-K elements are thought to have lost their ability to transpose, some HERV-K elements are known to retain the ability to code for viral proteins and particles.[4,21,22] These proteins are derived from the *gag-pol-env* open reading frames, which encode the retroviral antigen, reverse transcriptase, and the viral envelope protein, respectively. HERV-K viral particles have been detected in tumor-derived tissues as well as tissues associated with autoimmune and neurologic diseases.[23–27] The protein levels also have been found to be elevated in human immunodeficiency virus (HIV)-infected individuals with decreased protein levels in patients treated with anti-HIV therapy.[28] These findings, along with genome studies showing polymorphic HERV-K insertions, suggest HERV-K may still have potential for activity; however, there has been no direct evidence of *de novo* insertions and active infectivity.[16,22,29,30]

1.2.2 LINE-1

LINE-1, or L1, is the only active autonomous retrotransposon that possesses its own machinery necessary to mobilize itself in the genome (**Error! Reference source not found.**). L1 elements also have been part of the human genome for at least the past 150 million years of mammalian radiation.[31] Due to their property of self-propagation, there have been numerous L1 lineages throughout primate evolution.[32] For the past 40 million years, through a phenomenon of subfamily succession, these multiple lineages have been reduced to a single dominant lineage that is the modern L1 family, L1PA1 or L1Hs (L1 Homo sapiens).[32,33] With over 500,000 copies in the human genome, L1 occupies approximately 17% of human genomic DNA.[4]

L1 is a ~6 kb non-LTR retrotransposon that is composed of a 5'untranslated region (5' UTR), two open reading frames (ORFs), and a 3'UTR (Figure 1.2.1). "Full-length L1" refers to a complete L1 with all these features intact, and thus, potentially having retrotransposition activity. Since I will be focusing on L1Hs elements-and for the sake of clarity-I will refer to full-length L1s as FL-L1Hs going forward. Full-length L1s older than the modern L1PA1 have the same structure with only changes in their nucleotide sequences that define their subfamily. In chapter 3 I will discuss in detail the significance of a full-length L1 elements.

The first part of a FL-L1Hs element is the 5'UTR, which contains the internal RNA polymerase II promoter and transcription factor binding sites that affect the levels of L1 transcription, where the first 100 bp has been shown to be critical for transcription.[34] Between nucleotides 13-21 is the Yin Yang 1 (YY1) transcription factor binding site believed to be involved in initiation of L1 transcription.[35–37] Runt-related transcription factor 3 (*RUNX3*) binds to nucleotide 83-101 and increases L1 transcription.[38] Two transcription factor binding sites for *SOX* family proteins, *SRY A* at 472-477 and *SRY B* at 572-577, modulate levels of L1 transcripts.[39] Transcription factor binding of *p53* is shown to repress L1 activity.[40,41] There are other documented transcription factors that bind the L1 promoter that may also affect activity.[42]

The two ORFs encode an ORF1 protein (ORF1p or p40) and an ORF2 protein (ORF2p) that preferentially associate with and bind the L1 mRNA that produced them (referred to as "cis-preference") to form a ribonucleoprotein (RNP) complex.[43–46] The ORF1p is a ~40kDa protein that functions as a nucleic acid chaperone for L1 integration.[47] It has a coiled-coiled domain, RNA recognition motif (RRM), and a carboxyl-terminal domain. The coiled-coiled domain forms superhelices of α-helices created by seven units of seven amino acid repeats known as heptads; they are responsible for the formation of homotrimers.[45,48,49] The RRM and carboxyl-terminal domains of ORF1p work cooperatively to bind single stranded nucleic acids.[50,51] The N-terminal end of ORF1p also has an intrinsic disordered region (IDR) believed to be important for regulatory protein binding and contains phosphorylation sites important for retrotransposition.[52–54] ORF2p is a ~150kDa protein that contains an endonuclease (EN) domain and reverse transcriptase (RT) domain that are both important for L1 integration.[55,56] ORF2p also contains two other conserved domains, the Z domain and a cysteine-rich domain (Cys). The Z domain lies between the EN and RT domains and is a 100 amino-acid sequence that has an octapeptide sequence necessary for RT activity.[57] The Z domain also contains a PCNA (polymerase-delta-associated sliding DNA clamp) binding motif that contributes to L1 retrotransposition efficiency.[58] The Cys domain has high affinity for RNA and may assist in the observed cis-preference for its own transcript during the formation of the L1 RNP complex.[59]

The process of L1 integration occurs by a process called target-primed reverse transcription (TPRT)[60] The ORF2p endonuclease starts by creating a single-strand nick in the genomic DNA to generate a free 3' hydroxyl group.[55] The endonuclease has a domain homologous to the apurinic/apyrimidinic (AP) endonuclease family and targets a degenerate consensus 5'-TTTTT/AA-3' DNA motif.[55,61,62] The exposed 3' hydroxyl group functions as a primer for the ORF2p reverse transcriptase, which uses the L1 mRNA as a template for first-strand cDNA synthesis. The exact processes of second-strand cleavage and DNA synthesis remain somewhat unclear, although it is believed that the host DNA repair machinery may play a role.[63–67] The completion of this "copy-and-paste" mechanism results in an L1 copy at a new genomic site that is flanked by direct repeats, known as target-site duplications (TSD's).[60,68]

The LINE-1 mRNA also includes a poly(A) signal, AATAAA, at the end of the 3'UTR which directs the addition of a poly(A) tail to the transcript. This signal is typically weak and allows a significant portion of polymerase to continue transcribing the

DNA downstream (also known as readthrough) of the L1 until it comes across an adequate poly(A) signal.[69,70] When this additional DNA sequence is transcribed and inserted with the L1 at the new integration site it is known as a 3'transduction.[70] It is estimated through *in silico* experiments that 15% of FL-L1Hs elements have generated 3'transductions in the human genome.[71]

1.2.3 SINE

The most prevalent SINE elements in humans are *Alu* elements, which are non-autonomous retrotransposons that rely on L1 to mobilize their own sequences in the genome. *Alu* elements make up ~11% of the human genome.[72] *Alu* elements were first identified more than 40 years ago when renaturation techniques using the *Alu*I restriction enzyme reproducibly produced a ~300 nucleotide renatured sequence.[73] *Alu* is ancestrally derived from the 7SL RNA gene that codes for the signal recognition particle of the ribonucleoprotein complex.[74] Its origin dates back about 65 million years ago (Mya) during the primate radiation when its prolific amplification occurred after acquiring an internal deletion of the central S domain, a 3' poly-A tract, and the tandem duplication of folding units. [75,76] *Alu* is predominantly found in introns, 3'UTRs, and intergenic genomic regions. *Alu* is the most abundant mobile element in the human genome by copy number.[4]

The dimeric structure of *Alu* is made up of a left and right monomeric folding unit separated by a short A-rich region (Figure 1.2.1). The left monomer consists of an internal RNA polymerase III promoter on the 5' half that is split into an A-box and B-box responsible for its transcription initiation[77], whereas the right monomer contains a long A-rich region at the 3' end. The transcribed *Alu* RNAs form ribonucleoprotein particles with Srp9p/14p heterodimers, poly(A)-binding proteins (PABP), and other unknown binding proteins.[78,79] These folding units form the secondary structure necessary for binding the Srp9p/14p and for retrotransposition to occur.[80] Both monomers are flanked by identical repeat sequences; however, since *Alu* lacks its own termination signal, the polymerase continues downstream until it finds a termination signal consisting of four or more Ts.

After transcription and formation of the *Alu* RNA and Srp9p/14p heterodimer complex, the complex can dock on a host ribosome and recruit L1 ORF2p.[61,81,82] *Alu* does not have its own ORFs, and relies on ORF2p TPRT for the final integration step resulting in TSDs. *Alu* only requires ORF2p for mobilization, whereas L1 requires both ORF1p and ORF2p.[82] In addition, L1s with their ORF1 truncated can, in certain cases, still generate a functional ORF2p.[83] These reasons could explain *Alu* having greater success in amplification leading to twice the copy number of L1 and causing more disease.[4,6,84,85]

1.2.4 SVA

SINE-VNTR-*Alu* (SVA) is a ~2kb nonautonomous retrotransposon that is a composite of three separate elements and makes up 0.6% of the human genome with at least 2,700 copies in the reference genome.[72,86] Their origins date back 18-25 Mya, making them the youngest family of retrotransposons that are active in human and primate genomes.[86] SINE-R (SINE-Retrovirus) is a 630bp nonviral retrotransposon discovered while examining the extinct HERV-K10 in 1987.[87] A few years later, a variable-number-of-tandem-repeats (VNTR) locus was found associated with a SINE-R in human complement C2 gene.[88] In 1994, a SINE-R with multiple copies of VNTRs was discovered associated with an *Alu*-like element in the complement C2 gene. SVA also has a hexameric repeat of CCCTCT located at the 5' end of the antisense *Alu* element. A full length active SVA that has undergone L1-mediated retrotransposition also has a poly(A) tail and flanking TSDs.

The mechanism of SVA transcription prior to *trans* L1 retrotransposition is still under study. The poly(A) tail that follows a canonical poly(A) signal suggests transcription is performed by RNA polymerase II. *In vitro* experiments demonstrate that the CT hexameric repeat and *Alu*-like domain have some promoter activity.[89] Initiation of transcription appears to occur from various positions on the 5' end of SVA and can even occur upstream of the SVA, which can produce a 5' transduction.[90] The transcribed SVA RNA associates with a ribosome or L1 RNP presumably by its *Alu*-like domain.[91,92] Integration of the SVA by ORF2p is believed to occur in the same fashion as in L1.[92] This would also mean SVA is susceptible to 3' transductions, 5' truncations, and 5' inversions that are hallmarks of L1-mediated retrotransposition.[92] The SINE-R domain also can contain a premature termination signal leading to 3' truncation.[90] Overall, the majority of SVAs in the human genome are full length with high sequence similarity of >89%.[92]

1.2.5 Processed pseudogenes and splicesomal RNAs

L1 can reverse transcribe and integrate other RNA transcripts found in the cell. Splicesomal RNAs have been discovered attached to the start of a truncated L1 transcript forming a chimeric retrotranscript.[93] The most commonly seen splicesomal RNA is the U6 small nuclear RNA (snRNA) forming a U6-L1 chimera with direct repeats flanking both ends suggesting insertion via TPRT. It is hypothesized to occur when the L1 RNP complex binds the splicesomal RNA during reverse transcription of the L1 mRNA and switches templates midway.[93,94] The chimeras appear to form independently through time without common ancestry as the fused L1s range from the young L1 subfamily, L1Hs, to ancient L1MA4.[93,95]

The L1 machinery also can be diverted to reverse transcribe and insert cDNAs from processed mRNAs. These inserted cDNAs are referred to as processed pseudogenes (they belong to a broader group known as retrocopies or retroduplications).[46,96,97] These processed pseudogenes lack an internal promoter to initiate transcription and typically possess mutations that render them unable to generate a functional polypeptide.[97] However, in mammals at least 10% of processed pseudogenes insert close enough to an RNA polymerase II promoter to undergo transcription producing a noncoding RNA.[98] A cancer environment also can provide the necessary factors to activate these pseudogenes.[99] Like all insertions produced by the L1 machinery, processed pseudogenes have TSDs, but also can have 5' truncations and mutations that prevent protein expression. There have been ~8,000 processed pseudogenes originating from ~2,500 genes identified in the draft human reference genome with high sequence similarity for coding DNA and amino acids.[100] A recent effort examining the 1000 Genomes Project Phase 3 cohort identified 15,642 processed pseudogenes from 2,533 whole exome sequences with six novel processed pseudogenes per individual.[101] In addition, with the increased use of long-read sequencing technology (discussed further below), there will be more processed pseudogenes discovered.[102]

1.3 Mobile elements shape the human genome

Amplification of mobile elements over millions of years has shaped the structure and evolution of hominid primate genomes. MEs occupy a significant portion of the human genome, so much so that the most abundant ME, *Alu*, is found on average once every 3kb of the human genome.[4] In fact, at least 45% of the human genome is derived from MEs and additional portions of the genome that originally harbored MEs appear to have become unrecognizable after millions of years of evolution.[72,103] Since they exist in such high copy numbers, MEs are categorized as repetitive elements and have actually increased the human genome size by ~750Mb–in just the past 6 million years the human genome has grown by 8Mb.[4,104] In order for the human genome to accumulate such a large addition of DNA over time, the MEs that are retained would have to be innocuous to their host or render a positive impact on host evolutionary fitness.

On the other hand, MEs historically have been referred to as “junk” DNA as they appear functionally inert and otherwise would have been eliminated by natural selection if found in a deleterious site. Thus, most insertions in our genome likely reflect evolutionary dynamics and a small fraction are “selection free” *de novo* insertions. Aside from the loci rich in adenosine-thymine dinucleotides, where the consensus integration site is more likely found, the insertion preference may be influenced by factors such as the nuclear topology, chromatin state, nuclear tethering proteins, or histone marks.[105] However, large-scale *in vitro* L1 experiments show retrotransposition site selection appears to lack preference for specific regions of the genome.[62,106] The near random nature of insertion site preference creates various MEI-host dynamics that incur both positive and negative effects on the host.

1.3.1 Increasing genetic diversity and altering gene expression

MEIs, by their nature, increase genetic diversity when they insert somewhere new in the genome and insertions can occur at different rates in different populations depending on various factors, which aren't entirely clear (i.e., different MEI activity and mutation rates).[72,82,107–110] MEIs also can generate genetic diversity through additional mechanisms that affect gene expression (Figure 1.3.1). *Alu* elements contain regions rich in adenosine between its monomers and in its poly(A) tail. These adenosine-rich tracts are prone to contraction and expansion creating new microsatellite repeats.[111] Microsatellite repeats, or short tandem repeats (STRs), are implicated in various roles of controlling gene expression.[112] More recent data using newer sequencing technology (i.e., PacBio long-read and 10X Genomics link-read sequencing) show that STRs and variable number tandem repeats (VNTRs or minisatellites) derived from retrotransposition are associated with differential gene expression and transcript splicing.[113]

Internal features of MEIs can contribute to altering and diversifying the transcriptome as well. *Alu* can create splice variants through its own cryptic splice sites allowing its integration into mature mRNAs, or "exonization", altering expression or protein function.[114–116] In human embryonic stem cells, transcripts containing inverted *Alu* sequences in the 3'UTR can lead to

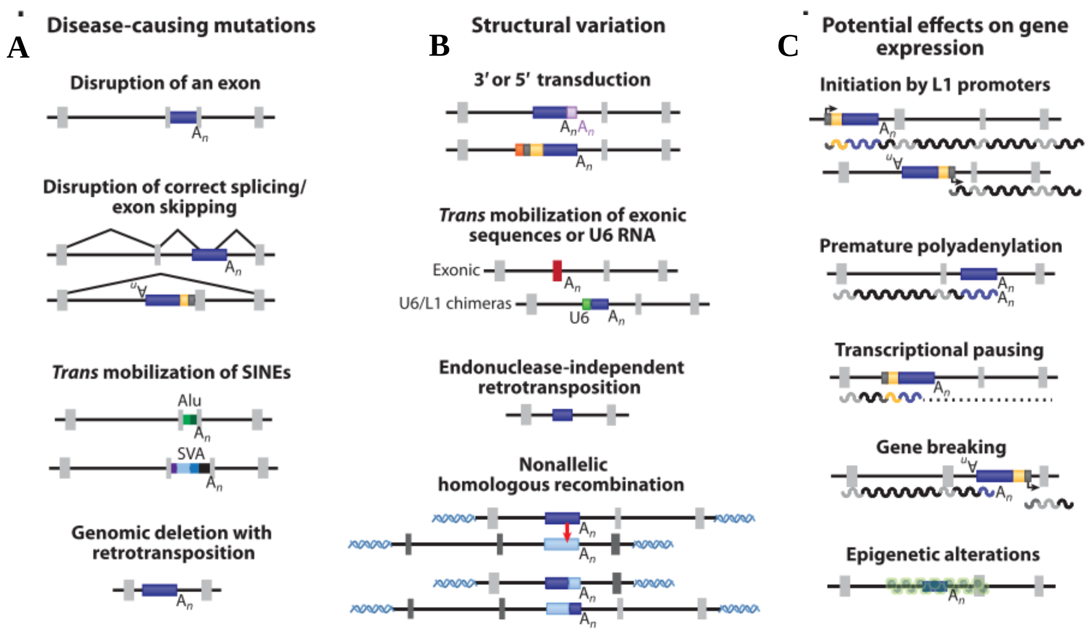


Figure 1.3.1 Mechanisms of altering gene expression and contributing to genetic diversity

Examples of mechanisms of MEI interaction to alter gene expression. **A.** Mobile element insertions can insert in genes leading to mutagenesis and disease. **B.** Mobile elements can directly or indirectly create structural variation. **C.** Mobile elements can alter regulation of gene expression. (Figure adapted from Beck et al.)[117]

preferential nuclear retention as a mechanism of gene regulation.[118] L1 insertions in introns can alter gene expression by causing truncated transcripts through premature polyadenylation or by compromising RNA polymerase elongation by slowing, pausing, or even dissociating the polymerase.[119,120] Further adding to transcript diversity, the L1 5’UTR contains an antisense promoter (ASP) that can drive transcription of adjacent genes in the opposite direction.[121–124] The L1 can also be positioned to partition the gene into two smaller transcripts that can give rise to novel proteins.[125] These truncated transcripts can be stable enough to contribute to additional mRNA isoforms diversifying the cell transcriptome. This generation of genetic diversity in the genome and transcriptome is ultimately advantageous to the host as a means of providing rapid evolutionary adaptation.[126]

However, not all insertions are advantageous to the host, and insertions that lead to the disruption of normal gene expression can, in the worst case, lead to disease. The most straightforward example is insertional mutagenesis caused by an MEI that is in or near a gene, such as the two independent L1 insertions in exon 14 of the *Factor VIII* gene described above.[5] MEIs can cause premature termination of a gene's transcript when inserted in an intron of a gene, especially L1 as it has multiple poly(A) signals upstream of the 3'UTR.[120] Cryptic splice sites, which allow *Alu* to exonize mRNAs, can also cause exon skipping or aberrant splicing that compromise protein structure and function.[127,128] *Alu* exonization also can cause frameshift mutations and protein truncations that lead to disease.[129–131] Furthermore, large deletions of the genome can occur when L1 and *Alu* sequences are substrates for nonallelic homologous recombination, which has been implicated in multiple cases of familial or hereditary syndromes.[132–136] Target site deletions of adjacent genomic sequences can occur during the retrotransposition process.[137,138] The L1 endonuclease can induce DNA double-strand breaks leading to genomic instability and, subsequently, apoptosis.[139–143] Despite all these potentially deleterious events that MEs can cause, the host maintains defense mechanisms to regulate MEIs in the genome.

1.3.2 Host repression and evasion

ME activity in the germline has facilitated the persistent existence and coevolution of MEs with the host for millions of years.[144] Retrotransposition of MEs in the germline is more likely to occur during embryogenesis or germ cell development due to genome-wide epigenetic de-repression.[145–147] It is estimated there is one new *Alu* insertion for every 20 human births, one new L1 insertion in about 20-200 births, and one new SVA insertion in every 900 births.[148–150] These germline insertions occur at varying allele frequencies (polymorphisms) in human populations.[55,151–154] Although quite rare, these heritable insertions can evade host defense and cause disease.[5–7,145,155,156] There are a multitude of host factors employed to target ME transcription, post-transcriptional processing, and integration in order to keep germline retrotransposition in check. As retrotransposition relies on an RNA intermediate, the host defense mechanisms are similar to the immune response to retroviruses (reviewed in Goodier et al. 2016).[157]

The germline dynamics of regulating MEs are complex and differ between the sexes due to strong sexual dimorphism. The majority of our understanding in this area is derived from animal models. Unchecked retrotransposition in the spermatogonia and spermatocytes leads to meiotic failure and eventual azoospermia in mutant mouse models.[158] At the heart of the gonadal genome defense are various RNA interference (RNAi) molecules, such as small interfering RNA (siRNA), microRNA (miRNA), and most importantly, PIWI-interacting RNA (piRNA). These molecules are formed by cleavage of long double stranded RNA (dsRNA) by the Dicer enzyme into short fragments of siRNAs or miRNA if the dsRNA is produced from endogenous RNA. The piRNAs are formed independently of the Dicer enzyme and specifically target transposons in germ cells.[159] These RNAi molecules are then incorporated into an RNA-induced silencing complex (RISC) that targets the complementary nucleic acid sequence for post-transcriptional gene silencing. For example, the mir-128 miRNA can bind to L1

RNA and thereby prevent L1 integration in stem cells.[160] In female reproductive germ cells, mutations in PIWI proteins do not affect fertility, and it appears that mature oocytes are able to tolerate hypomethylation and transcriptional expression of MEs.[161,162] There are far more restriction factors described in the literature than can be summarized here; for example, there are 22 RNAi pathway proteins that are associated with increased retrotransposon activity when they are disrupted in mutant mouse models.[157] Although the piRNA pathway is critical for germline defense against insertions, these RNAi molecules also are required to maintain aberrant reactivation in the soma.[163,164]

Somatic retrotransposition also can occur in terminally differentiated tissues of the human body. L1 has been shown to be active in neural progenitor cells in the brain as well as during embryogenesis, leading to somatic mosaicism.[165–167] It is believed that neurons are permissive of L1 activity and genetic variation introduced by L1 contributes to neuronal plasticity.[150,165,168–171] Somatic retrotransposition has mostly been studied in the context of disease pathogenesis (See Hancks et al. 2016 for review of both germline and somatic insertions associated with disease).[6] The rate of somatic retrotransposition varies greatly depending on tissue type, with epithelial tumors demonstrating the highest rates. These solid organ tumors have significant increases in somatic ME retrotranspositions linked to hypomethylation and reactivation of L1.[7,172,173] The deregulated tumor microenvironment gives rise to different rates of retrotransposition with esophageal adenocarcinoma, head-and-neck squamous carcinoma, lung squamous carcinoma, and colorectal adenocarcinoma being the most enriched.[152,173] The rate of retrotransposition for a given type of tissue can also vary from individual to individual.[174,175]

The main defense mechanism the host has evolved to suppress retrotransposition in the soma is believed to be CpG methylation.[176] SVA and *Alu* sequences are enriched in CpGs with almost a third of all CpGs in the human genome found in *Alu* sequences.[86,177,178] L1 sequences carry about 12% of the CpGs in the human genome and L1 CpGs in the 5' UTR are highly methylated.[179,180] In addition, with time, CpGs are susceptible to undergo spontaneous deamination of 5-methylcytosine to thymine; thus, the older fixed subfamilies are more likely to have CpG transitions.[181] Mutation of the YY1 transcription factor binding site can lead to hypomethylation of the 5'UTR in embryonic stem cells as well as liver and brain tissues.[182] Disruption of the ability to methylate these cytosines, such as in mutations in the family of DNA methyltransferases, can lead to hypomethylation and increase of ME transcription.[158,183–187]

Although rare, somatic retrotransposition events can occur when the various host defense mechanisms fail. Cell stress induced by genotoxic poisons, radiation, heat shock, viral infection, and heavy metals are all able to induce mobilization of SINE and L1s in mouse and cell models.[188,189] Epigenetic changes associated with aging and senescence also may be contributing factors that release MEs from repression, as *in vitro* models of aging and senescence showed increased copy numbers of non-LTR retrotransposons.[190–192] As mentioned above, siRNAs also can silence somatic L1 expression by increasing DNA methylation of the L1 5'UTR promoter.[193] As in germ cells, siRNAs and miRNAs can associate with their respective RISC to bind L1 RNA for cleavage in somatic tissue.[194,195] This likely functions as a second line of defense if L1 promoter is hypomethylated and L1 is transcribed.

In addition to DNA methylation, chromatin remodeling provides another redundant layer of epigenetic regulation. Retrotransposons are enriched for histone H3 lysine9 (H3K9) methylation consistent with a repressed chromatin state, and disruption of histone modification proteins and histone methylation can lead to reactivation of MEs.[196–199] For example, loss of histone methylation through retinoblastoma protein knockout causes reactivation of L1 in mouse fibroblasts.[200] Inhibition of histone methyltransferase SUV39H increased RNA polymerase III recruitment and transcription of *Alu* elements.[201] KRAB-associated protein1 (KAP1/TRIM28) is a transcriptional repressor of endogenous retroviruses in embryonic stem cells, and has been observed to bind older L1 subfamilies but not human-specific L1Hs elements.[202,203]

1.4 Summary

MEIs are an important part of our genome and active MEs continue to contribute to germline diversity or insertional mutagenesis. In the following chapters, I will describe the significance of studying mobile element insertions on a population scale, highlight the extent of their impact on genetic features associated with disease, and demonstrate insights on the newest subfamilies of L1. Prior to our effort, there was a lack of large-scale projects for mobile element discovery. Mobile elements have a major impact on shaping human genomes, and there is still much we do not know regarding their distribution in human populations. We will show that by studying MEIs on a large-scale in multiple diverse populations, we created the largest MEI resource that reveals patterns in their population distributions and shifts in subfamily proportions. We achieved this through the development of a new cloud-based tool to facilitate MEI discovery on an unprecedented scale. MEIs have been shown to be implicated in various diseases, and by merging our resource with other existing MEI collections, we demonstrate the potential impact of MEIs on human genome mutagenesis. Lastly, many MEs as described above rely on L1 for their mobilization in the genome. By developing a new sequencing pipeline, we were able to study L1Hs source elements on a deeper level and identify new families of active L1Hs elements. I conclude the dissertation with a summary of the results followed by a discussion of future implications and new research questions.

Chapter 2. Large-scale discovery of mobile elements

2.1 Abstract

Mobile elements have been present in humans for millions of years and their expansion has created significant genetic diversity. Due to limitations in sequencing technology, characterizing these elements has been difficult and somewhat limited to highly sequenced genomes, such as the reference genome. In this chapter, we describe methods for identifying MEIs on an unprecedented scale using a combination of cloud computing and software our lab developed previously called the Mobile Element Locator Tool (MELT). With a massive collection of new MEIs we are able to examine population differences, ME subfamily distributions, and demonstrate the extent that MEIs occupy genic regions in the human genome. Using curated ontology databases, we highlight the various diseases and disease systems associated with these MEIs.

2.2 Introduction

Despite the significance of mobile elements in shaping our genome, influencing gene regulation, and causing disease, large-scale studies to further study them have been lacking. The massive effort to sequence and construct the human reference genome has provided an excellent starting point to study the most common MEIs. Tools and databases such as RepeatMasker and Repbase have cataloged these MEIs for the reference genome.[204,205] It is difficult to capture all the missing information, or gaps, when sequencing a genome, and so the reference genome provides the most detailed example of the genetic sequence of a human genome. Although the current reference genome (build GRCh38, or hg38) should have 99% sequence similarity with other human genomes, it is just one example derived from the genetic data of primarily 11 individuals- of which, one person accounts for 70% of the total build.[206] Thus, the reference genome alone cannot account for the large diversity of variation that is found in “nonreference” human genomes.

The first large-scale MEI discovery effort to describe nonreference MEIs was performed by the 1000 Genomes Project (1kGP) international consortium, which identified 16,631 (10,595 novel) nonreference MEIs in 2,504 individuals across 26 populations from around the world (Table 2.2.1

).[153] An additional 6,089 MEI calls were discovered with an updated detection algorithm for rare MEIs.[207] These MEIs were not found in the reference genome, and thus, are considered polymorphic insertions. These collections could contain insertions that are found in only a few individuals and might cause disease. The 1kGP cohort is derived from individuals deemed in good health at the time of study inclusion but given that 26 MEIs were found in exon coding regions, the possibility of an associated disease or phenotype cannot be excluded. The genomes were sequenced at low coverage, where the average number of nucleotides that "cover" known reference bases were low, meaning there is reduced information to capture all the nonreference MEIs.[207] Larger studies with higher coverage and phenotype data are needed to further understand the true disease risk MEIs have on human health.

Since the 1kGP study in 2015, there have been few other large studies involving specific disease populations. Gardner et al discovered 1,129 MEIs in the exomes of 9,739 patients with developmental disorders as well as their parents, or 28,132 individuals, who are part of the Deciphering Developmental Disorders (DDD) cohort.[208] Compared to the whole genome sequence (WGS), the whole exome sequence (WES) is much smaller and

Population name	Population Abbreviation	Superpopulation	Population description
British	GBR	EUR	British in England and Scotland
Finnish	FIN	EUR	Finnish in Finland
Southern Han Chinese	CHS	EAS	Han Chinese South
Puerto Rican	PUR	AMR	Puerto Rican in Puerto Rico
Dai Chinese	CDX	EAS	Chinese Dai in Xishuangbanna, China
Colombian	CLM	AMR	Colombian in Medellin, Colombia
Iberian	IBS	EUR	Iberian populations in Spain
Peruvian	PEL	AMR	Peruvian in Lima, Peru
Punjabi	PJL	SAS	Punjabi in Lahore, Pakistan
Kinh Vietnamese	KHV	EAS	Kinh in Ho Chi Minh City, Vietnam
African Caribbean	ACB	AFR	African Caribbean in Barbados
Gambian Mandinka	GWD	AFR	Gambian in Western Division, The Gambia - Mandinka
Esan	ESN	AFR	Esan in Nigeria
Bengali	BEB	SAS	Bengali in Bangladesh
Mende	MSL	AFR	Mende in Sierra Leone
Tamil	STU	SAS	Sri Lankan Tamil in the UK
Telugu	ITU	SAS	Indian Telugu in the UK
CEPH	CEU	EUR	Utah residents (CEPH) with Northern and Western European ancestry
Yoruba	YRI	AFR	Yoruba in Ibadan, Nigeria
Han Chinese	CHB	EAS	Han Chinese in Beijing, China
Japanese	JPT	EAS	Japanese in Tokyo, Japan
Luhya	LWK	AFR	Luhya in Webuye, Kenya
African Ancestry SW	ASW	AFR	African Ancestry in Southwest US
Mexican Ancestry	MXL	AMR	Mexican Ancestry in Los Angeles, California
Toscani	TSI	EUR	Toscani in Italy
Gujarati	GIH	SAS	Gujarati Indians in Houston, TX

Table 2.2.1 Populations included in the 1000 Genomes Project.

Listed are the abbreviations and descriptions of the 26 global populations for the 2,504 individuals recruited for the study. Superpopulation abbreviations: EUR (European ancestry), EAS (Eastern Asian ancestry), AMR (American ancestry), AFR (African ancestry), SAS (South Asian ancestry).[153]

consists of exons and small portions of flanking introns. Four MEIs in this study were shown to be the likely causative disruptions leading to the patients' developmental disease. Torene et al. discovered 23,014 MEIs from 89,874 targeted clinical exomes (Agilent Clinical Research Exome or IDT xGen Exome v1.0 baits) from patients referred for genetic testing.[209] Three patients had *Alu* insertions in exons suggestive for their diseases. Collins et al discovered 77,582 MEIs in 14,237 high coverage WGS from the Genome Aggregation Database (gnomAD), which consists of a combination of healthy and disease cohorts.[210] These studies demonstrate that there are likely more MEIs to be discovered in broader populations.

There are many limitations to performing these sequencing projects, mainly the cost to gather the necessary samples, to sequence them to an adequate depth, and the computational costs for data storage and processing (to name a few). After acquiring the processed sequencing data, software is needed to identify, or "call", the MEIs in a desired genome. MELT was developed to address this need in the 1kGP study and also was used in two of the aforementioned studies.[208,210] MELT was demonstrated to be the fastest and most efficient MEI caller with high sensitivity and specificity when compared to four other callers.[207] MELT allowed our group to perform MEI discovery on the 1kGP cohort of 2,504 low coverage genomes in about 2.5 weeks on a local compute cluster.[153] This was achieved by launching many instances of MELT in parallel as "jobs" in order to dramatically reduce the time to run MELT 2,504 times sequentially. Despite the speed and efficiency of MELT, it cannot overcome the inefficiencies and limited resources associated with a local compute cluster shared with many other users.

Although larger studies are needed to study MEIs on a population level using high coverage (~30X) genomes, performing MELT on local compute clusters with such genomes can take many months. There are already many large studies looking at high coverage genomic sequencing data and this number has been doubling about every 18 months. Currently the largest project is the Trans-Omics in Precision Medicine (TOPMed) with a goal of sequencing ~155,000 from over 80 different studies designed to target disease treatment by examining an individual's -omic (e.g., genomic, RNA, protein, epigenomic, metabolomic, etc.) data. The amount of storage required to house TOPMed sequencing data will be over 2.3 petabytes (PB, or a million gigabytes, GB).[211] Potentially larger ongoing observational studies, such as the Precision Medicine Initiative and Million Veterans Project, will contribute even more available data for analysis.[212,213]

This highlights the need for a larger scale implementation of MELT that can address the growing size of genomic data available. If the main bottleneck for most research institutions is their local compute cluster sharing resources, then a dedicated cluster is needed. Renting a dedicated compute cluster is becoming easier with the accessibility of "cloud" computing. The formal definition of cloud computing according to the National Institute of Science and Technology (NIST) is "ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".[214] There are various cloud computing services that provide as much compute processing units as one requires, such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud.

2.3 Results

2.3.1 MEI discovery in TOPMed disease cohorts

As part of the TOPMed consortium we contributed to the initial MEI discovery of 342 individuals spanning multiple disease cohort studies. We sought to expand discovery to two large TOPMed studies focused on cardiac disease with one of European ancestry and the other of African ancestry. The Amish Family Calcification Study (AFCS) recruits members of the Old Order Amish in Lancaster County, Pennsylvania, and records information pertinent to cardiovascular disease (CVD) through demographic information, serum markers, and noninvasive imaging for vascular calcification.[215] The Jackson Heart Study recruits African-Americans from the Jackson, Mississippi metropolitan area and collects similar CVD history and noninvasive measurements as the AFCS.[216]

A. Amish Study (n=1,112)

MEI	Total	5'UTR	3'UTR	Exon	Intron
L1	1,195	0	5	0	444
Alu	9,422	8	62	3	4,186
SVA	389	0	4	1	186
HERVK	51	0	0	0	17
Total	11,057	8	71	4	4,833

B. Jackson Heart Study (n=3,331)

MEI	Total	5'UTR	3'UTR	Exon	Intron
L1	6,760	10	37	4	2,873
Alu	40,323	71	349	44	17,898
SVA	2,920	6	37	5	1,464
HERVK	195	0	0	0	47
Total	50,198	87	846	53	22,282

Table 2.3.1 TOPMed MEI discovery

MELT MEI discovery of the TOPMed (**A**) Amish Family Calcification Study and (**B**) Jackson Heart Study. The Total column is the sum of 5'UTR, 3'UTR, Exon, Intron, and intergenic insertions (intergenic not shown).

We performed MELT discovery on 1,112 Amish and 3,331 African-American Jackson Heart Study (JHS) genomes (Table 2.3.1). Although the JHS study sample size is approximately 3X larger than the AFCS study, there are 5X more total MEIs discovered.

The Amish population of eastern Pennsylvania represents a population with a gene pool derived from only about 250 founding families who emigrated from Switzerland in the mid-1700s.[217] This "Founder Effect" creates a limited gene pool reflected in their lower MEI counts. In contrast, the JHS African-Americans possess a much more diverse genetic pool, which originates from their African ancestry coupled with generations of admixture with North American ethnic groups.[218,219]

2.3.2 Large scale WES MEI discovery

The United Kingdom Biobank (UKBB) study is a prospective cohort study following approximately 500,000 40- to 69-year-old participants living in the UK recruited during 2006-2010. The UKBB study leveraged the medical records of these individuals and regularly collected blood, urine, and saliva samples for testing. They also had questionnaires to assess lifestyle factors that may impact the participants health-related records. We selected a cohort of 49,635 individuals who identified as British, Irish, or "Other White" and downloaded their WES data for MELT MEI discovery.

Compared to the 54,847 MEIs from 4,443 WGS of the TOPMed Amish and JHS, the ~11x greater sample size of the UKBB cohort generated 4,027 total MEIs (Table 2.3.2). This is to be expected with targeted sequencing of exome, which allowed for discovery of 386 exonic insertions compared to 55 from TOPMed. WES sequencing requires capture of flanking intron sequence around each exon, which is enriched for MEIs and accounts for 75% (3,027/4,027) of the discovered MEIs. Studying insertional mutagenesis in exons allows for straightforward interpretation of functional impact leading to Mendelian disease or trait.[220,221] 89% (3,588/4,027) of the discovered insertions are rare (minor allele frequency, or MAF<0.01), 78% (3,174/4,027) are very rare (MAF<0.001), and 29% (1,182/4,027) found in only a single individual (MAF=0.00001).

UKBB (n=49,635)					
MEI	**Total**	**5'UTR**	**3'UTR**	**Exon**	**Intron**
L1	1,480	55	157	128	1,101
Alu	2,374	81	211	239	1,793
SVA	170	5	9	19	130
HERVK	3	0	0	0	3
Total	4,027	141	377	386	3,027

Table 2.3.2 UKBB MEI discovery

MELT MEI discovery of the UKBB WES data. The Total column is the sum of 5'UTR, 3'UTR, Exon, Intron, and intergenic insertions (intergenic not shown).

2.3.3 CloudMELT and cloud-based MEI discovery

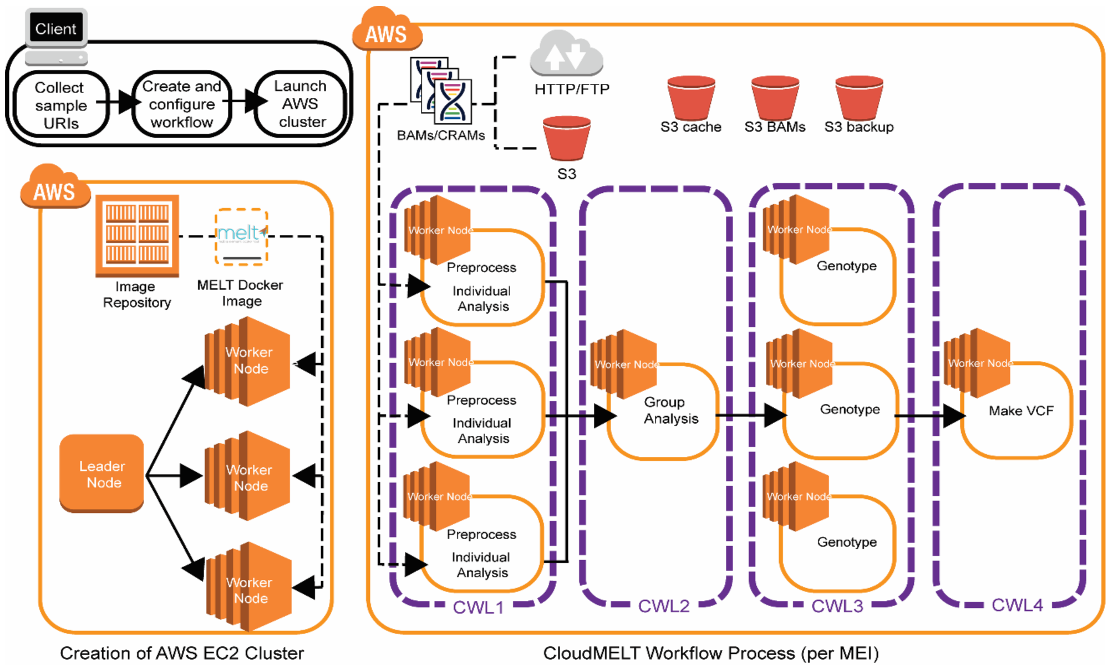


Figure 2.3.1 CloudMELT v1.0 workflow

The user configures the desired size of the cluster and supplies the universal resource identifier (URI or s3/http/ftp link) in Common Workflow Language (CWL) format for CloudMELT to create the AWS EC2 cluster. When the cluster is created, CloudMELT automates retrieval of genome data (BAMs or CRAMs) and coordinates all the “worker” nodes to process MEI discovery in parallel. Each of MELT’s discovery steps (Preprocess, Individual Analysis, Group Analysis, Genotype, Make VCF) are encapsulated in CWL steps 1-4. AWS S3 storage is used to store downloaded data as well as save working files for debugging.

CloudMELT is a cloud-specific workflow developed to address the limits of running massively parallel MEI discovery on local compute clusters. We discovered that increasing the amount of genomic data (from low coverage to high coverage genomes) while running the same number of jobs, drastically increased processing time due to limited resources of a local cluster. Running MELT on 4,443 TOPMed genomes took approximately eight months on our local cluster. Prior to running MELT, the sequencing data had to be downloaded first, which took approximately three months on our cluster.

Transferring terabytes of sequencing data using high speed protocols requires not only a reliable connection with available bandwidth, but also free compute units to coordinate the transfers. Therefore, we knew that developing a cloud port of MELT would allow fast and efficient MEI discovery of thousands of genomes without the limitations of sharing a local compute cluster.

We developed CloudMELT on AWS as it is the most widely used cloud computing platform (Figure 2.3.1). CloudMELT relies on Toil to create a compute cluster on Amazon Elastic Compute Cloud (Amazon EC2) and coordinate workflow distribution to the available virtual centralized processing units (vCPU) of the cluster.[222] CloudMELT allows users to define the size of their own compute cluster on Amazon EC2 to perform MELT analyses larger than possible on a local cluster. In addition, CloudMELT includes a significant performance improvement in the MELT Individual Analysis step reducing runtime from 387 min 32 sec to 19 min 33 sec resulting in a 20.4x speedup. Many consortiums host their genome sequencing data on the AWS cloud storage, also known as Amazon Simple Storage Service (Amazon S3). Having the large amounts of data on the same cloud platform as CloudMELT allows for fast, stable, and easy transfers.

As members of the Human Genome Structural Variation Consortium (HGSVC) we used CloudMELT to perform MEI discovery on 3,202 1kGP genomes sequenced at a targeted depth of 30X coverage (mean 34X). These 3,202 genomes are derived from the original 1kGP 2,504 unrelated samples from 26 diverse populations that were analyzed by Sudmant et al. (1kGP Phase 3 callset) and an additional 698 related samples.[153,223] Utilizing computing clusters composed of, on average, 15 worker nodes each with 16 vCPUs (storage optimized i3.4xlarge instances) capable of up to 10 Gbps (Gigabits per second) transfer speeds, we completed 3,202 genomes in approximately 4-5 weeks. In addition, we also performed discovery on 639 genomes from the Genotype-Tissue Expression project version 8 (GTEx) on CloudMELT that took approximately 11 days.[224,225] CloudMELT allowed us to significantly reduce the time to process thousands of genomes from almost a year on our local cluster to about 1.5 months on AWS.

2.3.4 Merged dataset

After completing MEI discovery on these large datasets, we merged them together to create the largest non-reference MEI dataset to date. We discovered 104,350 non-redundant MEIs (including 80,562 *Alu*, 16,525 L1, 6,956 SVA, and 307 HERV-K Insertions) in 57,919 human samples derived from the aforementioned studies (Figure 2.3.2A). When compared to previous MEI discovery efforts in the literature performed with earlier versions of MELT, 62.3% (65,192/104,350) of our MEIs are novel and had not been identified previously (Figure 2.3.3B). Examining existing repositories specific for L1, only 200 of our L1s were found in dbRIP (Database of Retrotransposon Insertion Polymorphisms, version 3) and 2,019 were found in euL1db (The European database of L1Hs retrotransposon insertions in humans, version 1.0).[227,228] These findings indicate that MEI discovery is far from complete in humans and there is still an unmet need for highly-scalable MEI discovery tools such as CloudMELT.

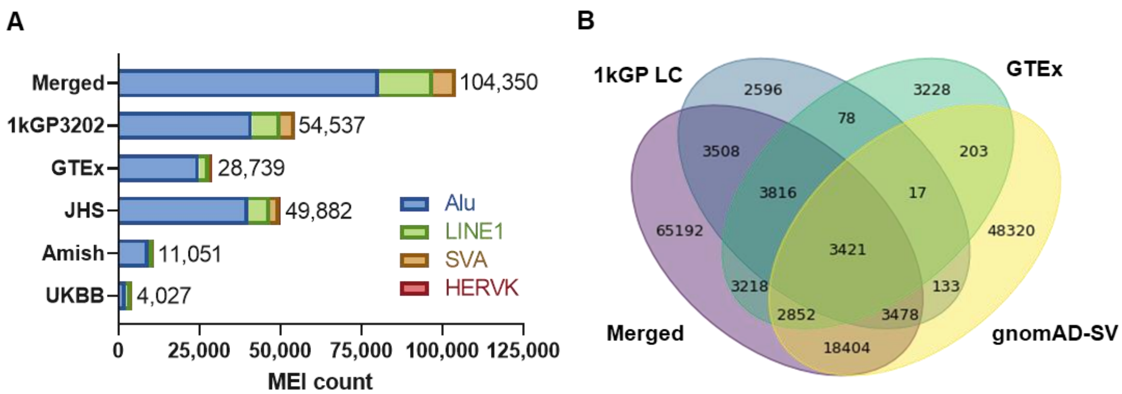


Figure 2.3.2 MEI discovery of five diverse populations

A. MEIs discovered displayed by their respective studies with bars colored by MEI type. Total discovered non-redundant MEIs, labeled "Merged", amount to 104,350 MEIs. **B.** Comparison of our study with other MEI discovery projects displayed as a 4-way Venn diagram shows that 65,192/104,350 (62.3%) of our MEIs are novel. (Merged, our merged dataset from multiple cohorts; 1kGP LC, 1000 Genomes Project 2,504 low coverage genomes dataset from Sudmant et al.; GTEx, GTEx dataset using rSD and PASS filter from Cao et al.; gnomAD-SV, gnomAD-SV dataset from Collins et al.)[153,207,210,226]

A. 1kGP High Coverage of Unrelated Samples (n=2,504)					
MEI	**Total**	**5'UTR**	**3'UTR**	**Exon**	**Intron**
L1	8,179	18	59	9	444
Alu	38,643	55	343	51	4,186
SVA	4,006	10	65	10	186
HERVK	72	0	1	0	17
Total	50,900	83	468	70	4,833

B. 1kGP High Coverage of All Samples (n=3,202)					
MEI	**Total**	**5'UTR**	**3'UTR**	**Exon**	**Intron**
L1	8,835	20	62	10	1,486
Alu	41,189	58	362	55	17,910
SVA	4,430	11	17	11	2,274
HERVK	83	0	1	0	18
Total	54,537	89	442	76	21,688

Table 2.3.3 1kGP High Coverage MEI discovery.

CloudMELT MEI discovery in high coverage 1kGP (**A**) 2,504 unrelated samples and (**B**) 3,202 samples including the additional 698 related samples. The Total column is the sum of 5'UTR, 3'UTR, Exon, Intron, and intergenic insertions (intergenic not shown).

2.3.5 MEI analysis of 1kGP high coverage data

CloudMELT discovery of the 1kGP 2,504 high coverage genomes showed an approximately three-fold increase in MEIs discovered (50,900 MEIs: 8,179 LINE-1s, 38,643 *Alu*, 4,006 SVA, 72 HERV-K) with ~30X coverage compared to initial discovery with ~7.4X of the same individuals (Table 2.3.3A).[153] 37,296 MEIs are novel (73%) compared to the original discovery call set. Merging the additional 698 related genomes increased the number of discovered MEIs by 3,637. Overall, MEI discovery of the 3,202 high coverage genomes resulted in 54,537 total MEIs with 8,835 LINE-1, 41,189 *Alu*, 4,430 SVA, 83 HERV-K (Table 2.3.3B).

We next focused our analysis on the 2,504 unrelated individuals and counted the number of MEIs per individual of the 26 diverse 1kGP populations. We found higher numbers of MEIs per individual compared to previous studies with the same 26 populations.[153,207] This outcome is not surprising, given the higher WGS coverages in this study, an improved reference human genome build (GRCh38), and other technical improvements. The highest counts per individual were observed in the AFR populations, which is consistent with the original 1kGP Phase 3 data.[153] We also observed a range of MEI count variation within each of the 26 populations primarily due to variation in *Alu* counts (Figure 2.3.3A). Sharing analysis of the MEIs across the 26 diverse populations revealed MEIs that are found in all 26 populations (ALL), more than one (SHARED), or that were unique to a specific population (UNIQUE) (Figure 2.3.3B). Each population had similar profiles of shared sites (Figure 2.3.3B), and each population had close to 1,000 unique sites (with exception of AFR populations, LWK and GWD, at around 2,400 and 1,500 unique sites, respectively) that are mostly singletons. MEIs that are unique to a single population include singletons, which have an allele count of 1, and MEIs with non-singleton frequencies (Figure 2.3.3C). AFR populations have lower proportions of MEIs found in all populations, but higher proportions that are shared. The other superpopulations have higher proportion of MEIs that are found in all populations, which is further support of lower genetic diversity compared to the AFR populations. The proportion of unique MEIs are similar to the other superpopulations, and the majority of MEIs unique to a population are also singletons found in only a single individual.

We performed the same analyses comparing the Amish, JHS, and UKBB populations and found that the Jackson Heart Study had higher MEI counts than the Amish and UKBB populations (Figure 2.3.4A). Similar to findings discussed above in Table 2.3.1, the JHS study of African-Americans have more total MEIs and more MEIs per individual compared to the Amish population (Figure 2.3.4A).[153] Thus, the lower levels of MEI variation in the Amish population compared to the JHS population is consistent with the population histories of these two groups. The UKBB population had much lower MEI counts because MEI discovery was performed in whole exome sequences instead of whole genome sequences (Figure 2.3.4C). Shared MEI analysis revealed only 23 MEIs that were shared by all three populations (Figure 2.3.4D). In contrast to the 1kGP populations, where MEI discovery was performed in ~100 individuals from each of the 26 diverse populations, we performed MEI discovery in 1,112 Amish genomes, 3331 JHS genomes, and 49,635 UKBB exomes. Thus, as expected, there was more unique variation in the Amish, JHS, and UKBB samples that was not shared between these diverse populations (Figure 2.3.4E).

2.3.6 *Alu* and SVA subfamily analyses

We performed subfamily analyses on *Alu*, SVA, and L1 (Please see next chapter for results and discussion of L1 subfamilies). In agreement with previous studies, the majority of discovered *Alu* elements belong to young *Alu*Y subfamilies that are known to be polymorphic and active in humans (e.g., 81% of the elements are *Alu*Y and *Alu*Ya, *Alu*Yb, and *Alu*Yc lineages similar to what has been observed previously; Figure 2.3.5).[207] We compared the SVA subfamily distribution of the build GRCh38 reference genome with our SVA subfamily distribution from non-REF populations. The youngest subfamilies, SVA-E and SVA-F, have undergone a recent expansion and have tripled in proportion compared to the REF SVA-E and SVA-F elements (Figure 2.3.6). These data combined with L1 subfamily analysis in the next chapter demonstrate there are only a few dominant lineages of *Alu*, L1, and SVA that are responsible for the majority of MEI expansion and ongoing mutagenesis in human populations.

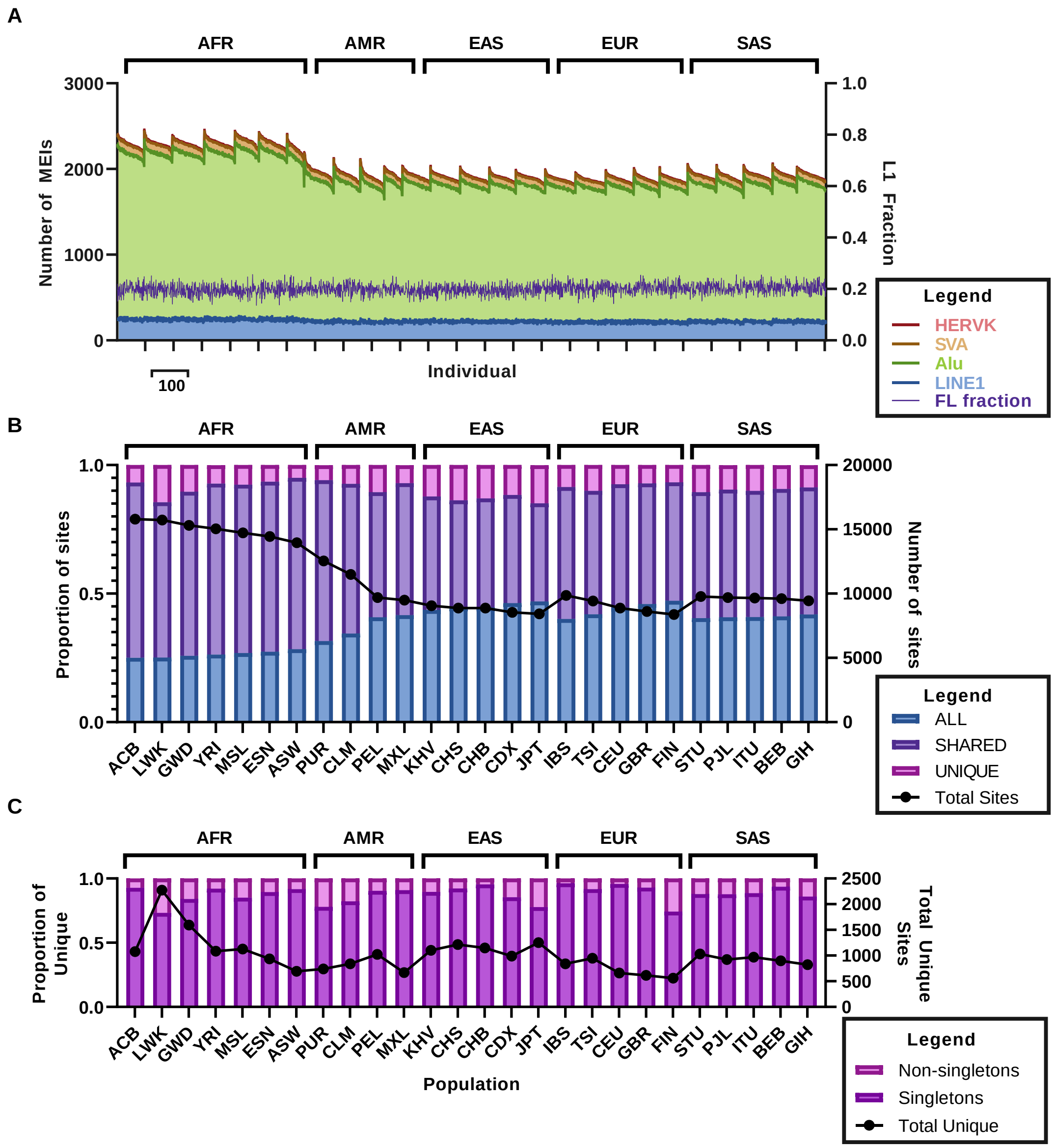


Figure 2.3.3 Distribution of MEI counts per individual and MEI sharing across 26 diverse 1kGP populations.

A. The number of MEIs per individual (L1 = light blue; *Alu* = light green; SVA = light brown; HERV-K = red) grouped by 1kGP-defined populations and sorted by superpopulation and highest average of MEIs. The dark lines of the same colors delineate the boundaries of each MEI type. The dark purple line represents the Full-length (FL) fraction, which shows the fraction of total L1s that are full-length for each individual. Note that the AFR populations have the highest MEI counts, consistent with previous studies.[153] **B.** Sharing of MEIs across the 26 diverse populations of 1kGP. The solid black line and the right Y-axis show the total number of MEIs for each population. **C.** MEIs that are unique to a single population of the 26 1kGP populations from panel B broken down into singletons and non-singletons. Total unique sites per population are shown by solid black line and right Y-axis.

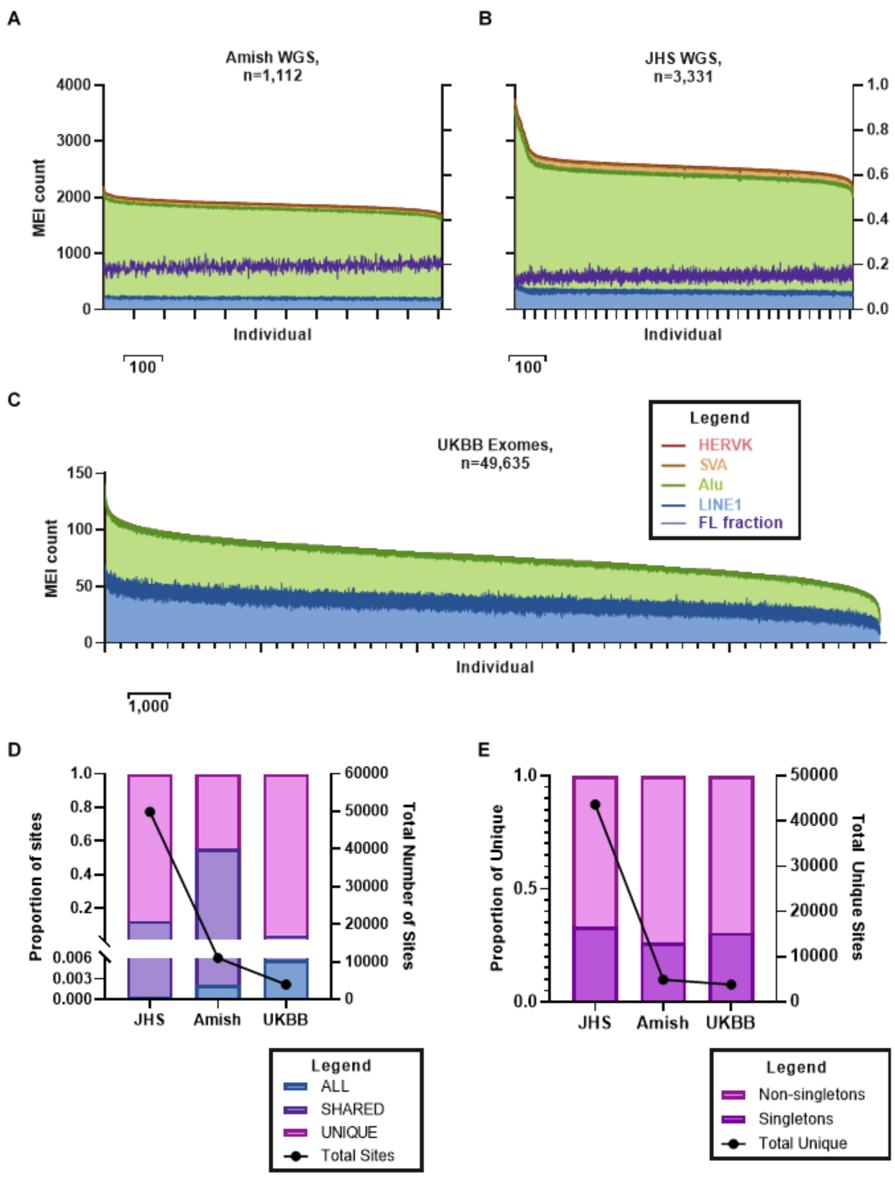


Figure 2.3.4 Distribution of MEI counts in TOPMed and UKBB individuals.

A-C. MEI counts (L1=light blue; *Alu*=light green; SVA=light brown; HERV-K=light red) are plotted for each individual in the three populations. The dark lines of the same colors delineate the boundaries of each MEI type. The dark purple lines indicate the fraction of non-REF L1s that are FL-L1Hs elements (plotted for whole genome data only). Note that the Jackson Heart Study (**B**), which is African Americans, has increased MEI counts compared to the Amish population (**A**). These results are consistent with a European bottleneck associated with the Amish population compared to the JHS African-American population.[217] Note also that the UKBB individuals (**C**) have fewer MEIs compared to the TOPMed Jackson Heart and Amish populations, consistent with the fact that only the exome portions of the genome were analyzed in the UKBB individuals. **D, E.** Analysis of shared MEIs across the three populations.

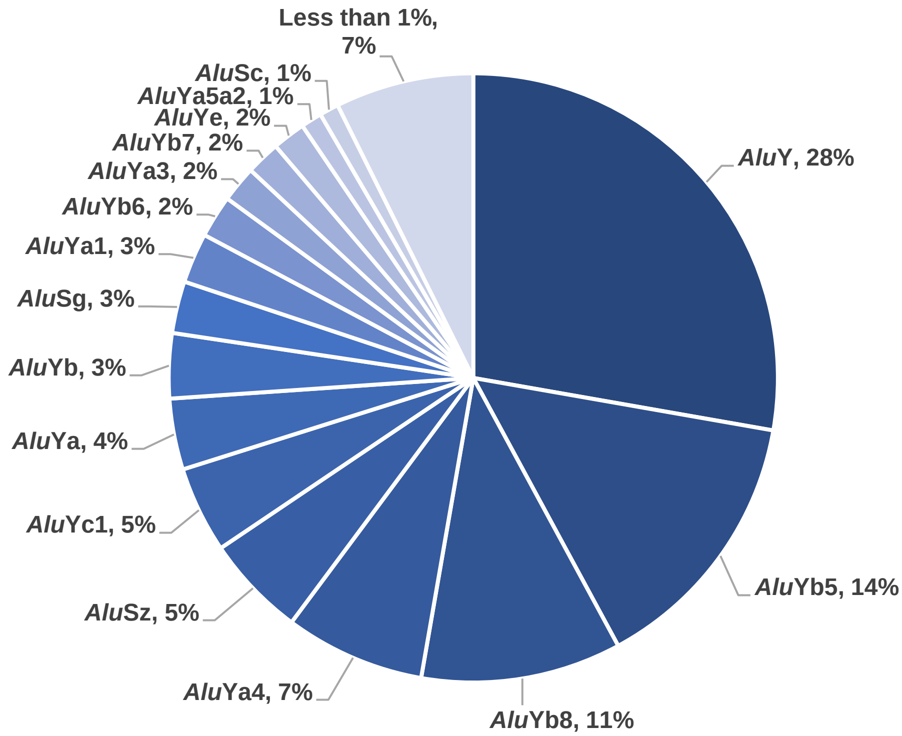


Figure 2.3.5 Subfamily analysis of *Alu* MEIs.

Note that the majority of subfamilies that we identified are young *Alu*Y subfamilies that are known to be polymorphic and active in humans (e.g., 81% of the elements belong to *Alu*Y and the *Alu*Ya, *Alu*Yb, and *Alu*Yc lineages similar to what has been observed in previous studies.[207] Nine percent of the elements belong to three *Alu*S subfamilies (*Alu*Sz, *Alu*Sg, *Alu*Sc), which are known to be polymorphic in humans as well.[91] *Alu* subfamilies were annotated using the CAlu package of MELT.[207]

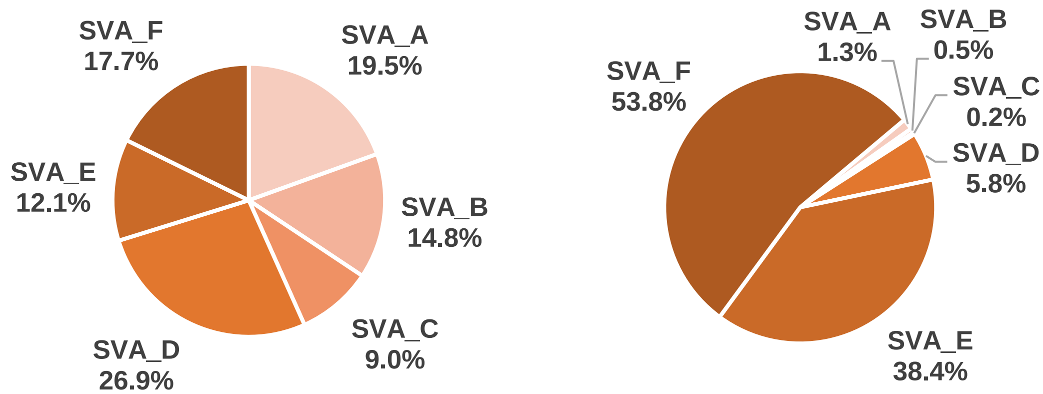


Figure 2.3.6 Subfamily analysis of SVA MEIs.

A. Classification of SVA subfamilies in the build GRCh38 reference genome. **B.** Classification of the non-REF SVAs that we discovered based on the approach outlined in Wang et al. 2005 using 3’ SINE sequences.[86] Note that the youngest SVA_E and SVA_F elements have undergone a recent expansion from 29.8% in the REF genome (**A**) to 92.2% of the non-REF SVAs (**B**).

2.3.7 Impact of MEI mutagenesis on genes and other important genomic features

We also leveraged our collection of *Alu*, L1, SVA, and HERV-K insertions to examine the impact of MEI mutagenesis on human genomes. We combined our MEIs with the others shown in Figure 2.3.2B to create a non-redundant set of 158,873 MEIs. Accessing two large databases of DNA elements (ENCyclopedia Of DNA Elements, or ENCODE) and gene features (GENCODE), we compared the coordinates of these MEIs with those of GENCODE gene features and ENCODE DNA elements (Figure 2.3.7).[229,230] 97,103 MEIs map to GENCODE-annotated transcripts with 274 MEIs disrupting 5'UTRs, 2,449 disrupting 3'UTRs, 4,639 disrupting exons, and 781 disrupting coding exons (CDS) of annotated genes (Figure 2.3.7A) Two MEIs disrupt start codons, 11 disrupt stop codons, and 101 disrupt splice sites (Figure 2.3.7A). MEIs also disrupt 11,586 ENCODE cis-Candidate Regulatory Elements (cCREs) (Figure 2.3.7B).

We next searched the Online Mendelian Inheritance in Man (OMIM) database to identify MEIs that disrupt genes implicated in human diseases.[231] We identified 16,606 MEIs that disrupt 2,335 OMIM-annotated genes. These genes were further grouped by disease, and we identified 20 clusters of disease genes that are disrupted by MEIs (Figure 2.3.7C). For example, 177 genes that have been implicated in diverse human cancers are disrupted by MEIs in our collection (Figure 2.3.7C). We next expanded this approach for the cancer group to include the Catalogue of Somatic Mutations In Cancer (COSMIC) database, which curates genes that have been implicated in cancers, the Genome Association Database (GAD), which curates loci that have been implicated in human cancers from GWAS studies, and the Candidate Cancer Gene Database (CCGD), which curates genes that have been implicated in mouse studies involving transposon mutagenesis.[232–234] We identified MEIs that disrupt 7,166 genes in these four databases. Among this collection, we found 376 MEIs that disrupt the CDS's of cancer-related genes, including well-defined tumor suppressor genes and oncogenes (e.g., *APC, BRCA2, DNMT1, FANCC, FANCI, FANCM, MSH6, and XRCC2*). Individuals who harbor such MEIs presumably would be at greater risk for developing cancers, particularly if the second allele of a tumor suppressor gene should be mutated or silenced during their lifetime. We also identified similarly relevant genes and mutations for the remaining disease groups in OMIM (Figure 2.3.7C).

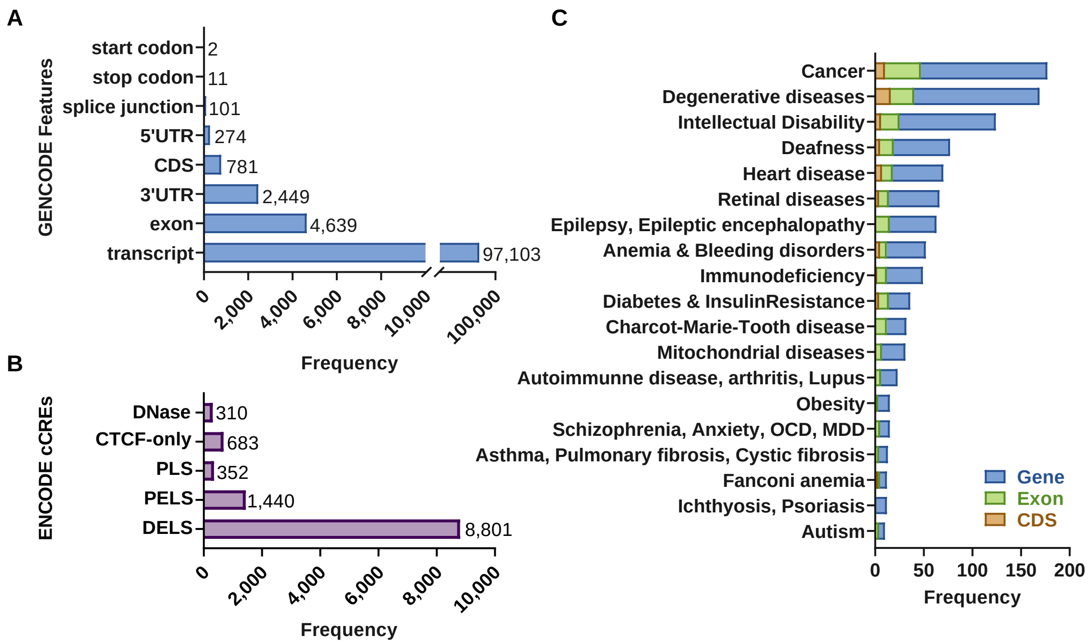


Figure 2.3.7 MEI mutagenesis patterns in GENCODE, ENCODE, and other databases

A. Comparisons of the combined collection of 158,783 MEIs depicted in Figure 2.3.1 revealed intersections with GENCODE v35 gene annotations. We identified insertions that disrupt various subregions of genes including 781 MEIs that disrupt CDS exon sequences. **B.** MEIs also disrupt ENCODE annotated cis-regulatory elements (cCREs): DNase-H3K4me3, high DNase and H3K4me3 element of unknown function; CTCF-only, CCCTC-binding factor only; PLS, promoter-like signature; PELS, proximal enhancer-like signature; DELS, distal enhancer-like signature. **C.** MEIs also disrupt genes that have been linked to various disease in the Online Mendelian in Man (OMIM) database.

2.4 Discussion

Our collection of 104,350 MEIs combined with the remaining MEI collections in Figure 2.3.2B (a total of 158,873 non-redundant MEIs) allowed us to examine MEI mutagenesis in humans on an unprecedented scale. We found 781 MEIs that disrupted the CDS's of GENCODE-annotated genes and we expect that most of these would disrupt gene function. We also identified MEIs in other annotated features of GENCODE genes, and a subset of these likely also affect gene function (Figure 2.3.7A). Finally, we identified 11,586 MEIs that disrupt ENCODE cCREs, and thus, might impact the expression of genes that are regulated by these cCREs (Figure 2.3.7B). Most of these MEIs belong to young subfamilies, are rare, and likely integrated into human genomes relatively recently. Many of these MEIs undoubtedly were mobilized by FL-L1Hs source elements belonging to three new subfamilies discovered in this study, which will be described in the next chapter. Such source elements are young and active, and they would be responsible for mobilizing not only new L1 insertions, but also new *Alu* and SVA insertions.

We examined the population distributions and subfamilies of the MEIs we discovered in this study. Our analysis indicates that there is substantial variation in the MEI counts per individual in these populations, largely due to variation in the number of *Alu* MEIs (Figure 2.3.3A; Figure 2.3.4A). Likewise, we identified MEIs in each population that are rare or unique to a single population (Figure 2.3.4C; Figure 2.3.5C). These population-specific variants might be useful to determine the genetic ancestries of individuals and might also help to account for population-specific phenotypes or susceptibility to certain diseases. We also examined the subfamily status of the *Alu* and SVA insertions in our MEI collections, and found that a few dominant lineages of *Alu*, L1, and SVA elements are responsible for the majority of ongoing MEI mutagenesis in human populations (Figure 2.3.5; Figure 2.3.6). Please see the next chapter for L1 analysis.

Analysis of the OMIM database revealed 16,606 MEIs that are located within 2,335 genes that have been implicated in a wide range of human diseases, including cancers, degenerative diseases, intellectual disability, deafness, heart disease, and others (Figure 2.3.7C). Overall, our data indicate that MEI mutagenesis impacts a wide range of genes and gene features that are relevant to a broad spectrum of human diseases. Since the majority of MEIs that we discovered in this study are rare (81,645/104,350 or 78.2%), MEIs that disrupt CDS's or other functionally important genomic features (i.e., loss-of-function, or LoF, variants) likely would fit a rare variant model of human disease.[235–237]

2.5 Methods

2.5.1 CloudMELT pipeline

CloudMELT version 1.0 is a port of MELT version 2.1.5 to the Amazon EC2 cloud computing environment.[207] The MELT 2.1.5 engine was further optimized to provide ~20X speedup in the MELT Individual Analysis step as well as provide deterministic results in the MELT Group Analysis step. The resulting new engine is termed MELTv2.1.5.fast. The pipeline was written with Toil version 3.17 to create a reproducible data analysis workflow.[222] Docker was used to create an image of the optimized MELTv2.1.5.fast engine with its required dependencies and the image was loaded to an AWS container repository.[238] CloudMELT requires users to configure initial parameters for their run: define the reference genome version, provide the uniform resource identifier (URIs) of the genomes of their dataset, set MELT flags/arguments, and optionally designate minimum storage and memory requirements. Users then launch a computation cluster on AWS with a "leader" node and a desired number of "worker" nodes. The CloudMELT workflow (Figure 2.3.1) is created and uploaded to the leader node. The MELT docker image and the necessary AWS permissions are distributed to the worker nodes. After this setup, the CloudMELT pipeline is ready to launch.

CloudMELT is then launched as follows. For phase 1 of the CloudMELT pipeline, the genome alignment files (i.e., BAM or CRAM) are downloaded to the respective worker nodes for MELT Preprocess and Individual Analysis steps to discover the user-defined MEIs (*Alu*, L1, SVA, and HERV-K). Please refer to the original MELT paper for detailed description of each MELT runtime step.[207] The data generated from the worker nodes are consolidated to the leader node. Phase 2 is the MELT Group Analysis step where the phase 1 output is processed on a single worker node. Phase 3 is the MELT Genotyping step where the phase 2 output is used by each worker node to determine the genotype of the candidate MEIs. Finally, the phase 4 Make VCF step aggregates the phase 3 output into a VCF file.

The MELT deletion caller, MELT-Deletion, also is included in CloudMELT and can be added as another "MEI" in the pipeline. The deletion caller identifies reference MEIs that are "deleted" in the sample being analyzed. CloudMELT phase 1 would include the MELT-Deletion Genotype step after the Preprocess and Individual Analysis steps. Similarly, the output would be transferred to the leader node and the Merge step proceeds in CloudMELT phase 2 after the Group Analysis step. The MELT-Deletion VCF is created during the Merge step.

Previously, we performed whole genome MEI discovery simulations and extensive PCR validation studies (on 400 independent MEI sites) and achieved an overall false discovery rate (FDR) below 5% with low coverage Illumina WGS data.[153,207] Our analysis of high coverage Illumina WGS data in this study indicated that the higher coverages yielded more accurate calls and lower FDRs. Therefore, we re-adjusted our filtering process to achieve an overall FDR of <5%. An initial 1kGP call set of PASS MEIs (39,830) was generated with the original (more stringent) criteria, and an extended call set was generated that includes 14,707 additional high quality MEIs (for a total of 54,537 calls; Figure 2.3.2A). This adjusted filtering was used for all other datasets. As we developed CloudMELT, we compared the results obtained with MELTv2.1.5 vs the improved MELTv2.1.5.fast and CloudMELT to ensure comparable results. We also examined 50 to 100 MEIs of each MEI type in the Integrative Genome Viewer (IGV) during various stages of development and in new datasets to ensure that there was strong evidence to support the calls and that our overall FDRs were below 5%. [239]

2.5.2 Improvements to MELT code

The improvements that we engineered into the MELT engine were focused on the Individual Analysis step. This is the second of six steps in the MELT pipeline, and we found that it was a major bottleneck with high coverage (30-40X) Illumina genome sequences. The Individual Analysis step is focused on identifying clusters of discordant read pairs that occur at sites with new mobile element insertions. The original Individual Analysis step was coded to scan along each chromosome to identify clusters of discordant read pairs using a set of rules that we developed and tested.[153,207] Although this approach worked well with low coverage genomes (~7X, which is the depth for the original 1kGP), it became a noticeable bottleneck as 30-40X coverages were examined due to the larger number of discordant read pairs at each site. To reduce the runtime, we re-coded this step to generate a hash table of all discordant read pairs (DRPs) that were discovered in the first step of MELT (Preprocess) and then identified clusters of DRPs in memory using the same rules that we developed previously in terms of the allowed sizes of cluster regions, minimal number of DRPs as a function of sequence coverage, and proximity to REF mobile elements. This approach generated comparable results in terms of MEI calls and their features, while dramatically reducing the runtime. We also achieved additional major improvements in speed by adapting this version of MELT to the Cloud, where we could rapidly examine many genomes in parallel using optimal hardware configurations.

2.5.3 SVA subfamily analysis

SVA subfamily classification was based on the approach described in Wang et al 2005.[86] The diagnostic nucleotides for each subfamily are determined by the SINE-R region. SVA subfamily consensus sequences were downloaded from the Wang et al. supplemental data for comparison. MELT analyzes each discovered SVA for nucleotide changes against a reference SVA consensus sequence. These changes are used to construct a nucleotide sequence for each non-REF SVA. Each non-REF SVA SINE-R sequence is then pair-wised aligned against each of the five SVA subfamily consensus sequences. The subfamily is determined by the highest match score.

2.5.4 Data sets used in this study

1kGP, UKBB, and GTEx BAM files aligned to build GRCh38 of the reference human genome were used for initial MEI discovery in these populations. TOPMed BAM files aligned to build GRCh37, or hg19, of the reference human genome were used for initial MEI discovery in these populations and the MEI coordinates were converted to build GRCh38 coordinates using the University of California Santa Cruz Genome Browser Database (UCSC GBD) LiftOver tool.[240] All downstream analysis was performed with the GRCh38 reference genome sequence and annotation.

The high coverage Illumina Amish, Jackson Heart Study, and GTEx WGS data were obtained from database of Genotypes and Phenotypes (dbGaP) under Institutional Review Board approvals (accession numbers phs000956.v4.p1, phs000286.v6.p2, and phs000424.v8.p2, respectively). The high coverage 1kGP WGS data were obtained from AWS (s3://1000genomes/1000G_2504_high_coverage/, Ebert et al. 2021).[241] The UKBB WES data were obtained from the UK Biobank resource (https://www.ukbiobank.ac.uk/). The data outlined in Figure 2.3.2B were obtained from Sudmant et al. 2015, Cao et al. 2020, and Collins et al. 2020.[153,210,226] GENCODE and ENCODE v35 data were obtained from the University of California Santa Cruz Genome Browser (https://genome.ucsc.edu/).[240] OMIM and GAD data were downloaded from https://david.ncicrf.gov/summary.jsp; COSMIC data were downloaded from http://cancer.sanger.ac.uk/cosmic, and CCGD data were downloaded from http://hst-ccgd-prd-web-oit.umn.edu/.

Chapter 3. Full-length L1Hs sequencing and subfamily analysis

3.1 Abstract

FL-L1Hs elements are integral to the mobilization of both L1 elements and non-autonomous retrotransposons. The “copy and paste” mechanism that is used by FL-L1Hs is generally understood, but the factors that differentiate FL-L1Hs activity levels are still poorly defined. A large reason is due to limitations in capturing the full internal sequences of FL-L1Hs elements. We designed a novel pipeline to sequence the internal sequences of 698 FL-L1Hs elements to examine patterns of variation that may contribute to their activity. Interior sequences allowed us to determine whether FL-L1Hs elements have two intact ORFs, belong to young subfamilies, or otherwise have internal mutations that might impact L1 mobilization. Internal sequences also allow for more accurate tracking of source and offspring relationships. Using the internal sequence data that we discovered from these 698 FL-L1Hs elements, we discovered three new subfamilies that are responsible for the majority of recent L1 insertions. We used previous FL-L1Hs element activity data from Beck et al to verify that these new subfamilies are the most active subfamilies.[242] We also show the frequency of sequence variation in the collection of FL-L1Hs elements that we sequenced. This dataset will be valuable to further study mechanisms that influence FL-L1Hs element activity.

3.2 Introduction

In the previous chapter we discussed MEI dynamics across multiple diverse populations. There is still much to be learned about MEI patterns and their effect on human health, but what is shared by these MEIs, with the exception of HERV-K, is their mode of mobilization to create these MEI patterns in each person and population. *Alu*, L1, and SVA require L1 to provide the necessary machinery for retrotransposition. In this chapter, we focus on understanding the FL-L1Hs source elements that drive the formation of new *Alu*, L1, and SVA insertions, and the determinants of an active FL-L1Hs elements.

For a FL-L1Hs element to be retrotransposition-competent, it relies on not just being 6 kb in length but also having an intact 5'UTR, ORF1, ORF2, 3' UTR, and poly(A) signal. As introduced earlier, this means the 5'UTR needs to retain binding motifs for initiating transcription, ORF1 and ORF2 need to generate functional ORF1p and ORF2p (*Alu* only needs ORF2p), and a poly(A) signal in the 3'UTR that can direct the addition of a poly(A) tail to the end of the transcript. Factors that inactivate L1 include 5'truncations, internal rearrangements, point mutations, deletions, and even the insertion of other MEs. The 5'UTR is the least conserved region and accumulates mutations quicker than the rest of L1, and sequences upstream of 5'UTR can either repress or enhance activity.[243,244] L1s have multiple non-canonical poly(A) sites upstream of the 3'UTR. These premature poly(A) signals can result in truncation of the L1 transcript and potential loss of L1 activity.[120] Splice donor and acceptor sites in the 5'UTR and ORF1 can lead to 5' truncation and internal deletion during posttranscriptional processing of the RNA, and when these shortened transcripts are integrated into the genome they are termed spliced integrated retrotransposed elements (SpIREs).[245,246] Even though the great majority of L1 copies in the human genome are inactivated, there are still estimated to be up to 100 retrotransposition competent FL-L1Hs elements in the reference human genome.[108,242]

A FL-L1Hs element that is active, or able to generate “offspring” insertions, is considered a “source” element. We are able to determine if an L1 is a source element through either identifying 3’transductions that map to an “offspring” L1 or through *in vitro* cell culture retrotransposition assays. The degree of activity can be measured by the number of offspring L1s a source element generates. For example, one of the most active known L1s, LRE3, accounted for 33.9% (41/121) of the detected offspring insertions with 3’transductions in the 1kGP Phase 3 dataset.[207] With a retrotransposition assay, the L1 of interest is ligated into a vector with a unique *mneoI* retrotransposition indicator cassette that allows for detection and quantification of L1 retrotransposition.[247] The retrotransposition assay allows identification of which FL-L1Hs sequences are active and to what extent they are active. Most L1s and even FL-L1Hs elements are inactive, and it is likely that the majority of L1 offspring (and MEIs) are being generated by a small subset of “hot” L1s .[108,248]

Knowing which L1s are 6 kb and have intact ORF1 and ORF2 is a fundamental step to tracking L1s that could be active source elements. MELT is an excellent tool for calling MEIs and estimating their sizes, but it cannot determine if the ORFs are intact. MELT was developed to call MEIs from short read paired-end sequencing data, and the inherent limitation of this technology is resolving long repetitive regions (Figure 3.2.1). Paired-end short reads are typically a pair of short 50-150 bp sequences captured at a set distance between them to create fragments of the desired “insert” size. For example, one of the most widely used sequencing platforms, Illumina HiSeq X, supports paired-end reads of 150 bp with a maximum insert size of 450 bp.[249] The paired-end short reads can map the ends of the L1 and provide information for MEI detection. However, since full-length L1s span a long stretch of DNA much greater than the insert size and their sequence is highly repetitive in the genome, paired-end short reads are unable to map the interior sequences of individual FL-L1Hs loci.

The interior sequence of FL-L1Hs elements is not only important for understanding which L1s are still active, but also important for studying the evolution and phylogeny of L1 subfamilies. L1 subfamilies are determined by the nucleotide differences and similarities found in their L1 progenitors. L1 has been part of the mammalian genome for over ~170 million years with multiple lineages of subfamilies.[4,250] During the evolution of ancestral primates over the past 70 Mya there were only three distinct L1 lineages, and during the past 40 Mya of anthropoid primate evolution a single dominant lineage emerged (Figure 3.2.2).[32] It is believed that these L1

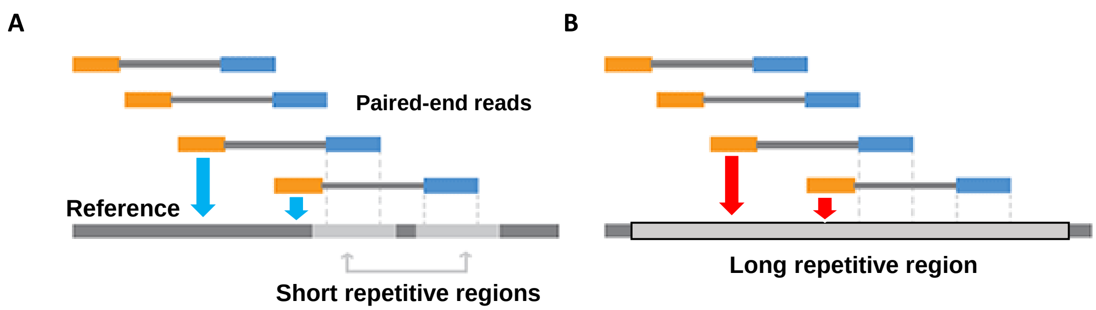


Figure 3.2.1 Limitations of short read paired-end sequencing for long repetitive regions

Repetitive elements are difficult for paired-end reads to distinguish in the genome if the read falls within the repetitive region. **A.** Short paired-end reads can map to the correct locus of the reference if one end aligns inside the repetitive region and its paired end aligns outside in a non-repetitive region. The non-repetitive region serves as an "anchor" to map the location of the paired-end reads within the reference sequence. **B.** The short paired-ends depicted are unable to anchor the sequence to a non-repetitive region if the repetitive region is longer than the length of the insert size. (Adapted from http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html) lineages compete with one another in the host (i.e., for host transcription factors) until one lineage becomes the most successful in self-propagation.[251,252]

Multiple lineages were able to coexist when their 5'UTRs did not necessitate competition for the same transcription factors.[253–255] Through time, the ORF1 and ORF2 amino acid sequences have remained relatively conserved, but the 5'UTR underwent the most change between lineages.[32] Around the time L1 lineages were reduced to a single lineage, there was coincidentally a period of increased amplification of L1, *Alu*Y ,and processed pseudogenes.[256,257] The increase in amplification can be attributed to the dominant L1 lineage acquiring an advantageous 5'UTR that reduced the fitness of competing lineages.[32] Studies of recent and most active L1s are limited and understanding the nucleotide changes could help understand the factors that have permitted the youngest L1s to have the most replicative success.

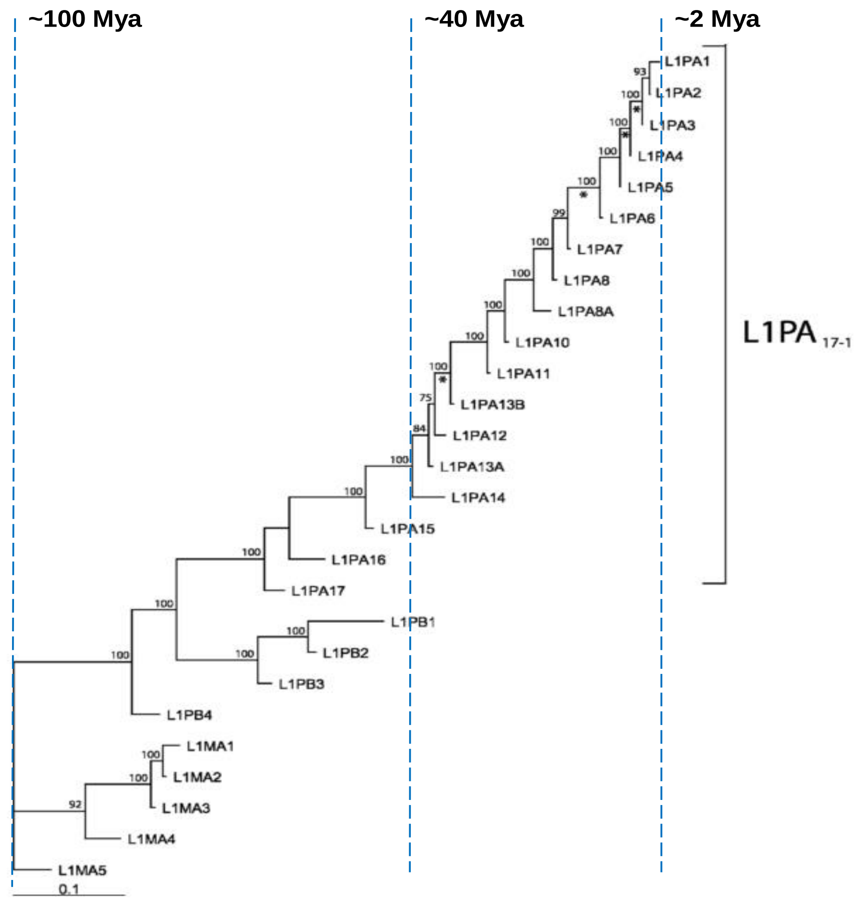


Figure 3.2.2 L1 sequential evolution through time.

Evolution of L1 from approximately 100 Mya is shown based on 27 subfamily consensus ORF1 and ORF2 sequences. By 40 Mya, L1PA has replaced L1PB and L1MA as the dominant subfamily. This single lineage has continued evolving to the present day L1PA1 subfamily. (Adapted from Khan et al. 2006)[32]

3.3 Results

3.3.1 Nonreference FL-L1Hs

Of the 14,066 non-REF L1s discovered from our collection, 3,728 (26.5%) were at least 6 kb in length and 12,797 (73.5%) were truncated (Figure 3.3.1A). Examining the size distribution of L1 shows a bimodal distribution with a small peak centered at 400 bp and another larger peak at 6 kb (Figure 3.3.1A). We next examined the FL-L1Hs copy number distribution in the four WGS populations in our study. We found that the FL-L1Hs per individual follows a uniform distribution for each study and the difference in mean of each study was statistically significant (Figure 3.3.1B). The JHS population had on average the greatest number of FL-L1Hs elements per individual, whereas the Amish population had the least number of FL-L1Hs elements per individual. Not only are there more potential source elements in the JHS population, but there are also more elements that have been demonstrated in cell culture or by transductions to have increased activity in JHS (Figure 3.3.2). The GTEx population is 89% Caucasian, and thus, its mean is closer to the mean of the Amish, as expected.

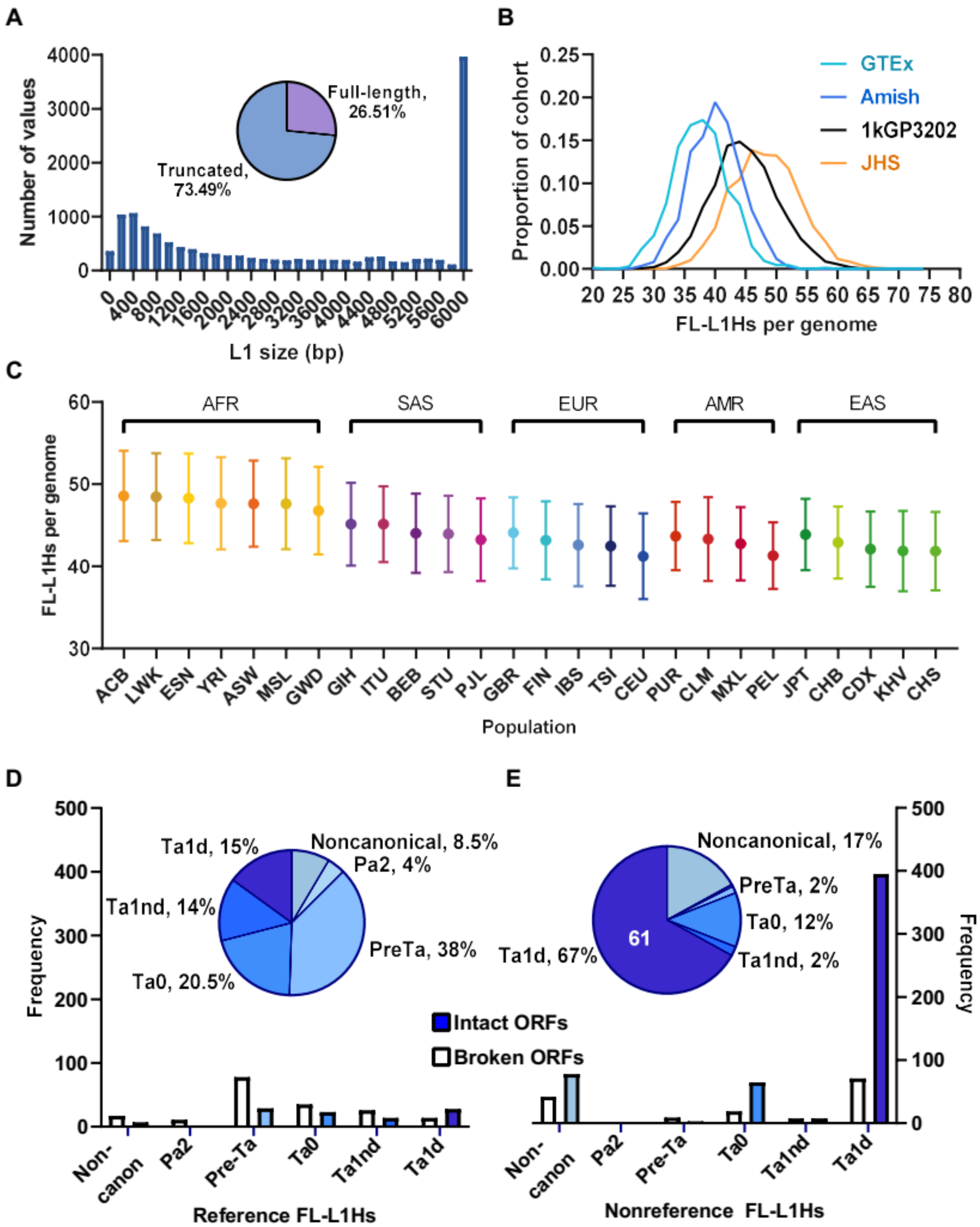


Figure 3.3.1 Nonreference Full-length L1s

A. There are 16,525 nonreference L1 insertion sizes obtained from the merged collection of MEIs with 3,728 FL-L1Hs elements that are at least 6 kb in size. There is a bimodal distribution centered around 400 bp and 6,000 bp size bins. **B.** Distribution of the proportion of FL-L1Hs elements per individual for each study. The difference of the mean of each distribution is statistically significant by One-way ANOVA with multiple testing correction by Tukey method. **C.** The FL-L1Hs elements per individual sorted by the 26 1kGP populations and grouped by their respective super populations. See Table 2.2.1 for 1kGP abbreviations. **D, E.** Comparison of reference FL-L1Hs to nonreference FL-L1Hs according to their subfamily distribution.

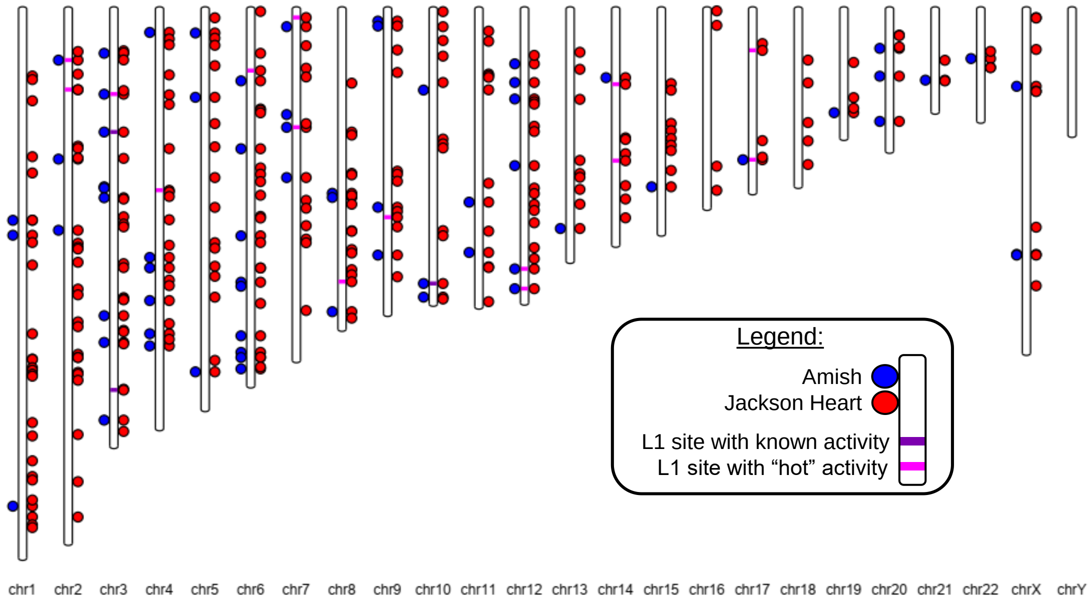


Figure 3.3.2 Non-REF FL-L1 distribution

Ideogram showing the distribution of non-REF FL-L1 elements between the Amish study and the Jackson Heart study. Amish FL-L1 elements are displayed on the left side of the karyotype and the Jackson Heart FL-L1 elements are displayed on the right side. Sites with known activity in the literature are highlighted purple and sites described as "hot" in the literature are highlighted pink.

Looking at the 1kGP data set, we observed variation in the number of FL-L1Hs elements per individual across the 26 diverse populations (Figure 3.3.1C). When broken down by the 26 populations and grouped by their superpopulations, the populations of African descent have the most FL-L1Hs elements per individual compared to the other superpopulations. Although there is on average 44.3 non-REF FL-L1Hs elements in 1kGP populations, the copy numbers per individual range from 25 to 63. These data suggest that the levels of L1 mutagenesis and L1-mediated disease could vary considerably both within and across diverse populations.

3.3.2 Full-Length L1 interior sequence

We wanted to examine the interior sequences of the non-REF FL-L1Hs elements that we discovered in order to study the extent of L1s with intact ORFs and determine whether they belong to active L1Hs subfamilies. We devised a targeted sequencing pipeline based on Pacific Bioscience long read sequencing technology and successfully sequenced a sampling of 698/3,728 (18.7%) FL-L1Hs elements. We chose a diverse and polymorphic collection of FL-L1Hs elements discovered from the 1kGP Phase 3 MELT callset as the genomic DNA samples can be easily acquired from the Coriell repository (Coriell Institute for Medical Research, Camden, NJ).[207] We used custom designed primers in conjunction with long-range PCR amplification to sequence these 698 non-REF insertions from 310 independent 1kGP Coriell samples plus one additional patient sample. There were 647/698 (92.7%) FL-L1Hs elements that are at least 6 kb in length. The remaining 51 elements have small 5' truncations or internal deletions, and 10 of them were found to be SpIREs.

Most (519/647 or 80.2%) of our sequenced 6 kb elements have two intact ORFs and the majority of elements with two intact ORFs (381/519 or 73.4%) belong to the youngest, most active Ta1d subfamily (Figure 3.3.1E). Moreover, we found that the Ta1d subfamily has undergone a substantial expansion in non-REF populations compared to other L1 subfamilies (Figure 3.3.1D,E; note that FL-L1Hs Ta1d elements (purple) expanded from 15% of REF elements (D) to 67% of non-REF elements (E)). We also examined published studies to determine whether our non-REF Ta1d FL-L1Hs elements have been active in i) cell culture assays, ii) the germline, or iii) somatic cancers.[173,207,242,258] We found evidence of activity from these studies for 91 of our sequenced FL-L1Hs elements, 61 of which belong to Ta1d subfamily (Figure 3.3.1E, white numeral). Overall, our data indicate that the Ta1d subfamily is the most rapidly expanding and active group of FL-L1Hs elements in human populations.

3.3.3 Three novel, highly-active FL-L1Hs subfamilies in human populations

We leveraged the data from our sequenced FL-L1Hs elements to determine whether we could identify new L1 subfamilies that might be particularly active. In fact, we identified three novel subfamilies that recently evolved from the highly active Ta1d subfamily (Figure 3.3.3A,B). These subfamilies are defined by interior sequence changes at positions 1026, 3337, and 3340 when aligned to a L1Hs-Ta subfamily consensus sequence (Figure 3.3.3A). We named these subfamilies Ta1d-CAT, Ta1d-CCA, and Ta1d-TCA based on the sequences that are present at these three positions. The Ta1d-TCA subfamily appears to be the youngest and most active subfamily, as it has undergone the largest recent expansion in non-REF populations (Figure 3.3.3C,D, dark purple; Note that the Ta1d-TCA subfamily went from 9% in REF (C) to 32% in non-REF (D)). However, the Ta1d-CAT and Ta1d-CCA subfamilies also are relatively abundant in non-REF populations (Figure 3.3.3D). We next identified collections of FL-L1Hs elements that belong to these three subfamilies and also have been sequenced and tested for activity in cell culture assays previously by Beck et al. 2010.[242] Two of our three new subfamilies (Ta1d-CCA, Ta1d-TCA) had the highest levels of activity in the cell culture assay compared to the other subfamilies tested, with the Ta1d-TCA subfamily having the highest level of activity (Figure 3.3.3A). We conclude that the Ta1d-TCA subfamily is the most active new subfamily in humans, followed by Ta1d-CCA and Ta1d-CAT (Figure 3.3.3B,C,D; Figure 3.3.4A).

We examined previous studies to determine whether any of the well characterized FL-L1Hs elements that are highly active or otherwise have been implicated in human diseases belong to our three novel subfamilies. We found that four of the most active elements in the Brouha et al. study belong to the Ta1d-TCA and Ta1d-CCA subfamilies (this study focused on REF elements, but included LRE3, which is a Ta1d-TCA element).[108] We also determined that the following elements all belong to our three novel subfamilies: a) 39 non-REF elements from the Beck et al. study, b) the L1.3 element from the Dombroski et al study, and c) the three FL-L1Hs elements that gave rise to somatic L1 insertions in a case of colorectal cancer from Scott et al.[7,242,259] In contrast, the well-studied L1RP element (from Schwahn et al.) does not belong to these subfamilies.[260] These data suggest that many (but not all) of the highly active FL-L1Hs source elements that have been studied previously belong to our three novel subfamilies.

Figure 3.3.3 Three novel Ta1d subfamilies of FL-L1Hs elements

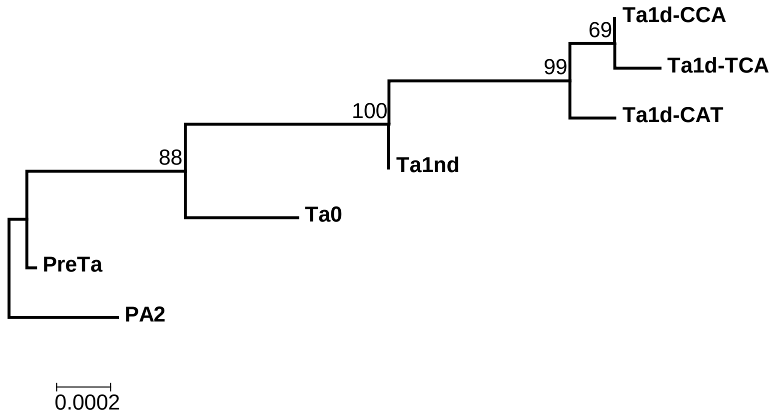
A. Table of canonical positions defining L1 subfamilies expanding on previously published studies (Boissinot et al. 2000, Brouha et al. 2003).[108,261] Positions in red are new canonical positions discovered in our 647 sequenced FL-L1Hs elements. Sequence changes from the ancestral state are highlighted in blue. **B.** A phylogenetic tree was constructed using subfamily consensus sequences to evaluate the relationship of known subfamilies versus our new subfamilies. The tree was calculated using the Neighbor-Joining method and distances were corrected using the Kimura 2-parameter model. The numbers at each node represent the percentage of replicate trees that clustered together in 1,000 bootstrap tests. **C,D.** Reference and non-reference proportions of Ta1d subfamilies that show expansion of the Ta1d-TCA subfamily in the non-reference populations (**D**). Note the expansion from 9% to 32% (dark purple) when comparing the Ta1d-TCA subfamily in Reference (**C**) vs non-reference (**D**). The white numerals in **D** indicate the number of FL-L1Hs elements for that subfamily that were found to be active in the literature.

Figure 3.3.3 Three novel Ta1d subfamilies of FL-L1Hs elements (continued)

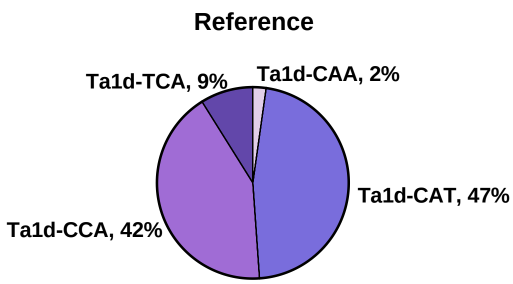
A

	5'UTR			ORF1		ORF2													3'UTR		
	74	353	705	1026	1813	3337	3440	3951	4249	4371	4554	4895	5003	5106	5388	5532	5535	5707	5926	5927	5928
Pa2	G	A	C	C	C	A	A	T	A	C	C	G	C	C	T	G	C	G	G	A	G
Pre-Ta	G	A	C	C	C	A	A	T	A	C	C	G	C	C	T	G	C	G	A	C	G
Ta0	G	A	C	C	C	A	A	T	G	C	T	G	C	C	T	G	C	A	A	C	A
Ta1nd	G	A	C	C	C	A	A	T	G	C	T	A	A	C	C	T	G	A	A	C	A
Ta1d	-	G	T	C/T 67%	T	C/A 57%	A/T 59%	C	G	G	T	A	A	T	C	T	G	A	A	C	A

B



C



D

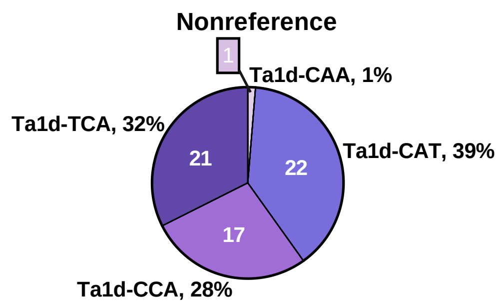


Table 3.3.1 Many of the active FL-L1Hs elements in the literature belong to the three novel subfamilies that we discovered in this study

Examining FL-L1Hs elements in the literature that are known to be active revealed that many belong to our three novel subfamilies. The majority of activity was determined through cell culture retrotransposition assays measured against a reference hot L1, such as L1.3 or L1RP. FL-L1Hs elements listed with zero activity were determined to be active through either 3'transductions or because they caused disease through *de novo* insertions (i.e., Hemophilia A and colorectal cancer). The asterisk denotes that although first described in Dombroski et al. 1993, L1.3 has been measured by cell assay to be a very hot FL-L1Hs element.[259] (Studies listed: Beck et al. 2010, Brouha et al. 2003, Dombroski et al. 1991, Dombroski et al. 1993, Scott et al. 2016, Schwahn et al. 1998)[7,108,242,259,260,262]

Table 3.3.1 Many of the active FL-L1Hs elements in the literature belong to the three novel subfamilies that we discovered in this study (continued)

Study	ID	Activity	Cell Assay Reference	Subfamily
Beck	3-40	0	L1.3	Ta1d-CAT
Beck	5-86	0	L1.3	Ta1d-CAT
Beck	2-7	0	L1.3	Ta1d-CAT
Beck	2-6	0	L1.3	Ta1d-CAT
Beck	3-17	0	L1.3	Ta1d-CAT
Beck	3-15	0	L1.3	Ta1d-CAT
Beck	2-39	0	L1.3	Ta1d-CAT
Beck	2-59	3	L1.3	Ta1d-CAT
Beck	6-113	4	L1.3	Ta1d-CAT
Beck	3-38	6	L1.3	Ta1d-CAT
Beck	4-32	11	L1.3	Ta1d-CAT
Beck	5-58	13	L1.3	Ta1d-CAT
Beck	3-3	15	L1.3	Ta1d-CAT
Beck	5-66	23	L1.3	Ta1d-CAT
Beck	6-97	36	L1.3	Ta1d-CAT
Beck	3-18	62	L1.3	Ta1d-CAT
Beck	4-27	79	L1.3	Ta1d-CAT
Beck	4-8	161	L1.3	Ta1d-CAT
Beck	4-35	191	L1.3	Ta1d-CAT
Beck	3-1	0	L1.3	Ta1d-CCA
Beck	2-24	0	L1.3	Ta1d-CCA
Beck	1-2	0	L1.3	Ta1d-CCA
Beck	5-36	1	L1.3	Ta1d-CCA
Beck	3-39	41	L1.3	Ta1d-CCA
Beck	2-30	46	L1.3	Ta1d-CCA
Beck	4-15	54	L1.3	Ta1d-CCA
Beck	2-53	86	L1.3	Ta1d-CCA
Beck	3-4	98	L1.3	Ta1d-CCA
Beck	6-102	129	L1.3	Ta1d-CCA
Beck	2-21	133	L1.3	Ta1d-CCA
Beck	6-107	139	L1.3	Ta1d-CCA
Beck	3-25	170	L1.3	Ta1d-CCA
Beck	3-34	68	L1.3	Ta1d-TCA
Beck	1-4	107	L1.3	Ta1d-TCA
Beck	2-32	118	L1.3	Ta1d-TCA
Beck	1-(2-1)	118	L1.3	Ta1d-TCA
Beck	3-31	120	L1.3	Ta1d-TCA

Beck	2-12	130	L1.3	Ta1d-TCA
Beck	2-1	141	L1.3	Ta1d-TCA
Brouha	AL137845.9	42.3	L1RP	L1Ambig
Brouha	AL356438.15	73.8	L1RP	L1Ambig
Brouha	AC004200.1	80	L1RP	Ta1d-TCA
Brouha	AC021017.4	89.4	L1RP	Ambig-Ta1d-CCA
Brouha	AC002980.1	132	L1RP	Ta1d-CCA
Brouha	AL512428.13	112.7	L1RP	Ta1nd
Brouha	LRE3	150	L1RP	Ta1d-TCA
Dombroski[1]	L1.2			Ta1d-CAT
Dombroski[2]	L1.3	*		Ta1d-CAT
Dombroski[2]	L1.4			Ta1d-CAT
Scott	Chr17 source element	5 transductions		Ta1d-CCA
Scott	Chr14 source element	12 transductions		Ta1d-TCA
Scott	Chr12 source element	2 transductions		Ta1d-CAT
Schwahn	L1RP	223	L1.3	L1Ambig

3.3.4 FL-L1Hs element internal variation may affect its activity

We next examined CpGs in the promoter regions of our sequenced FL-L1Hs elements and found an inverse relationship between subfamily age and the number of CpGs in the promoters of these elements. In particular, the older subfamilies (i.e., L1-Pre-Ta and L1-Ta0) had fewer CpGs compared to the younger subfamilies (L1-Ta1nd and our novel Ta1d subfamilies). In fact, our three novel subfamilies had the most CpGs, with an average of one or two additional CpGs compared to the older subfamilies (Figure 3.3.4B; Figure 3.3.5). This finding is counter-intuitive and may suggest that the birth of a new, highly active L1 subfamily requires an increased number of CpGs to ensure that these elements become more tightly regulated. These data could also suggest that older elements may be less active from loss of CpGs over time due to spontaneous deamination of methylated cytosines.[176,263]

We identified additional internal sequence changes in our 647 FL-L1Hs elements that could potentially impact L1 function (Figure 3.3.4C,D; Figure 3.3.5). In addition to CpG changes, mutations that impact transcription factor binding sites in L1 promoters or cis-regulatory RNA binding sites also could be envisioned to modulate L1 activity. Likewise, a large number of synonymous and non-synonymous changes have occurred within the ORF1 and ORF2 sequences of these elements (Figure 3.3.4D). Stop codons, frameshift mutations, and non-synonymous changes that introduce dissimilar amino acid substitutions would be expected to influence L1 the greatest. Many of the interior sequence changes are unique to a specific L1 locus and can be used to track source/offspring relationships or the expression of a given FL-L1Hs locus.[7] Together with the Brouha et al. and Beck et al. collections, our collection of sequence-resolved FL-

L1Hs elements will serve as a useful resource to study the impact of internal sequence variation on L1 regulation, activity, and evolution.[108,242]

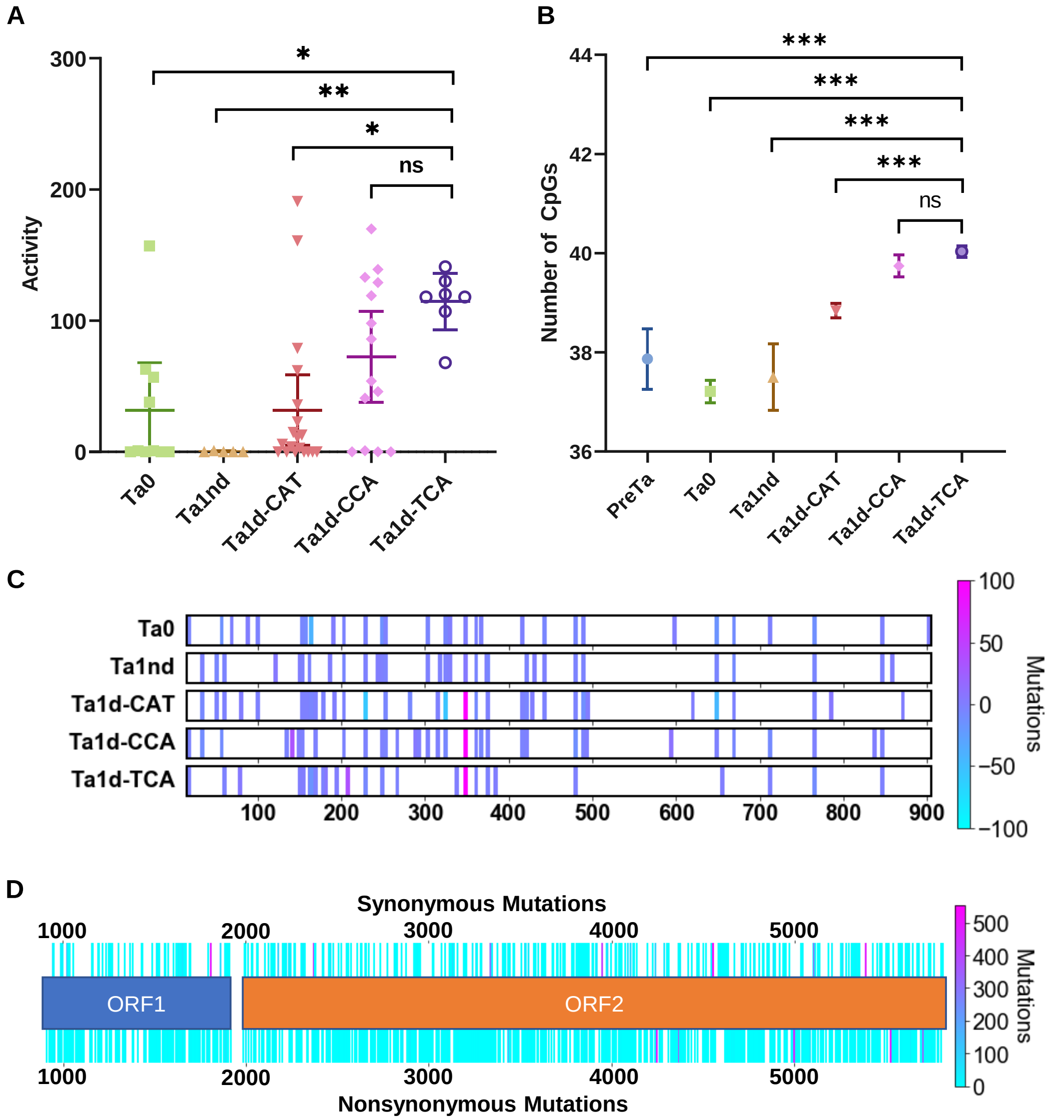
3.3.5 3' Transduction activity in germline and somatic tissues

We also identified 3' transductions that are associated with our sequenced FL-L1Hs elements (Figure 3.3.6). Illumina short read sequencing can capture a small portion of the FL-L1Hs flanks for 3' transduction interrogation. However, with PacBio long read sequencing we are able to capture larger flanking sequences, and thus, detect new 3' transductions that are much larger (Figure 3.3.6). Together with other published 3' transduction data (Gardner et al. 2017, Tubio et al. 2014, Rodriguez-Martin et al. 2020), we determined that 76 (10.9%) of our 698 sequenced FL-L1Hs elements generated 3' transductions.[173,207,258] 31 (40.7%) were active only in the germline, 28 (36.8%) were active only in somatic tissues, and 17 (22.4%) were active in both germline and somatic tissues. Thus, these 3' transductions will likely be useful to further study the expression, regulation, and mobilization of FL-L1Hs elements in various cellular contexts.

Figure 3.3.4 Analysis of new subfamilies, CpGs, and other interior sequence changes in our sequenced FL-L1Hs elements

A. We identified elements that previously had been sequenced and tested in the cell culture retrotransposition assay belonging to various L1 subfamilies (including our three novel subfamilies).[242] The new Ta1d-TCA subfamily has the highest levels of activity followed by the new Ta1d-CCA subfamily. The Ta1d-CAT and Ta0 subfamilies had similar activities. Significance was calculated by one-way ANOVA corrected for multiple comparisons by the Tukey method. Error bars represent 95% confidence intervals. **B.** We also plotted the number of CpGs by subfamily in our sequenced FL-L1Hs elements and found that our three new subfamilies (Ta1d-CAT, Ta1d-CCA, Ta1d-TCA) have 1-2 additional CpGs in their promoter regions compared to older subfamilies. **C.** Mutations in the promoter region causing a gain or loss of CpGs for each L1 subfamily. The color shows the frequency gain or loss as denoted by the diverging color map. **D.** Mutations within the two L1 ORFs. Note the synonymous and non-synonymous mutations above and below the ORF map, respectively. The frequency of the mutations is also shown with diverging color map.

Figure 3.3.4 Analysis of new subfamilies, CpGs, and other interior sequence changes in our sequenced FL-L1Hs elements (continued)



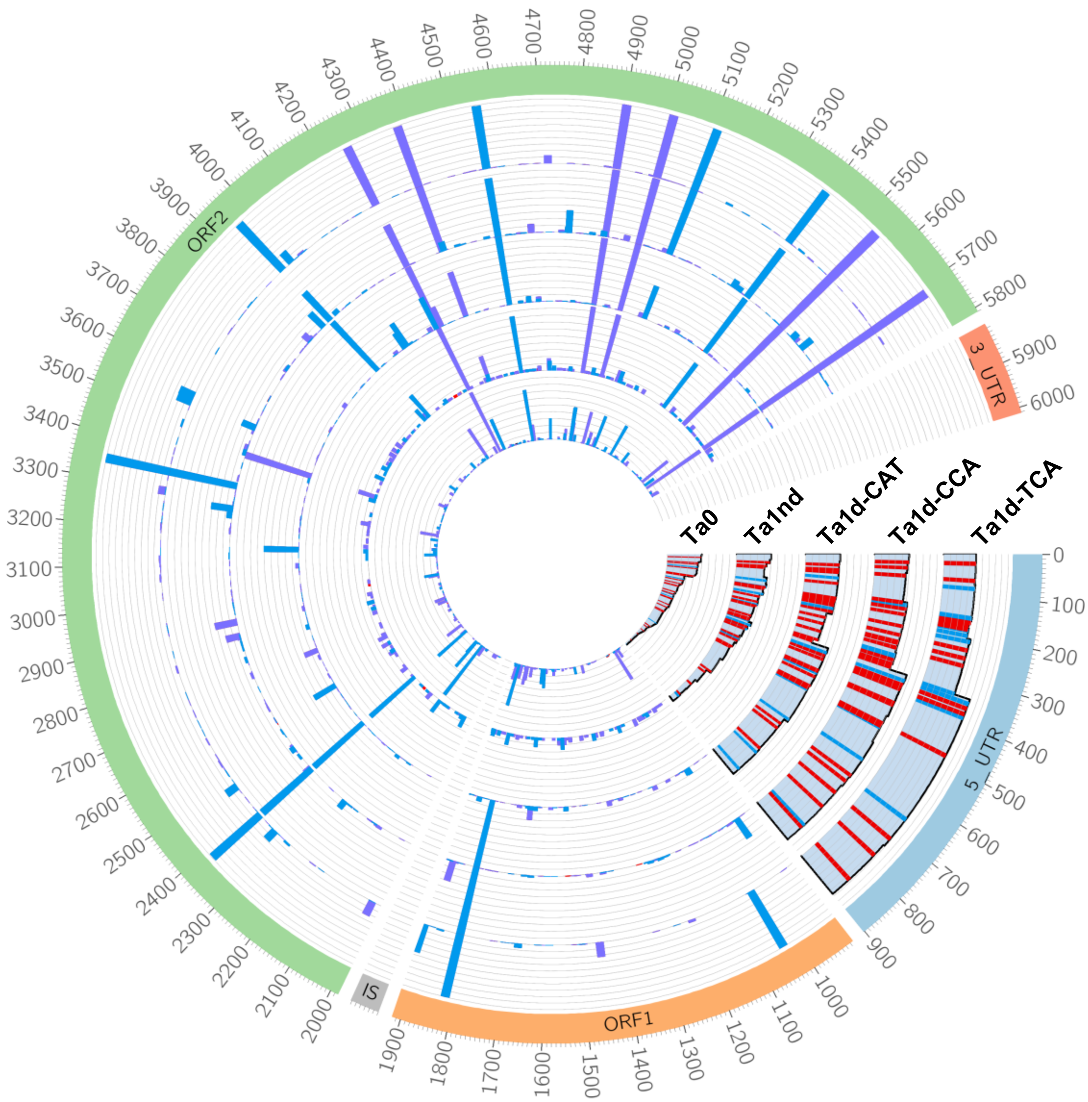


Figure 3.3.5 Circos plot demonstrating the frequency of variation within our collection of sequenced FL-L1Hs elements arranged by subfamily

All subfamilies displayed were aligned to the PreTa subfamily as the reference. The inner track is the oldest subfamily after PreTa, Ta0, and the youngest is Ta1d-TCA on the outermost track. The 5'UTR shows the frequency of CpG changes starting from the reference number of CpGs found in the PreTa consensus sequence. The black line shows the change in the number of CpGs moving downstream with blue bars indicating gain of CpGs and red bars indicating loss of CpGs. Note the decrease in CpGs in the older subfamilies, Ta0 and Ta1nd, and the increases in the younger Ta1d subfamilies. For the ORF1 and ORF2 regions, the fractions of elements with changes at each position are shown as blue bars for synonymous mutations and purple bars for nonsynonymous mutations. Red bars are nonsense mutations.

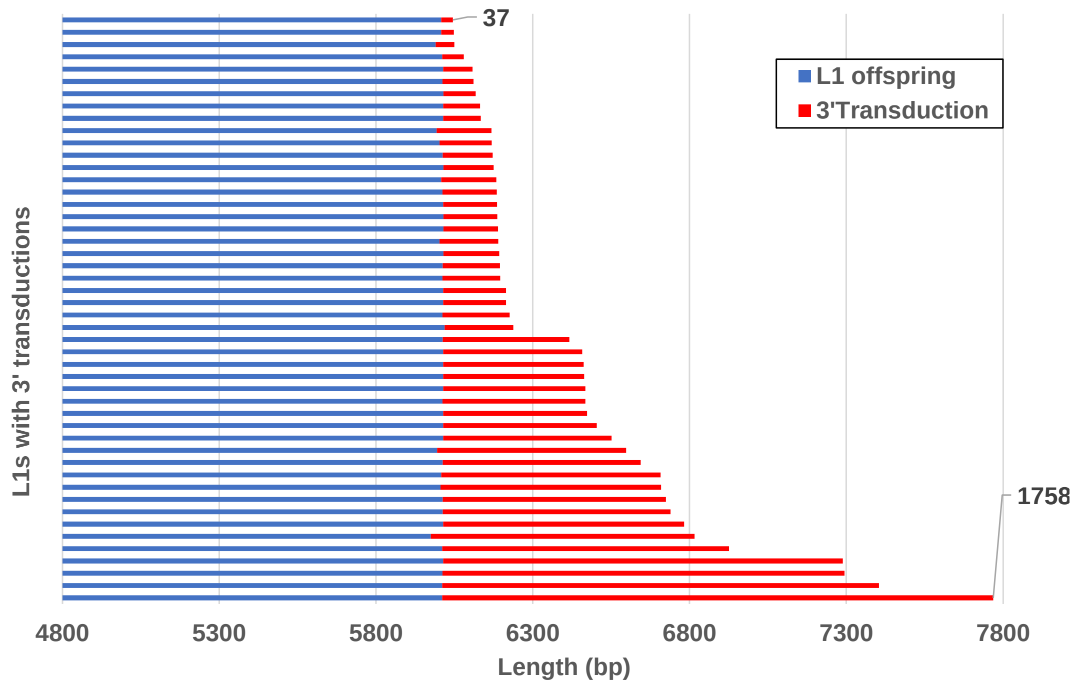


Figure 3.3.6 3’ transductions associated with FL-L1Hs elements

The 3’ ends of the 46 FL-L1Hs elements that are associated with 3’ transductions from our PacBio sequenced elements are depicted (blue) along with the 3’ transduced sequences (red). The full 3’ transduced sequences were determined along with the adjacent FL-L1Hs elements. Transduction size ranged from 37 bp to 1758 bp in length. A total of 91 of our sequenced FL-L1Hs elements have generated MEIs with 3’ transductions.

3.4 Discussion

We discovered 3,728 non-REF FL-L1Hs elements and sequenced 698 of these elements using a novel PacBio-based approach. This provided us with new opportunities to determine whether these elements i) have two intact ORFs, ii) belong to active subfamilies, or iii) belong to new subfamilies that are rapidly expanding in human populations. Since the majority of the non-REF FL-L1Hs elements in our collection have two intact ORFs and are very young (mostly belonging to the three novel Ta1d subfamilies that we identified in this study), we would expect many of these elements to be active in humans. This conclusion is supported by the Beck et al. study, where their collection of non-REF FL-L1Hs elements likewise was enriched for younger, non-REF elements that were highly active in cell culture-based retrotransposition assays.[242]

Our findings demonstrate that the number of non-REF FL-L1Hs copies varies considerably both within and across populations. For example, the number of non-REF FL-L1Hs elements in the 1kGP populations ranged from 25 to 63. The 1kGP individual with the lowest number of non-REF FL-L1Hs elements (i.e., 25) might be expected to have lower levels of MEI mutagenesis than the individual with the highest copy number (i.e., 63; Figure 3.3.1B). However, the individual with copy number 25 could have a large fraction of very hot L1's, whereas the individual with copy number 63 could have mostly inactive copies. Even a single detrimental mutation can severely diminish or eliminate L1 activity.[55,247] Thus, more work is necessary to understand the mutagenic threat that is posed by the full collection of FL-L1Hs elements in an individual's genome.

We discovered three novel L1 subfamilies that have evolved from the highly active L1Ta1d subfamily, namely, Ta1d-CAT, Ta1d-CCA, and Ta1d-TCA (Figure 3.3.3;

Figure 3.3.4). We named these subfamilies based on the three new canonical positions in L1 that define these elements (1026, 3337, and 3440) and the sequences that are found at those three positions (CAT, CCA, TCA). All three of these positions are located within L1 ORFs (1026 is in ORF1, whereas 3337 and 3440 are in ORF2). However, only one of the changes (at position 3440) creates a non-synonymous amino acid substitution (Lysine to Methionine) in the Ta1d-CAT subfamily. This substitution occurs in a critical region of the Z domain known as the octapeptide sequence, which is important for RT activity.[57] Since the change at 3337 is synonymous it should not affect RT function, whereas the change at 3440 causes an amino acid substitution of a larger positively charged, polar amino acid to a smaller neutral, nonpolar amino acid. This could alter protein folding due to the change in size, charge, and polarity, and consequently, the efficiency of retrotransposition.

This reduced RT efficiency may explain why Ta1d-CAT appears to be a relatively active subfamily but has overall lower activity compared to the Ta1d-CCA and Ta1d-TCA (Figure 3.3.4A). In fact, Ta1d-CAT cell culture assay activity is not significantly different from the older subfamily, Ta0 (Figure 3.3.4A). These conclusions are limited by the small sample size of activity data available. Nevertheless, Ta1d-CAT, along with Ta1d-CCA and Ta1d-TCA, share many acquired canonical nucleotide changes that likely contributed to their replicative success (i.e., outcompete older subfamilies for host factors) over time.

Ta1d-CCA may have higher activity simply from not having the 3440 mutation, or it may have increased transcriptional efficiency if there is a codon usage bias.[264] Ta1d-TCA, just like Ta1d-CCA, does not have the 3440 mutation, has the 3337 mutation, and in addition, has the synonymous mutation at 1026 in ORF1. This last mutation is found in the IDR of the N-terminal region of ORF1 where regulatory proteins and kinases interact to affect retrotransposition.[265] Given the high activity observed within the Ta1d-TCA subfamily, it is possible the 1026 mutation confers an advantage in facilitating retrotransposition.

The synonymous changes of positions 1026 and 3337 might be responsible for driving the recent population expansion that we observed for the Ta1d-CCA and Ta1d-TCA subfamily in non-REF populations (Figure 3.3.3D) and the high levels of activity in the cell culture assay (Figure 3.3.4A). These two changes could boost L1 activity, for example, through eliminating host factor binding sites that mediate L1 suppression or by optimizing translation efficiency.[264,266] Additional mechanistic studies will be necessary to determine how sequence changes at these positions impact FL-L1Hs activity.

3.5 Methods

3.5.1 Long-range amplification and sequencing of FL-L1Hs elements

We sequenced the full interior regions of 698 FL-L1Hs elements as follows (Figure 3.5.1). Genomic DNA samples were purchased from Coriell (Coriell Institute for Medical Research, Camden, NJ). PCR primers flanking the FL-L1Hs elements were designed with Primer3 within the 500 bp regions flanking each insertion and purchased from Integrated DNA Technologies (IDT).[267,268] The target size of the primer was 22-26 bases with a melting temperature of 63-68°C. If no primers were found initially, the amount of flanking DNA was increased. In some cases, primers were manually designed to target higher melting temperatures (68-74°C). PCR amplification was performed using LA Taq DNA polymerase (TaKaRa Bio USA, Inc) and a PCR protocol for long-range amplification: 90 sec at 94°C, followed by 32 cycles for 30 sec at 94°C, for 30 sec at 57°C, for 8 min 30 sec at 68°C, with a final elongation step for 10 min at 68°C. This protocol has a relatively low error rate (See next section 3.5.2).

Although amplification was attempted on 789 FL-L1Hs sites, 49 sites could not be amplified even with repeat attempts and 42 sites could not be sequenced (Figure 3.5.2). Each PCR filled site was obtained from a single individual. We required FL-L1Hs candidates for sequencing to be at least 6,000 bp in length as predicted by MELT. Several dozen amplicons were pooled in approximate equimolar concentrations for Pacific Bioscience sequencing with each SMRT cell. Amplicon pooled samples were size selected to filter small fragments. A sequencing library then was prepared and sequenced with the Pacific Bioscience RSII platform. The raw sequencing data were aligned with

BLASR 1.3.1 and assembled with ConsenseTools as part of the SMRT Analysis v2.3.0 software suite (Pacific Bioscience of California, Inc).[269]

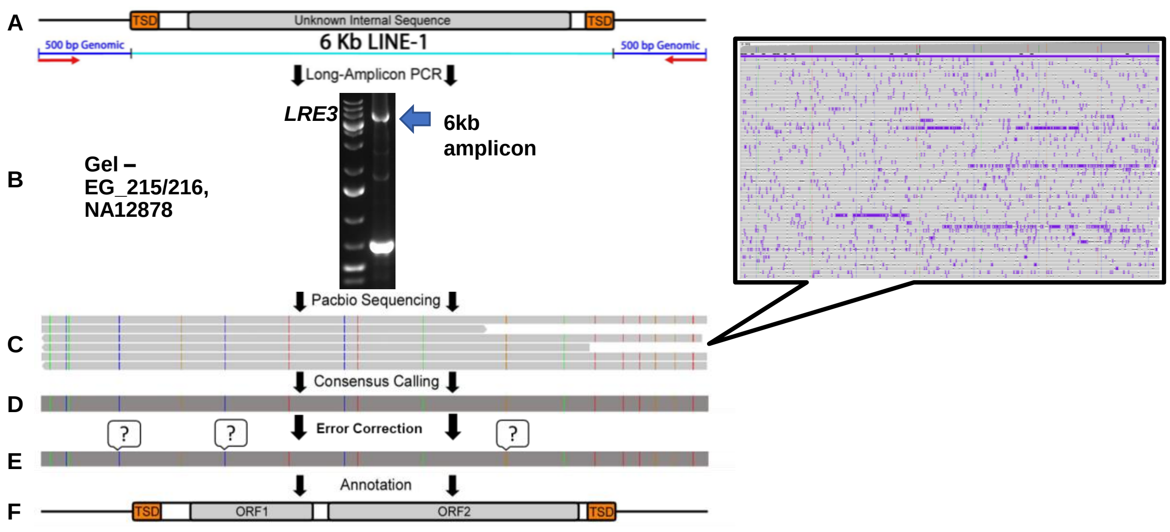


Figure 3.5.1 PacBio amplification and sequencing of FL-L1Hs elements

A. Long-range PCR primers are designed to amplify each target FL-L1Hs element (red arrows). **B.** The initial long PCR for the LRE3 FL-L1Hs element is shown on an agarose gel to determine whether a 6 kb FL-L1Hs fragment was successfully amplified (blue arrow). **C.** Several dozen successful amplicons from unrelated sites are then pooled and the 6 kb FL-L1Hs fragments are eluted from an agarose gel to eliminate the unoccupied allele. The eluted 6 kb pools are then subjected to PacBio sequencing. A sampling of the actual PacBio reads for LRE3 is depicted in the upper right. **D.** The long reads for each FL-L1Hs site are identified using the unique flanking genomic sequences, aligned, and the consensus sequences are determined with the SMRT Analysis v2.3.0 software. Vertical colored lines indicate internal sequence changes within a given FL-L1Hs element compared to the reference FL-L1Hs element sequence. **E.** Systematic error correction is performed on the consensus sequences, and **F**. the final sequence is annotated with a custom analysis pipeline.

Sequenced FL-L1Hs elements were aligned to the L1Ta consensus sequence, and all coordinates are listed relative to this consensus sequence. L1 subfamilies were annotated using previously-reported diagnostic positions (Boisinott et al. 2000, Brouha et al. 2003) along with our newly-defined diagnostic positions (Figure 3.3.3A).[108,261] Ambiguous (Ambig) elements do not perfectly match the diagnostic positions of any known L1 subfamily.

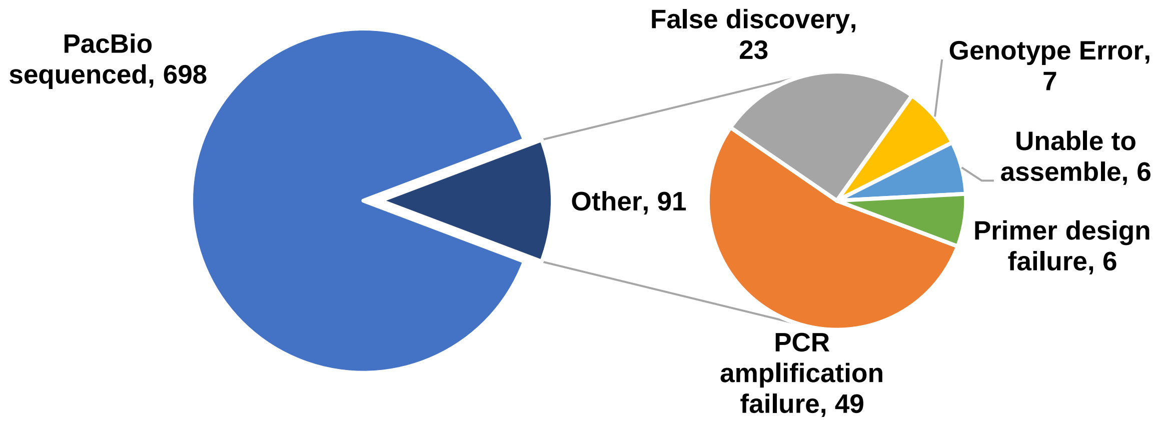


Figure 3.5.2 Number of sites that failed PacBio long read sequencing of 789 attempted

We selected 789 FL-L1Hs elements for long read sequencing. Of those, only 698 sites were successfully sequenced. The 91 sites that were unsuccessful were due to PCR amplification failure (49), false discovery (23), genotype error (7), assembly failure (6), and unable to design a successful primer (6).

3.5.2 Accuracy of long PCR and PacBio sequencing of FL-L1Hs elements.

The type of long PCR that we used from Takara is highly accurate with an error rate of 8.7 x 10-6 due to the 3' to 5' exonuclease (proofreading) activity of the polymerase (TaKaRa). We routinely obtained at least 500 traces for each FL-L1Hs element and any sequencing errors that might occur would be cancelled out upon trace assembly. Likewise, we re-sequenced a sample of FL-L1Hs elements with ABI capillary Sanger sequencing using separate PCR preps, and the two methods were in good agreement (except for a small number of mononucleotide counts within homopolymer tracts, where PacBio makes systematic INDEL errors that we corrected as discussed in the next section below 3.5.3).

3.5.3 Insertion-deletion correction of Pacific Bioscience sequencing errors

Pacific Bioscience sequencing has an estimated INDEL error rate close to 15%.[270] Surveying the collection of FL-L1 sequences, we discovered that there was a high rate of single nucleotide deletion in homopolymer tracts (HPT) with many of them causing frameshift mutations. We used ABI capillary electrophoresis to perform targeted resequencing of 14 randomly selected FL-L1 sites to verify the HPT deletion errors. Comparing these 14 L1 sequences to their respective PacBio sequences we were able to find 19 positions that had spurious HPT deletions. These HPT deletions were also found frequently in the collection of PacBio sequenced sites. Applying correction to these 19 positions removed the sites with high frequency of HPT deletions. The corrections also restored many ORF1 and ORF2 by eliminating frame shifts caused by a missing nucleotide and, subsequently, protein truncation.

3.5.4 Subfamily Analysis

Using the updated Table in Figure 3.3.3A of canonical positions, we sorted the FL-L1Hs elements from our PacBio sequencing. We created consensus sequences for each subfamily and marked positions that did not meet a majority consensus of 60%. SeaView version 4 and Jalview were used to create the ungapped consensus sequences.[271,272] The addition of non-reference elements from other available studies helped achieve consensus for some of the non-majority positions.[242,273] The Ta1d subfamily had two positions that were in non-majority, and when the consensus threshold was increased to 70% there was a third non-majority position. Using these three positions as haplotype markers, we split the Ta1d FL-L1Hs elements into three major haplotypes. Consensus sequences were created for the three new subfamilies and aligned with the other known subfamilies to define new canonical positions (Figure 3.3.3A).

Chapter 4. Discussion and future directions

4.1 Introduction

The field of MEIs is important for understanding the impact of mobile elements on human health and genetic diversity. In this dissertation, I have reviewed the literature and described our own research demonstrating the major role that MEIs have played in human evolution as well as their potential for causing disease. We set out to show that there is still much to be uncovered that large scale sequencing projects will reveal in the coming years. We directed our efforts at studying the engine that drives the mobilization of these MEIs by performing long-range sequencing of FL-L1Hs elements. Altogether, we have described the largest collection of MEIs in diverse human populations as well as the largest collection of fully sequenced FL-L1Hs elements to date. This massive dataset provides a useful resource for the MEI research community to not only answer existing questions but will be an important resource for new questions and new avenues of MEI research.

4.2 Estimating the true extent of MEI association with disease

Despite compiling the largest collection of nonreference MEIs from diverse populations, we are still far from understanding the impact of MEIs on human health. By using this novel collection to cross-reference well established databases such as GENCODE, ENCODE, and OMIM, we demonstrated that there are insertions that may influence human traits and diseases. Insertions in exons, splice junctions, start codons, or stop codons have a straightforward mechanism for disrupting the expression of a normal protein through direct alteration of the coding sequence, transcript, or transcription process. It is likely that even if an insertion is found in these portions of genes, a given insertion may not necessarily be pathogenic or the disease process may not be as simple as a Mendelian disorder such as in Kazazian et al.'s hemophilia patients.[5] Examining our MEIs in these large databases only shows the potential that an isolated insertion can have on disease pathogenesis. Studying these insertions further in a large cohort of whole genome data with associated phenotype data to determine the significance of these insertions would be necessary.

We did attempt to determine the extent MEIs have on human disease and health using the UKBB cohort of almost 50k exomes with corresponding phenotype data. We performed this using the Omics Analysis Search and Information System (OASIS) platform.[274] However, we did not find any statistically significant insertions through association testing. It is likely that most of the insertions in exons are so rare that the study was underpowered to detect an association. If we believe that the study was adequately powered for the more common insertions, then perhaps most germline insertions in exons are highly deleterious and those that remain have little impact on fitness. Given that the exome represents only a small fraction of the genomic space, it is possible there could still be pathogenic insertions not discovered in introns or intergenic regions. Larger studies using whole genome data would be needed to truly understand the extent MEIs impact human health.

4.3 CloudMELT enabling the future of MEI discovery

In order to meet the increased challenges of performing MEI discovery in tens or even thousands of Illumina genomes, CloudMELT allows users, particularly those without access to a large research institution computation cluster, to perform their own large-scale MEI discovery projects. We provided an updated MEI discovery of 1kGP diverse populations using high coverage data and CloudMELT, which increased the number of discovered MEIs three-fold over the 1kGP Phase 3 low coverage data (Table 2.3.3).[153] To our knowledge, CloudMELT is the most up-to-date MEI discovery tool that is specifically engineered and optimized for the cloud. The gnomad-SV caller is a similar cloud-based pipeline; however, it is designed with an older version of MELT without the newest discovery algorithms and optimizations.[210] There are two long read MEI callers, PALMER (Pre-mAsking Long reads for Mobile Element inseRtion) and xTea, that were recently released and used only on a small number of genomes. With the growing number of sequencing projects and availability of human genome sequencing data, MEI discovery will similarly grow. More important will be the increase in availability of long read sequencing data that will allow identification of interior sequence changes to distinguish subfamily, activity, and multiple alleles.[241,275,276] We are also looking to further adapt our rigorously tested MELT and CloudMELT to include long read sequence data compatibility.[277–279] Currently, CloudMELT is only available on AWS, and to increase its access and availability to users, we will need to port the software to the other two large cloud services, Google Cloud and Azure. There is still much of the human reference genome to resolve, and even more so the multitude of different polymorphic MEIs that exist from population to population.[241,275] Cloud based tools, like CloudMELT, will be a necessity for performing MEI discovery on the massive scale we foresee in the near future.

4.4 Tracking subfamily distribution of MEIs

To determine if an MEI belongs to a young subfamily, it is necessary to check the interior nucleotide sequence for diagnostic nucleotides required for subfamily classification. *Alu* elements are relatively short (~280 bp), and we often recover the full interior sequences of these elements with short-read sequencing (Figure 2.3.5). With longer MEs, such as L1, SVA, and HERV-K, that are several kb in length, MELT can only provide limited interior sequence information near the insertion junctions with short-read sequencing data. However, we were able to leverage internal sequences that we recovered with MELT near the 3' end of the SVA insertions to determine the subfamily status for these elements (Figure 2.3.6).

In principle, the SVA and HERV-K sites that we identified might serve as a foundation for future studies that also use long PCR and targeted sequencing to recover more of the interior sequences of these elements. Such studies may allow us to identify additional SVA and HERV-K subfamilies and other useful internal sequence variation. In contrast to the three active classes of mobile elements in humans (i.e., *Alu*, L1, and SVA elements), since HERV-K elements are thought to be extinct, it is unlikely that this would lead to the identification of any functionally intact HERV-K copies.[22] However, our HERV-K (and SVA) sites could be used to study genes located near these sites to determine whether these polymorphic MEIs impact the expression or regulation of these nearby genes. Having the interior sequences of these longer MEIs also would fuel investigations of mechanisms that disrupt gene function (i.e., cryptic splice sites or poly(A) signals) in suspected cases.

4.5 Active FL-L1Hs and three novel subfamilies

We identified three new FL-L1Hs subfamilies with Ta1d-TCA being the emerging dominant lineage, which suggests it likely has greater replicative success in addition to its high activity in comparison to the other two subfamilies. It may seem paradoxical for a highly active FL-L1Hs element to be successful at propagating in the human genome without the risk of a deleterious insertion to the host. It is conceivable that there is a "master" source element that showers the genome with insertions, thus increasing the chance for any given offspring insertion to reduce host fitness.[280] This would naturally subject the source element to negative selection. However, it has been proposed that there could be a "stealth" source element that generates offspring that are intact and possess low level of activity allowing propagation without compromising host fitness.[137] The former model would suggest the higher risk for causing host mutagenesis. Of note, the offspring of either a master or stealth source element can switch to the other model as their activity could be affected by their genomic context or mutations (i.e., a master source element can give rise to a stealth source element and vice-versa). This means that a hot master element can disseminate an offspring to a location that greatly attenuates its activity, thus changing its model to a stealth one. A parsimonious explanation explored by Beck et al. describes a model with similarity to the Red Queen hypothesis, where the hot L1s are generating offspring faster than the host defensive measures can suppress.[242,281] Given there could be upwards of 80-100 active or retrotransposition-competent L1s in the average human genome, the success of an L1 to become the dominant lineage would also hinge on its ability to compete for the necessary resources for its own replication.[33,108]

A limitation of attempting to create prediction models for FL-L1Hs activity would be the presence of mutations within the FL-L1Hs elements and the retrotransposons that they mobilize (i.e., *Alu*, SVA). A single mutation in a given FL-L1Hs element can render the element inactive, and similarly, mutations in non-autonomous retrotransposons could prevent their mobilization as well (e.g., inability for *Alu* to recruit RNA polymerase III or form necessary RNA structure to bind ORF2p).[108,282] Likewise, there are several studies indicating that FL-L1Hs loci can have multiple alleles at the same site with internal sequence changes that influence L1 mobilization. For example, Lutz et al. discovered two alleles of the L1.2 locus (L1.2A and L1.2B) with internal sequence differences that caused a 16-fold change in mobilization in a cell culture-based assay.[283] Likewise, Seleme et al. discovered new alleles of three hot L1 source elements that also impacted L1 mobilization in cell culture assays.[107] Finally, Sanchez-Luque reported allelic variants of a chromosome 13 FL-L1Hs source element where two of the three alleles were inactive in cell culture assays due to an internal stop codon and a missense mutation.[182] Conversely, a mutation could restore activity if it restores a broken ORF or increases activity if it prevents host repression.[108] Thus, additional alleles of our 698 sequenced loci likely exist in human populations, and at least some of these may harbor internal sequence changes that impact retrotransposition activity.

Another factor briefly mentioned before that can influence FL-L1Hs activity is the genomic context in which the element is located. Lavie et al. found that the genomic sequences upstream of individual FL-L1Hs loci can influence L1 promoter activity and can enhance or repress FL-L1Hs expression.[243] Likewise, Philippe et al. found that a relatively small subset of FL-L1Hs loci accounted for the bulk of L1 expression in 12 commonly-used cell lines.[284] Moreover, they found that these expressed loci were differentially regulated across the 12 cell lines. An integrated approach was used to study these loci that included RNA-seq analysis of 3' bypass transcription and 5' antisense promoter transcripts, along with analysis of active H3K4me3 and H3K27ac chromatin marks. Collectively, these data indicate that locus-specific upstream genomic sequences and cell-specific factors contribute to FL-L1Hs regulation.

Our collection of FL-L1Hs elements will be useful for expanding these studies to more broadly examine the impact of diverse genomic contexts on FL-L1Hs activity. As outlined above, the flanking genomic sequences could be examined along with L1 promoter methylation and other chromatin features to better understand how FL-L1Hs element regulation varies across diverse genomic loci.[182,243,284,285] Our PacBio long read targeted sequencing of FL-L1Hs elements captured, on average, 301 bases on the 5'flank and 370 bases on the 3'flank, which could contain evidence of these features. The expression of each FL-L1Hs locus could be studied using one of several possible approaches described by Phillipe et al. and reviewed by Lanciano et al.[284,286] For example, we previously showed that the internal mutation profiles of FL-L1Hs elements can be leveraged to measure the expression of specific FL-L1Hs loci using RNA-seq.[7] The unique internal mutations we identified in our sequenced FL-L1Hs elements (Figure 3.3.4C,D) could be used to differentiate the expressed transcripts in tissues of interest. Similarly, there are 91 of our 698 sequence-resolved FL-L1Hs elements (13%) that have partnered 3' transductions (Figure 3.3.6), which can be used to track activity but also provide another unique signature for sorting transcripts. This collection of fully sequenced FL-L1Hs elements with their flanking sequences could be used to answer future questions regarding the expression, regulation, and mobilization of these elements.[284,286]

4.6 Future Directions

Obtaining a large and comprehensive collection of MEIs is the first step towards studying the impact of MEIs on human health. The majority of MEIs we discovered are rare (78.2% have allele frequency <1%) and the number of MEIs in directly disruptive regions mentioned above are even more rare (Figure 2.3.7A). Unless these rare insertions are highly penetrant or have high effect sizes, the studies to determine association with a trait would require fairly large cohorts to achieve adequate power.[287,288] The standard approach used in Genome Wide Association Studies (GWAS) are single-variant tests, and if an insertion had a MAF of 0.01, the sample size would have to be over 50,000 (80% power and 5% disease prevalence) to detect an odds ratio of 1.4.[287] There are already many large disease-cohorts with high coverage sequencing data under investigation, such as the TOPMed program now with 130,000 WGS available on dbGaP as of this writing.[289] Performing MEI discovery on these massive disease-cohorts will be a next step to understanding the extent MEIs play in human disease.

Large scale assessment of FL-L1Hs activity with either a cell culture-based assay or 3’ transduction detection will assist in identifying the FL-L1Hs elements with mutagenic potential.[7,62] A comprehensive approach would be to perform cell culture retrotransposition assays of all the non-REF elements found in a diverse cohort of individuals, such as the 1kGP cohort. This would provide activity information on a large range of FL-L1Hs elements with varying allele frequencies. Combined with long read sequencing data to obtain subfamily classification, we can also use these activity data to confirm our activity findings of the three novel subfamilies from Beck et al.[242] Furthermore, these data could be used to build FL-L1Hs profiles for each individual to assess their retrotransposon mutagenic potential with increased weight given to pathogenic or hot elements (i.e., Figure 3.3.2 shows possible profiles for Amish and JHS populations). These profiles could be further integrated with REF FL-L1Hs element activity in a given genome. Another goal moving forward will be to develop better approaches to predict which elements are highly active vs. less active. This activity data could also allow development of a model that can be used to predict the activity of each new element or profiles of elements.

References

1. Orgel, L. E. & Crick, F. H. C. Selfish DNA: The ultimate parasite. *Nature* **284**, 604–607 (1980).

2. Fedoroff, N. V. McClintock's challenge in the 21st century. *Proc. Natl. Acad. Sci.* **109**, 20200–20203 (2012).

3. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).

4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

5. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).

6. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).

7. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* gr.201814.115- (2016). doi:10.1101/gr.201814.115

8. Scott, E. & Devine, S. The Role of Somatic L1 Retrotransposition in Human Cancers. *Viruses* **9**, 131 (2017).

9. Kazazian, H. H. & Moran, J. V. Mobile DNA in health and disease. *N. Engl. J. Med.* **377**, 361–370 (2017).

10. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).

11. Goodier, J. L. & Kazazian, H. H. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell* **135**, 23–35 (2008).

12. Mitra, R. *et al.* 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 (2013).

13. Ray, D. A. *et al.* Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res.* **18**, 717–728 (2008).

14. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).

15. Tristem, M. Identification and Characterization of Novel Human Endogenous

Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. *J. Virol.* **74**, 3715–3730 (2000).

16. Marchi, E., Kanapin, A., Magiorkinis, G. & Belshaw, R. Unfixed Endogenous Retroviral Insertions in the Human Population. *J. Virol.* **88**, 9529–9537 (2014).

17. Naveira, H., Bello, X., Abal-Fabeiro, J. L. & Maside, X. Evidence for the persistence of an active endogenous retrovirus (ERVE) in humans. *Genetica* **142**, 451–460 (2014).

18. Barbulescu, M. *et al.* Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **9**, 861–868 (1999).

19. Ross, S. R. MMTV infectious cycle and the contribution of virus-encoded proteins to transformation of mammary tissue. *J. Mammary Gland Biol. Neoplasia* **13**, 299–307 (2008).

20. Belshaw, R. *et al.* Rate of Recombinational Deletion among Human Endogenous Retroviruses. *J. Virol.* **81**, 9437–9442 (2007).

21. Löwer, R., Löwer, J. & Kurth, R. The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 5177–5184 (1996).

22. Wildschutte, J. H. *et al.* Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E2326–E2334 (2016).

23. Hervé, C. A., Lugli, E. B., Brand, A., Griffiths, D. J. & Venables, P. J. W. Autoantibodies to human endogenous retrovirus-K are frequently detected in health and disease and react with multiple epitopes. *Clin. Exp. Immunol.* **128**, 75–82 (2002).

24. Dervan, E., Bhattacharyya, D. D., McAuliffe, J. D., Khan, F. H. & Glynn, S. A. Ancient Adversary – HERV-K (HML-2) in Cancer. *Front. Oncol.* **11**, (2021).

25. Küry, P. *et al.* Human Endogenous Retroviruses in Neurological Diseases. *Trends Mol. Med.* **24**, 379–394 (2018).

26. Greenig, M. HERVs, immunity, and autoimmunity: understanding the connection. *PeerJ* **7**, (2019).

27. Gonzalez-Cao, M. *et al.* Human endogenous retroviruses and cancer. *Cancer Biology and Medicine* **13**, 483–488 (2016).

28. Michaud, H. A. *et al.* Trans-activation, post-transcriptional maturation, and induction of antibodies to HERV-K (HML-2) envelope transmembrane protein in

HIV-1 infection. *Retrovirology* **11**, (2014).

29. Hohn, O., Hanke, K. & Bannert, N. HERV-K(HML-2), the best preserved family of HERVs: endogenization, expression, and implications in health and disease. (2013). doi:10.3389/fonc.2013.00246

30. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annual Review of Genetics* **42**, 709–732 (2008).

31. Smit, A. F. A., Tóth, G., Riggs, A. D. & Jurka, J. Ancestral, Mammalian-wide Subfamilies of LINE-1 Repetitive Sequences. *J. Mol. Biol.* **246**, 401–417 (1995).

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

33. Boissinot, S. & Furano, A. V. Adaptive evolution in LINE-1 retrotransposons. *Mol. Biol. Evol.* **18**, 2186–94 (2001).

34. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10**, 6718–29 (1990).

35. Becker, K. G., Swergold, G. D., Ozato, K. & Thayer, R. E. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum. Mol. Genet.* **2**, 1697–1702 (1993).

36. Minakami, R. *et al.* Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res.* **20**, 3139–45 (1992).

37. Athanikar, J. N., Badge, R. M. & Moran, J. V. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* **32**, 3846–3855 (2004).

38. Yang, N., Zhang, L., Zhang, Y. & Kazazian, H. H. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res.* **31**, 4929–4940 (2003).

39. Tchénio, T., Casella, J. F. & Heidmann, T. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res.* **28**, 411–415 (2000).

40. Wylie, A. *et al.* p53 genes function to restrain mobile elements. *Genes Dev.* **30**, 64–77 (2016).

41. Harris, C. R. *et al.* P53 responsive elements in human retrotransposons. *Oncogene* **28**, 3857–3865 (2009).

42. Sun, X. *et al.* Transcription factor profiling reveals molecular choreography and

key regulators of human retrotransposon expression. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5526–E5535 (2018).

43. Kulpa, D. A. & Moran, J. V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum. Mol. Genet.* **14**, 3237–3248 (2005).

44. Kulpa, D. A. & Moran, J. V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat. Struct. Mol. Biol.* **13**, 655–660 (2006).

45. Hohjoh, H. & Singer, M. F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO Journal* **15**, 630–639 (1996).

46. Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001).

47. Martin, S. L. & Bushman, F. D. Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. *Mol. Cell. Biol.* **21**, 467–475 (2001).

48. Martin, S. L., Branciforte, D., Keller, D. & Bain, D. L. Trimeric structure for an essential protein in L1 retrotransposition. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 13815–13820 (2003).

49. Khazina, E. & Weichenrieder, O. Human LINE-1 retrotransposition requires a metastable coiled coil and a positively charged N-terminus in L1ORF1p. *Elife* **7**, 1–29 (2018).

50. Khazina, E. & Weichenrieder, O. Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 731–736 (2009).

51. Januszyk, K. *et al.* Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. *J. Biol. Chem.* **282**, 24893–24904 (2007).

52. Cook, P. R., Jones, C. E. & Furano, A. V. Phosphorylation of ORF1p is required for L1 retrotransposition. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4298–4303 (2015).

53. Fong, J. H. & Panchenko, A. R. Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. Biosyst.* **6**, 1821–1828 (2010).

54. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002).

55. Feng, Q., Moran, J. V, Kazazian, H. H. & Boeke, J. D. Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. *Cell* **87**, 905–916 (1996).

56. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362 (1990).

57. Clements, A. P. & Singer, M. F. The human LINE-1 reverse transcriptase:effect of deletions outside the common reverse transcriptase domain. *Nucleic Acids Res.* **26**, 3528–35 (1998).

58. Taylor, M. S. *et al.* Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155**, 1034–1048 (2013).

59. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3**, 433–437 (2013).

60. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**, 5899–5910 (2002).

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

62. Flasch, D. A. *et al.* Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication In Brief Article Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* **177**, 837-851.e28 (2019).

63. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357**, 1383–1393 (2006).

64. Gasior, S. L., Roy-Engel, A. M. & Deininger, P. L. ERCC1/XPF limits L1 retrotransposition. *DNA Repair (Amst).* **7**, 983–989 (2008).

65. Suzuki, J. *et al.* Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. *PLoS Genet.* **5**, (2009).

66. Coufal, N. G. *et al.* Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20382–7 (2011).

67. Yamaguchi, K., Kajikawa, M. & Okada, N. LINE retrotransposition and host DNA repair machinery. *Mob. Genet. Elements* **5**, 92–97 (2015).

68. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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 (1993).

69. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. A new retrotransposable human L1 element from the LRE2 locus on chromosome

1q produces a chimaeric insertion. *Nat. Genet.* **7**, 143–8 (1994).

70. Moran, J. V, DeBerardinis, R. J. & Kazazian, H. H. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).

71. Pickeral, O. K., Makałowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–5 (2000).

72. Cordaux, R. & Batzer, M. a. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).

73. Houck, C. M., Rinehart, F. P. & Schmid, C. W. A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* **132**, 289–306 (1979).

74. Ullu, E. & Tschudi, C. Alu sequences are processed 7SL RNA genes. *Nature* **312**, 171–172 (1984).

75. Richard Shen, M., Batzer, M. A. & Deininger, P. L. Evolution of the master Alu gene(s). *J. Mol. Evol.* **33**, 311–320 (1991).

76. Quentin, Y. Fusion of a free left alu monomer and a free right alu monometer at the origin of the alu family in the primate genomes. *Nucleic Acids Res.* **20**, 487–493 (1992).

77. Perez-Stable, C. & Shen, C. K. Competitive and cooperative functioning of the anterior and posterior promoter elements of an Alu family repeat. *Mol. Cell. Biol.* **6**, 2041–2052 (1986).

78. West, N. C., Roy-Engel, A. M., Imataka, H., Sonenberg, N. & Deininger, P. L. Shared protein components of SINE RNPs. *J. Mol. Biol.* **321**, 423–432 (2002).

79. Muddashetty, R. S. *et al.* Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. *J. Mol. Biol.* **321**, 433–445 (2002).

80. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18**, 1875–1883 (2008).

81. Mathias, S. L., Scott, A. F., Jr, H. H. K. & Boeke, J. D. Reverse Transcriptase Encoded by a Human Transposable Element Published by : American Association for the Advancement of Science Stable URL : https://www.jstor.org/stable/2880092 American Association for the Advancement of Science is collaborating with JS. **254**, 1808–1810 (1991).

82. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).

83. Belancio, V. P., Roy-Engel, A. M., Pochampally, R. R. & Deininger, P. Somatic

expression of LINE-1 elements in human tissues. *Nucleic Acids Res.* **38**, 3909–3922 (2010).

84. Deininger, P. L. & Batzer, M. A. Alu Repeats and Human Disease. *Mol. Genet. Metab.* **67**, 183–93 (1999).

85. Belancio, V. P., Deininger, P. L. & Roy-Engel, A. M. Introduction to mammalian transposable elements. *Genome Med.* **1**, 97 (2009).

86. Wang, H. *et al.* SVA elements: A hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005).

87. Ono, M., Kawakami, M. & Takezawa, T. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res.* **15**, 8725–8737 (1987).

88. Zhu, Z. B., Hsieh, S. L., Bentley, D. R., Campbell, R. D. & Volanakis, J. E. A Variable Number of Tandem Repeats Locus within the Human Complement C2 Gene Is Associated with a Retroposon Derived from a Human Endogenous Retrovirus. *J. Exp. Med.* **175**, 1783–1787 (1992).

89. Zabolotneva, A. A. *et al.* Transcriptional regulation of human-specific SVAF1 retrotransposons by cis-regulatory MAST2 sequences. *Gene* **505**, 128–136 (2012).

90. Damert, A. *et al.* 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009).

91. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).

92. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *Am. J. Hum. Genet.* **73**, 1444–1451 (2003).

93. Buzdin, A. *et al.* A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3′ terminus of L1. *Genomics* **80**, 402–406 (2002).

94. Hasnaoui, M., Doucet, A. J., Meziane, O. & Gilbert, N. Ancient repeat sequence derived from U6 snRNA in primate genomes. *Gene* **448**, 139–144 (2009).

95. Buzdin, A. *et al.* The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res.* **31**, 4385–4390 (2003).

96. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. **24**, 363–367 (2000).

97. Vanin, E. F. Processed pseudogenes: characteristics and evolution. *Annu. Rev.*

Genet. **19**, 253–272 (1985).

98. Kazazian, H. H. Processed pseudogene insertions in somatic cells. *Mobile DNA* **5**, 20 (2014).

99. Staszak, K. & Makałowska, I. Cancer, retrogenes, and evolution. *Life* **11**, 1–16 (2021).

100. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).

101. Zhang, Y., Li, S., Abyzov, A. & Gerstein, M. B. Landscape and variation of novel retroduplications in 26 human populations. *PLOS Comput. Biol.* **13**, e1005567 (2017).

102. Feng, X. & Li, H. Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies. *Mol. Biol. Evol.* **38**, 2958–2966 (2021).

103. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.* **7**, e1002384 (2011).

104. Mikkelsen, T. S. *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).

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

106. Sultana, T. *et al.* 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 (2019).

107. Seleme, M. C. *et al.* Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6611–6616 (2006).

108. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5280–5285 (2003).

109. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).

110. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural Genetic Variation Caused by Transposable Elements in Humans. *Genetics* **168**, 933 (2004).

111. Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. Alu repeats: A source for the genesis of primate microsatellites. *Genomics* **29**, 136–144 (1995).

112. Bagshaw, A. T. M. Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.* **9**, 2428–2443 (2017).

113. Sulovari, A. *et al.* Human-specific tandem repeat expansion and differential gene expression during primate evolution Human Genome Structural Variation Consortium. doi:10.1073/pnas.1912175116

114. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ Splice-site selection in Alu exons. *Science (80-.).* **300**, 1288–1291 (2003).

115. Lev-Maor, G. *et al.* Intronic Alus Influence Alternative Splicing. *PLoS Genet.* **4**, e1000204 (2008).

116. Shen, S. *et al.* Widespread establishment and regulatory impact of Alu exons in human genes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2837–2842 (2011).

117. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).

118. Chen, L. L. & Carmichael, G. G. Altered Nuclear Retention of mRNAs Containing Inverted Repeats in Human Embryonic Stem Cells: Functional Role of a Nuclear Noncoding RNA. *Mol. Cell* **35**, 467–478 (2009).

119. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268–74 (2004).

120. Perepelitsa-Belancio, V. & Deininger, P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat. Genet.* **35**, 363–366 (2003).

121. Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21**, 379–386 (2003).

122. Nigumann, P., Redik, K., Mätlik, K. & Speek, M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628–634 (2002).

123. Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, (2020).

124. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of

adjacent cellular genes. *Mol. Cell. Biol.* **21**, 1973–85 (2001).

125. Wheelan, S. J., Aizawa, Y., Han, J. S. & Boeke, J. D. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.* **15**, 1073–1078 (2005).

126. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics* **18**, 71–86 (2017).

127. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. The impact of multiple splice sites in human L1 elements. *Gene* **411**, 38–45 (2008).

128. Taniguchi-Ikeda, M. *et al.* Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature* **478**, 127–131 (2011).

129. Varon, R. *et al.* Partial deficiency of the C-terminal-domain phosphatase of RNA polymerase II is associated with congenital cataracts facial dysmorphism neuropathy syndrome. *Nat. Genet.* **35**, 185–189 (2003).

130. Vorechovsky, I. Transposable elements in disease-associated cryptic exons. *Hum. Genet.* **127**, 135–154 (2010).

131. Sorek, R., Ast, G. & Graur, D. Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060–1067 (2002).

132. Segal, Y. *et al.* LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am. J. Hum. Genet.* **64**, 62–69 (1999).

133. Burwinkel, B. & Kilimann, M. W. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *Journal of Molecular Biology* **277**, 513–517 (1998).

134. Lehrman, M. A. *et al.* Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science (80-.).* **227**, 140–146 (1985).

135. Halling, K. C. *et al.* Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. *Hum. Hered.* **49**, 97–102 (1999).

136. Nyström-Lahti, M. *et al.* Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat. Med.* **1**, 1203–1206 (1995).

137. Han, K. *et al.* Under the genomic radar: The Stealth model of Alu amplification. *Genome Res.* **15**, 655–664 (2005).

138. Callinan, P. A. *et al.* Alu retrotransposition-mediated deletion. *J. Mol. Biol.* **348**, 791–800 (2005).

139. Sinibaldi-Vallebona, P., Lavia, P., Garaci, E. & Spadafora, C. A role for endogenous reverse transcriptase in tumorigenesis and as a target in differentiating cancer therapy. *Genes Chromosomes and Cancer* **45**, 1–10 (2006).

140. Hedges, D. J. & Deininger, P. L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **616**, 46–59 (2007).

141. Morrish, T. A. *et al.* Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**, 208–212 (2007).

142. Belgnaoui, S. M., Gosden, R. G., Semmes, O. J. & Haoudi, A. Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer Cell Int.* **6**, 13 (2006).

143. Morrish, T. a *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31**, 159–65 (2002).

144. Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).

145. van den Hurk, J. A. J. M. *et al.* L1 retrotransposition can occur early in human embryonic development. *Hum. Mol. Genet.* **16**, 1587–1592 (2007).

146. Garcia-Perez, J. L. *et al.* LINE-1 retrotransposition in human embryonic stem cells. *Hum. Mol. Genet.* **16**, 1569–1577 (2007).

147. Branciforte, D. & Martin, S. L. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol. Cell. Biol.* **14**, 2584–2592 (1994).

148. Xing, J. *et al.* Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res.* **19**, (2009).

149. Cordaux, R., Hedges, D. J., Herke, S. W. & Batzer, M. A. Estimating the retrotransposition rate of human Alu elements. *Gene* **373**, 134–137 (2006).

150. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).

151. Papkou, A. *et al.* The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution. *Proc. Natl. Acad. Sci.* **116**, 923–928 (2019).

152. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. *Nat. Genet.* **52**, 179705 (2018).

153. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human

genomes. *Nature* **526**, 75–81 (2015).

154. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, (2011).

155. SR, R. *et al.* Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res.* **27**, 1395–1405 (2017).

156. Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: Variation and disease. *Current Opinion in Genetics and Development* **22**, 191–203 (2012).

157. Goodier, J. L. Restricting retrotransposons: a review. *Mob. DNA* **7**, 16 (2016).

158. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96–99 (2004).

159. Grimson, A. *et al.* Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**, 1193–1197 (2008).

160. Hamdorf, M. *et al.* MiR-128 represses L1 retrotransposition by binding directly to L1 RNA. *Nat. Struct. Mol. Biol.* **22**, 824–831 (2015).

161. Peaston, A. E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).

162. Lucifero, D. *et al.* Coordinate regulation of DNA methyltransferase expression during oogenesis. *BMC Dev. Biol.* **7**, 36 (2007).

163. Goodier, J. L., Zhang, L., Vetter, M. R. & Kazazian, H. H. LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. *Mol. Cell. Biol.* **27**, 6469–6483 (2007).

164. Goodier, J. L., Cheung, L. E. & Kazazian, H. H. Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res.* **41**, 7401–7419 (2013).

165. Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239 (2015).

166. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–10 (2005).

167. Kano, H. *et al.* L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev.* **23**, 1303–1312 (2009).

168. Singer, T., McConnell, M. J., Marchetto, M. C. N., Coufal, N. G. & Gage, F. H. LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes?

Trends Neurosci. **33**, 345–354 (2010).

169. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).

170. Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).

171. Erwin, J. A. *et al.* L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci. 2016 1912* **19**, 1583–1591 (2016).

172. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).

173. Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).

174. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).

175. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2328–38 (2012).

176. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).

177. Rubin, C. M., Vandevoort, C. A., Teplitz, R. L. & Schmid, C. W. Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res.* **22**, 5121–5127 (1994).

178. Xing, J. *et al.* Alu element mutation spectra: Molecular clocks and the effect of DNA methylation. *J. Mol. Biol.* **344**, 675–682 (2004).

179. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227–234 (1997).

180. RA, R. *et al.* Large-scale structure of genomic methylation patterns. **16**, 157–163 (2006).

181. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13994–14001 (2004).

182. Sanchez-Luque, F. J. *et al.* LINE-1 Evasion of Epigenetic Repression in Humans. *Mol. Cell* **75**, 590-604.e12 (2019).

183. Elmer, J. L. *et al.* Genomic properties of variably methylated retrotransposons in

mouse. *Mob. DNA* **12**, 6 (2021).

184. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation [4]. *Nature Genetics* **20**, 116–117 (1998).

185. Kato, Y. *et al.* Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Hum. Mol. Genet.* **16**, 2272–2280 (2007).

186. Min, B., Park, J. S., Jeong, Y. S., Jeon, K. & Kang, Y. K. Dnmt1 binds and represses genomic retroelements via DNA methylation in mouse early embryos. *Nucleic Acids Res.* **48**, 8431–8444 (2020).

187. Yang, F. *et al.* TEX15 associates with MILI and silences transposable elements in male germ cells. *Genes Dev.* **34**, 745–750 (2020).

188. Li, T. H., Spearow, J., Rubin, C. M. & Schmid, C. W. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239**, 367–372 (1999).

189. Kale, S. P., Carmichael, M. C., Harris, K. & Roy-Engel, A. M. The L1 retrotranspositional stimulation by particulate and soluble cadmium exposure is independent of the generation of DNA breaks. in *International Journal of Environmental Research and Public Health* **3**, 121–128 (Int J Environ Res Public Health, 2006).

190. De Cecco, M. *et al.* Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging (Albany. NY).* **5**, 867–83 (2013).

191. De Cecco, M. *et al.* Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. *Aging Cell* **12**, 247–56 (2013).

192. Meter, M. Van *et al.* SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat. Commun.* **5**, (2014).

193. Chen, L., Dahlstrom, J. E., Lee, S. H. & Rangasamy, D. Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. *Epigenetics* **7**, 758–771 (2012).

194. Yang, N. & Kazazian, H. H. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.* **13**, 763–771 (2006).

195. Tristán-Ramos, P. *et al.* The tumor suppressor microRNA let-7 inhibits human LINE-1 retrotransposition. *Nat. Commun.* **11**, 5712 (2020).

196. Hunter, R. G. *et al.* Acute stress and hippocampal histone H3 lysine 9 trimethylation, a retrotransposon silencing response. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17657–17662 (2012).

197. Di Giacomo, M., Comazzetto, S., Sampath, S. C., Sampath, S. C. & O'Carroll, D. G9a co-suppresses LINE1 elements in spermatogonia. *Epigenetics and Chromatin* **7**, (2014).

198. Martens, J. H. A. *et al.* The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* **24**, 800–812 (2005).

199. Kondo, Y. & Issa, J. P. J. Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J. Biol. Chem.* **278**, 27658–27662 (2003).

200. Montoya-Durango, D. E. *et al.* Epigenetic control of mammalian LINE-1 retrotransposon by retinoblastoma proteins. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **665**, 20–28 (2009).

201. Varshney, D. *et al.* SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation. *Nat. Commun.* **6**, (2015).

202. Collins, P. L., Kyle, K. E., Egawa, T., Shinkai, Y. & Oltz, E. M. The histone methyltransferase SETDB1 represses endogenous and exogenous retroviruses in B lymphocytes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8367 (2015).

203. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* **28**, 1397–1409 (2014).

204. Smit, A. F, Hubley, R., Green, P. RepeatMasker Home Page. *Repeat Masker 3.0* Available at: https://www.repeatmasker.org/. (Accessed: 29th March 2021)

205. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

206. Frequently Asked Questions - Genome Reference Consortium. Available at: https://www.ncbi.nlm.nih.gov/grc/help/faq/. (Accessed: 19th February 2022)

207. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).

208. Gardner, E. J. *et al.* Contribution of retrotransposition to developmental disorders. *Nat. Commun.* **10**, (2019).

209. Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261 (2010).

210. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

211. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. (2018). doi:10.1038/nrg.2017.113

212. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

213. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

214. Mell, P. M. & Grance, T. *The NIST definition of cloud computing*. (2011). doi:10.6028/NIST.SP.800-145

215. Post, W. *et al.* Determinants of coronary artery and aortic calcification in the old order Amish. *Circulation* **115**, 717–724 (2007).

216. Wyatt, S. B. *et al.* Prevalence, awareness, treatment, and control of hypertension in the Jackson heart study. *Hypertension* **51**, 650–656 (2008).

217. Sorkin, J. *et al.* Exploring the genetics of longevity in the Old Order Amish. in *Mechanisms of Ageing and Development* **126**, 347–350 (Elsevier, 2005).

218. Campbell, M. C. & Tishkoff, S. A. African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9**, 403–433 (2008).

219. Parra, E. J. *et al.* Estimating African American admixture proportions by use of population- specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).

220. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12**, 745–755 (2011).

221. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annual Review of Medicine* **63**, 35–61 (2012).

222. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology* **35**, 314–316 (2017).

223. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 2021.02.06.430068 (2021). doi:10.1101/2021.02.06.430068

224. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet. 2013 456* **45**, 580–585 (2013).

225. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (80-.).* **369**, 1318–1330 (2020).

226. Cao, X. *et al.* Polymorphic mobile element insertions contribute to gene expression

and alternative splicing in human tissues. *Genome Biol.* **21**, 185 (2020).

227. Wang, J. *et al.* dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation* **27**, 323–329 (2006).

228. Mir, A. A., Philippe, C. & El Cristofari, G. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.* **43**, 43–47 (2014).

229. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

230. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

231. OMIM - Online Mendelian Inheritance in Man. Available at: https://omim.org/. (Accessed: 12th July 2021)

232. Abbott, K. L. *et al.* The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* **43**, D844–D848 (2015).

233. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The Genetic Association Database [1]. *Nature Genetics* **36**, 431–432 (2004).

234. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

235. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. Rare Allele Hypotheses for Complex Diseases. *Curr. Opin. Genet. Dev.* **19**, 212 (2009).

236. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**, 695–701 (2008).

237. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).

238. Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal* **2014**, 2 (2014).

239. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–92 (2013).

240. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, (2006).

241. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (80-.).* **372**, eabf7117 (2021).

242. Beck, C. R. *et al.* LINE-1 Retrotransposition Activity in Human Genomes. *Cell* **141**, 1159–1170 (2010).

243. Lavie, L., Maldener, E., Brouha, B., Meese, E. U. & Mayer, J. The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* **14**, 2253–2260 (2004).

244. Lee, J., Mun, S., Meyer, T. J. & Han, K. High Levels of Sequence Diversity in the 5′ UTRs of Human-Specific L1 Elements. *Comp. Funct. Genomics* **2012**, (2012).

245. Belancio, V. P., Hedges, D. J. & Deininger, P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* **34**, 1512 (2006).

246. Larson, P. A. *et al.* Spliced integrated retrotransposed element (SpIRE) formation in the human genome. *PLOS Biol.* **16**, e2003067 (2018).

247. Moran, J. V *et al.* High Frequency Retrotransposition in Cultured Mammalian Cells. *Cell* **87**, 917–927 (1996).

248. Szak, S. T. *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3**, research0052 (2002).

249. Illumina. *HiSeq X Series of Sequencing Systems Specification Sheet*. (2016).

250. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748 (1996).

251. Casavant, N. C. & Hardies, S. C. The dynamics of murine LINE-1 subfamily amplification. *J. Mol. Biol.* **241**, 390–397 (1994).

252. Cabot, E. L., Angeletti, B., Usdin, K. & Furano, A. V. Rapid evolution of a young L1 (LINE-1) clade in recently speciated Rattus taxa. *J. Mol. Evol.* **45**, 412–423 (1997).

253. Goodier, J. L., Ostertag, E. M., Du, K. & Kazazian, J. A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.* **11**, 1677–1685 (2001).

254. Adey, N. B. *et al.* Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol. Biol. Evol.* **11**, 778–789 (1994).

255. Mears, M. L. & Hutchison, C. A. The evolution of modern lineages of mouse L1 elements. *J. Mol. Evol.* **52**, 51–62 (2001).

256. Deininger, P. L. & Batzer, M. A. Mammalian retroelements. *Genome Res.* **12**, 1455–65 (2002).

257. Ohshima, K. *et al.* Whole-genome screening indicates a possible burst of

formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.* **4**, R74 (2003).

258. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).

259. Dombroski, B. A., Scott, A. F. & Kazazian, H. H. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 6513–6517 (1993).

260. Schwahn, U. *et al.* Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nat. Genet.* **19**, 327–32 (1998).

261. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History. *Mol. Biol. Evol.* **17**, 915–928 (2000).

262. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).

263. Zhou, W., Liang, G., Molloy, P. L. & Jones, P. A. DNA methylation enables transposable element-driven genome expansion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 19359–19366 (2020).

264. Brule, C. E. & Grayhack, E. J. Synonymous Codons: Choose Wisely for Expression. *Trends Genet.* **33**, 283–297 (2017).

265. Furano, A. V. & Cook, P. R. The challenge of ORF1p phosphorylation: Effects on L1 activity and its host. *Mob. Genet. Elements* **6**, e1119927 (2016).

266. Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30**, 308–321 (2014).

267. Untergasser, A. *et al.* Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**, 1–12 (2012).

268. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).

269. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* **13**, (2012).

270. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* **13**, 278–289 (2015).

271. Gouy, M., Guindon, S. & Gascuel, O. SeaView Version 4: A Multiplatform

Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building | Molecular Biology and Evolution | Oxford Academic. in *Molecular Biology and Evolution* **27**, 221–224 (2010).

272. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

273. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).

274. Perry, J. A., Gaynor, B. J., Mitchell, B. D. & O'Connell, J. R. An Omics Analysis Search and Information System (OASIS) for Enabling Biological Discovery in the Old Order Amish. *bioRxiv* 2021.05.02.442370 (2021). doi:10.1101/2021.05.02.442370

275. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).

276. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675.e19 (2019).

277. Zhou, W. *et al.* Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020).

278. McDonald, T. L. *et al.* Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun. 2021 121* **12**, 1–13 (2021).

279. Chu, C. *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* **12**, (2021).

280. Deininger, P. L., Batzer, M. A., Hutchison, C. A. & Edgell, M. H. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311 (1992).

281. Van, V. & Valen, L. Van. A new evolutionary law. (1973).

282. Deininger, P. Alu elements: Know the SINEs. *Genome Biology* **12**, 1–12 (2011).

283. Lutz, S. M., Vincent, B. J., Kazazian Jr., H. H., Batzer, M. A. & Moran, J. V. Allelic Heterogeneity in LINE-1 Retrotransposition Activity. *Am. J. Hum. Genet.* **73**, 1431 (2003).

284. Philippe, C. *et al.* Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife* **5**, (2016).

285. Ewing, A. D. *et al.* Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol. Cell* **80**, 915-928.e5 (2020).

286. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. *Nature Reviews Genetics* **21**, 721–736 (2020).

287. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. *Rare-variant association analysis: Study designs and statistical tests*. *American Journal of Human Genetics* **95**, 5–23 (Elsevier, 2014).

288. Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genet.* **11**, e1005165 (2015).

289. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).