Identification and characterization of polymorphic mobile elements (MEs) in humans

Zakia Dahi, B.Sc. (Honours)

Biological Sciences

(Cell and Molecular Biology)

Submitted in partial fulfillment

of the requirements for the degree of

Master of Science

Faculty of Mathematics and Science, Brock University

St. Catharines, Ontario

Abstract

Retrotransposons are mobile elements (MEs) that propagate in a “copy and paste” fashion in the genomes via RNA intermediates. In the human genome, retrotransposons consist of long terminal repeats (LTRs), long interspersed elements (LINEs), short interspersed elements (SINEs), SINE-VNTR- Alus (SVAs), and processed pseudogenes (PPSGs), and they collectively contribute close to 50% of the genome. Some members of these MEs continue to undergo retrotransposition, thereby generating a type of structural variations (SVs) within and between human populations by the presence and absence of ME insertions at specific genomic locations. A large number of such polymorphic MEs have been previously reported and documented, including cases associated with diseases, but with limited sequence characterization and genotype analysis. In this study, we performed extensive computational analysis and compilation of polymorphic MEs from multiple sources. We focused on characterization of complete sequences representing the insertion alleles and pre-integration alleles of ME polymorphic loci, using methods including local sequence assembly based on rich personal genome sequence data for many entries. Further, we performed *in silico* genotyping and population distribution for these polymorphic MEs for 2600 human subjects representing 28 well recognized populations around the world, as well as phylogenetic analysis of these human subjects using these polymorphic MEs as markers. We identified a total of 4400 polymorphic MEs with full sequence characterization for both the pre-integration and insertion alleles. Among these, 1267 entries represent new insertions not previously documented in the Database of Retrotransposon Insertion Polymorphisms in humans (dbRIP), and 1777 entries represent ME insertions outside the current human reference genome. By individual populations and all samples as whole, all 5 ME types displayed a similar allele distribution pattern with the majority having an allele frequency at 0.5,

while differences across ME types are also seen at the very low frequency range. Nevertheless, polymorphic MEs do show substantial geographic differentiation, with numerous continent-specific loci identified. Polymorphic ME-based clustering of human subjects seems to correlate well with what we know about the history and relationship of human populations, indicating the usefulness of polymorphic MEs as markers for studying human evolution. Furthermore, polymorphic MEs were found to participate in both coding and regulatory sequences, signifying their potential contribution to the phenotypic diversity present among human populations and individuals. In conclusion, polymorphic MEs represent a significant source of human genetic diversity with potentials on impacting the structure, function, and evolution of the human genome.

Acknowledgments

I am very grateful to my supervisor, Dr. Ping Liang, for his endless support and guidance throughout my time in his lab. I will always be appreciative of his patience and encouragement to teach me many fundamental things regarding this subject. I would also like to thank my committee members, Dr. Charles Després and Dr. Adonis Skandalis for all their helpful insights and discussions. I am very appreciative of all my colleagues in the lab, including Daniel Tang, Radesh Nattamai, Aditya Joshi, Jina Nanyakkara, and Ilona Hilson for always being there for me when I needed. Finally, I would like to thank with gratitude, the love and unconditional support from my family and friends.

List of Abbreviations

Base pair (bp)
Basic Local alignment search tool (BLAST)
BLAST-like alignment tool (BLAT)
Coding DNA Sequence (CDS)
Database of Retrotransposon Insertion Polymorphisms (dbRIP)
Database of Single Nucleotide Polymorphisms (dbSNP)
Deoxyribonucleic acid (DNA)
Endogenous Retrovirus (ERV)
Hierarchical clustering (HCL)
Human Endogenous Retrovirus (HERV)
Human specific (HS)
Identical by descent (IBD)
Identical by state (IBS)
Small Insertions and deletions (INDELs)
Insertion-mediated deletion (IMD)
Long Interspersed Nuclear Element (LINE)
Long Terminal Repeat (LTR)
Messenger Ribonucleic acid (mRNA)
Micro Ribonucleic acid (miRNA)
Mobile Elements (MEs)
Open Reading Frame (ORF)
Polyadenylation (poly-A)
Polymerase chain reaction (PCR)
Precursor miRNA (pre-miRNA)
Principle component analysis (PCA)
Processed Pseudogene (PPSG)
Pseudogene (PSG)
Ribonucleic acid (RNA)
RNA interference (RNAi)
RNA-Sequencing (RNA-seq)
Sequence read archive (SRA)
Short Interspersed Nuclear Element (SINE)
SINE-VNTR-Alu (SVA)
Single nucleotide polymorphisms (SNPs)
Structural variants (SVs)
Target site duplication (TSD)
Target-site-primed-reverse transcription (TRPT)
Transcription factor (TF)
Transcription factor binding site (TFBS)
Transfer RNA (tRNA)
University of California at Santa Cruz (UCSC)
Untranslated Region (UTR)

Contents

Chapter 1: Introduction

1.1 Mobile elements in the human genome

The biological importance of “jumping genes”, also known as mobile elements (MEs) began with its discovery in the 1950s and has continuously been applied to the study of genome evolution (McClintock, 1950). As DNA sequences that are capable of moving or propagating themselves and integrating into new sites in the genome, MEs constitute a large portion in many eukaryotic genomes (Ayarpadikannan & Kim, 2014). Recent reports based on the current reference genome sequences indicate that MEs account for approximately 48% of the human genome (Tang et al., 2018), a revision higher than previously reported (Cordaux & Batzer, 2009; Deininger et al., 2003; Lander et al., 2001).

Depending on their method of transposition, MEs can be grouped into two main classes known as DNA transposons and retrotransposons. DNA transposons are MEs which achieve mobility by a cut-and-paste mechanism through its self-encoded transposase enzyme (Kazazian & Moran, 2017). This class of MEs are considered inactive and may no longer transpose in humans, and its percentage in the human genome was stated to be approximately 3.5% (Tang et al., 2018). Therefore, in the human genome, MEs are primarily represented by retrotransposons, which constitute approximately 44.5% of the genome, and this is more or less the same in other primate genomes (Tang et al., 2018). Retrotransposons mobilize by a copy-and-paste mechanism where MEs are first transcribed into RNA and then the RNA intermediates are reverse-transcribed into DNA sequences, which are then integrated into new genomic sites (Kazazian & Moran, 2017). This class of MEs can be further divided into two categories based on the presence or absence of long terminal repeats (LTRs): LTR and non-LTR retrotransposons. The group of LTR retrotransposons comprises endogenous retroviruses (ERVs), whereas non-LTR retrotransposons consists of long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and SINE-VNTR-Alus (SVAs). Together, these three types of non-LTR retrotransposons represent the majority of MEs present in the human genome, with a combined percentage of 35.4% (Tang et al., 2018).

In humans, LTR retrotransposons are mainly represented by human ERVs (HERVs) and account for approximately 9.1% of the human genome (Bannert & Kurth, 2004). Their structural features include LTRs that are separated by sequences analogous to the *gag*, *pol*, and *env* genes of retroviruses. These retroviral genes code for products that are essential to generate the viral proteins, which are also required for their retrotransposition (Finnegan, 2012).

The family of LINEs are one of the most abundant and successful MEs in humans by total length, constituting about 22% of the genome, as well as by number with more than one million copies all together (Tang et al., 2018). Among these, LINE-1 (L1) is the only currently active class of autonomous retrotransposons. A full-length L1 element is approximately 6 kilobases (kb), with unique features including a 5' untranslated region (UTR) containing an internal RNA polymerase II promoter, two opening readings frames (ORF1 and ORF2), and a 3' UTR comprising a polyadenylation (poly-A) signal (Ayarpadikannan & Kim, 2014). ORF1 encodes an RNA-binding protein, while ORF2 produces a protein with reverse transcriptase and endonuclease activities, and together these ORF proteins are responsible for the process of L1-mediated retrotransposition (Elbarbary, Lucas, & Maquat, 2016).

SINEs are considered one of the most successful MEs in the human genome by copy number, totaling more than 1.5 million copies and contributing to 13.4% in the genome (Tang et al., 2018). This group of retrotransposons are mainly represented by Alu elements, a size of approximately 300 bp in length (Kazazian & Moran, 2017). The structure of Alu elements consist of two monomers containing an internal polymerase III and a poly-A tail of variable length. Unlike L1 elements, the inability to encode proteins labels SINEs as non-autonomous elements. As a result, their retrotransposition depends on the proteins encoded by L1 elements (Kazazian & Moran, 2017).

SVAs have emerged as the youngest group of retrotransposons in humans and other *Hominidae* primates, accounting for approximately 0.1% of the genome (Hancks & Kazazian, 2010; Tang et al., 2018). A full-length SVA element is about 2 kb in length and its sequence is constructed of various segments consisting of a $(CCCTCT)_n$ hexamer repeat region at the 5' end, an Alu-like region, a variable number of tandem repeats (VNTR) region, a SINE element of retroviral origin (SINE-R) region, followed by a poly-A signal and poly-A tail. SVA elements are also non-autonomous and depend on L1 elements for retrotransposition (Hancks & Kazazian, 2010; Raiz et al., 2012).

Other MEs found in the human genome also include processed pseudogenes (PPSGs), the most common type of pseudogenes (PSGs), and they account for less than 1% of the human genome. Due to their means of derivation, PPSGs are also termed as retrotransposons or retrogenes and may be classified into the group of non-LTR retrotransposons (Zhang, Harrison, & Gerstein, 2002). The progression of PPSGs is dependent on the reverse transcription of processed mature mRNAs and the incorporation of the resulting complementary DNA sequence back into the genome (Mighell et al., 2000). As a result, PPSGs contain features of mRNA and can be differentiated from other types of PSGs by having a complete lack of introns and 5' regulatory elements, but possessing 3' poly-A tails and flanking direct repeats (Li, Yang, & Wang, 2013). PPSGs can be classified with Alu and SVAs as non-autonomous elements (Kazazian & Moran, 2017).

1.2 Mechanism of LINE-1 Retrotransposition

The mechanism of L1 retrotransposition mediates the integration of non-LTR transcripts back into the human genome. This process begins in the nucleus where a L1 DNA is transcribed from a promoter located within its 5’ UTR (Kazazian & Moran, 2017). The transcribed L1 RNA is exported into the cytoplasm where it undergoes translation and encodes ORF1 and ORF2 proteins. The ORF proteins identify and bind to non-LTR RNA transcripts and form a cytoplasmic complex. This complex is transported from the cytoplasm to the nucleus where a process known as target-site-primed reverse transcription (TPRT) occurs and begins with the activity of the ORF2 protein (Luan et al., 1993). The enzymatic action of endonuclease encoded by ORF2 nicks one strand of the target DNA, generating an exposed 3’ hydroxyl that serves as a primer for reverse transcription of the non-LTR RNA. Detailed analysis of human L1 retrotransposons have shown that it preferentially cleaves target sites containing A-T junction sequences (5’-TTTT/AA-3’) (Fujiwara, 2015). Following target site cleavage, by using the poly-A tail for binding to the above overhanging sequence as the primer, the reverse transcriptase enzyme, also encoded by ORF2, copies a non-LTR retrotransposon transcript into double stranded RNA/DNA hybrid. The RNA is then degraded, while the complementary DNA strand is then inserted into the chromosome via a second nick on the other strand near and most often slightly downstream of the first nick site. The synthesis of the complementary strand is thought to be done by the DNA repair mechanisms (Finnegan, 2012; Kazazian, 2004; Ostertag & Kazazian, 2001). As a result of TPRT and the nature of the staggered cleavage sites, L1-mediated retrotransposition incorporates unique characteristics of canonical non-LTR retrotransposons, which include frequent 5’ truncations, presence of a poly-A rich tail at the 3’ end, and 2-20 base pair (bp) long target site duplications (TSDs) (Cordaux & Batzer, 2009).

TSDs, as a hallmark of DNA transposition, are a pair of short repeats of sequence, and as part of the integration site are used to define the boundaries of an insertion (Hancks & Kazazian, 2012).

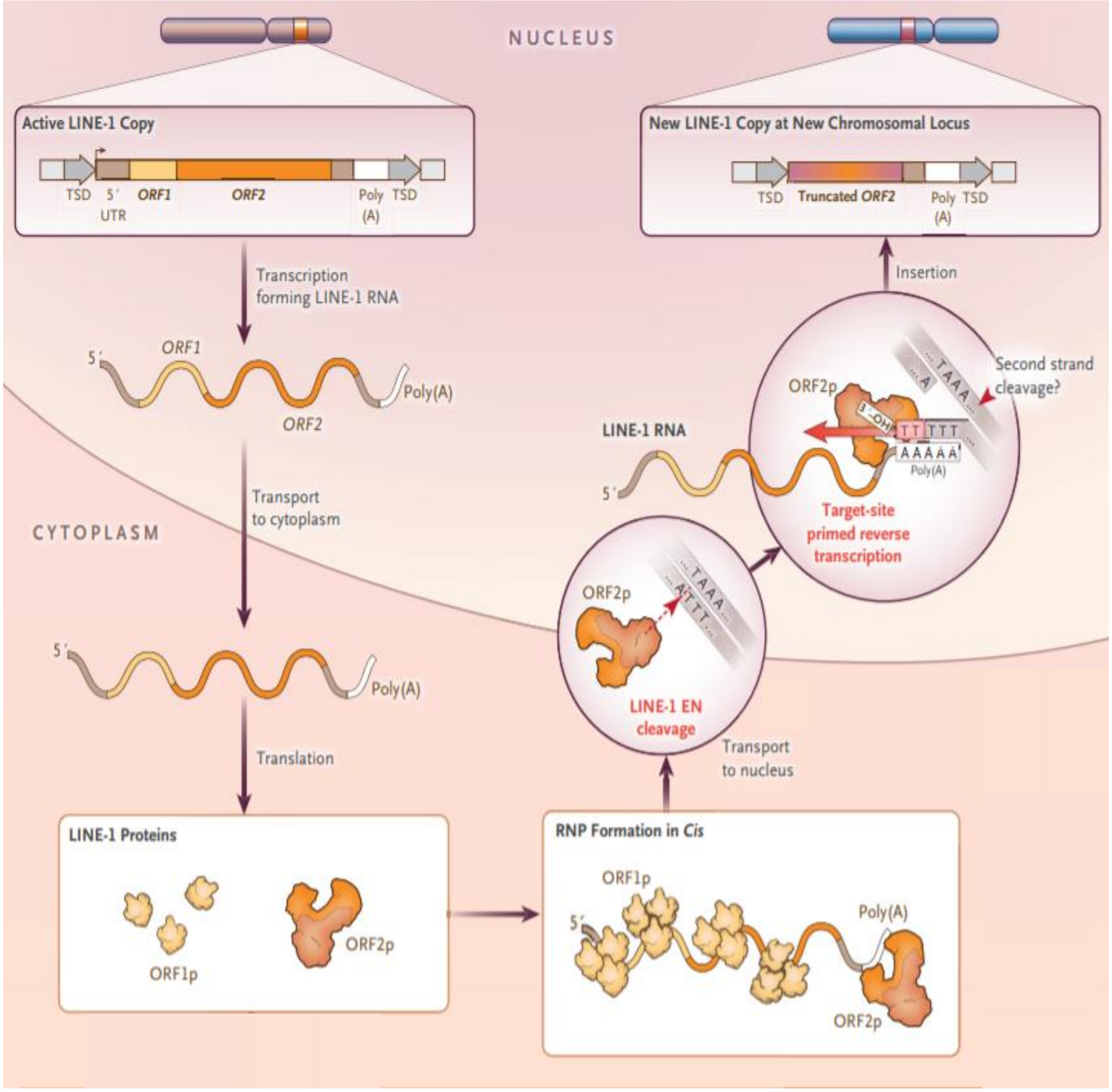

Figure 1. Mechanism of LINE-1 retrotransposition. This figure was adapted from "Mobile DNA in Health and Disease" by Haig H. Kazazian, Jr., and John V. Moran (Kazazian & Moran, 2017). Schematic model illustrating the mechanism of L1-mediated retrotransposition, a process also termed as target-site-primed reverse transcription (TPRT).

Understanding the mechanism of L1 retrotransposition has led to the knowledge of how other non-LTR retrotransposons use L1 machinery for their mobilization in the human genome. The process of L1 retrotransposition is critical to mediate the integration of non-autonomous MEs including Alus, SVAs, and PPSGs. Nevertheless, the full extent of how MEs identify a genomic location for integration remains to be clarified. Previous reports state that it is highly probable that chromatin states, importantly open chromatin, in addition to L1 endonuclease activity, may dictate target-site preference (Bourque et al., 2018; Flasch et al., 2019).

In contrast, LTR retrotransposition has a more complicated mechanism. The process of reverse transcription of LTR retrotransposons begins with the function of a cellular tRNA molecule that is used as a primer for the copying of the 5' of the ERV RNA into double stranded complementary DNA. Following the completion of the second-strand synthesis, integrase mediates the insertion of the DNA into the host chromosome at random locations (Sverdlov & Wiley, 2000).

1.3 Impact of MEs on human genome evolution and gene function

Genome evolution has been largely driven and shaped by MEs through a variety of mechanisms. Recent and ongoing events of retrotransposition contribute to be a major source of structural genetic diversity within and between individuals (Tang et al., 2018). One of the most intuitive outcomes of retrotransposon insertions is their contribution to increasing genome size. It has been found that the collection of all human-specific MEs contribute to approximately 14.2 million bp of sequence length to the human genome (Tang et al., 2018). Among ME types, L1s made the largest net genome size increase, followed by Alus, SVAs, LTRs, and PPSGs (Tang et al., 2018). For many years, numerous studies have focused on the activities of MEs and their relative retrotransposition levels in the human genome. Retrotransposition events mediated by L1 elements are estimated to occur minimally in 1 of 20 meioses for Alu, 1 of 20-200 meioses for L1, and 1 of 900 meioses for SVA (Kazazian & Moran, 2017; Tang et al., 2018). ME insertions were previously thought to be limited to the germline, but recent studies have found evidence of MEs in somatic tissues as well (Platt, Vandewege, & Ray, 2018). Such mobilization of MEs in the human genome can result in severe consequences, ranging from cancer to the dysfunction of aging cells (Ewing, 2017). Overall, the continuous accumulation and activity of retrotransposons in the human genome is an ongoing process which signifies their importance and impact during genome evolution (Cordaux & Batzer, 2009).

On the small scale, retrotransposons can impact genome evolution through several evolutionary and adaptive processes (Lonnig & Saedler, 2002; Solyom & Kazazian, 2012). These associations may range from generating genomic instability or genome rearrangements such as deletions, duplications, inversions, and transductions, in which ME insertions create structural variation in the genome (Konkel & Batzer, 2010). ME insertions have been observed to cause insertion-mediated deletions (IMDs), where the event of the insertion is associated with the deletion of adjacent genomic sequence (Belancio, Hedges, & Deininger, 2008). Both Alu and L1 retrotransposition-mediated deletions can lead to large-scale removal of genomic sequences, with reported cases of deletions spanning from 1 bp to over 130,000 bp (Callinan et al., 2005; Cordaux & Batzer, 2009). Deletions of genomic DNA in genic regions have the potential to be associated with disease. A large deletion of sequence involving the *PDHX* gene by a L1 element has been reported to result in pyruvate dehydrogenase deficiency, affecting the *NF1* gene, causing neurofibromatosis (Mine et al., 2007; Wimmer et al., 2011). Other well-known examples of this phenomenon include Alu-mediated deletions in the low-density lipoprotein (LDL) receptor locus resulting in familial hypercholestrolemia (Hoffmann et al., 2015).

In addition to IMDs, the event of ME retrotransposition can transduce additional flanking genomic sequences, either upstream or downstream of the insertion (Goodier, Ostertag, & Kazazian, 2000; Moran, DeBerardinis, & Kazazian, 1999; Pickeral et al., 2000). In the case of 3' transduction, the RNA processing machinery may skip the weak L1 poly-A signal and instead use a second polyadenylation site located downstream in the 3' flanking sequence. In the event of 5' transduction, an external promoter located upstream in the 5' flanking sequence of the ME insertion can be used to drive its transcription (Cordaux & Batzer, 2009).

1.3.1 Impact of MEs in regulating gene expression

Several studies have revealed the role of MEs in the regulatory regions. MEs integrated upstream of protein-coding genes may provide recognizable binding sites for transcription factors (TFs) and function as potential promoters. SINE-embedded TF binding sites (TFBS) have been identified to modulate gene transcription either in a positive or negative manner (Chuong, Elde, & Feschotte, 2017; Elbarbary et al., 2016). Additionally, gene regulation is influenced by ME-derived alternative promoters. Internal promoters carried by MEs can impact the expression of nearby genes, which can interfere with the biological functions of the gene product. ME-mediated activation of promoters can drive inappropriate expression of genes, with the potential for disease and cancer association (Anwar, Wulaningsih, & Lehmann, 2017). More specifically, there is evidence that ME insertions into UTRs may function as a promoter for the downstream gene, thereby generating an alternative transcript (de Souza, Franchini, & Rubinstein, 2013). Various studies have identified the involvement of MEs in reshaping the human transcriptional landscape by providing insights on the contribution of ME-derived sequences to novel regulatory elements.

1.3.2 Contribution of MEs to protein-coding regions in the human genome

Aside to genomic alterations, MEs have shaped genomes by providing sequences for a number of protein-coding regions of genes (Finnegan, 2012). In some cases, MEs influence genetic diversity by creating genes with novel functions inserted within human genomes. The predominant mechanism of MEs' ability to interrupt genes is through insertional mutagenesis that can be coupled with other processes including genomic transductions, deletions, or aberrant splicing (Solyom & Kazazian, 2012). Detailed analysis of retrotransposons has also given rise to identifying population-specific diseases. For instance, a homozygous Alu insertion in the male germ-cell-associated kinase (*MAK*) gene was identified in patients of Jewish ancestry who were diagnosed with retinitis pigmentosa (Tucker et al., 2011). Another population-specific disease was discovered to be caused by an ancestral insertion of an SVA element in Japan and Northeast Asian populations, leading to Fukuyama-type congenital muscular dystrophy (Kobayashi et al., 1998; Watanabe et al., 2005). However, it is important to note that not all insertional events have a negative effect in the genome. For instance, an Alu insertion polymorphism in the angiotensin I converting enzyme (*ACE)* has been associated with protection from the dry and atrophic form of age-related macular degeneration (Hamdi et al., 2002). Taken together, the integration of MEs in the genome can lead to genetic differences between human individuals, which in turn can lead to phenotypic differences as a result of gene function and expression differences.

1.3.3 Impact of MEs on splicing mechanisms

Alternative splicing is a mechanism where exons are extracted in different ways to generate variations of mRNA transcripts, many of which lead to protein diversity (Cowley & Oakey, 2013). Depending on the site of insertion, MEs can cause various forms of mutations that result in either beneficial or deleterious effects on the host genome. When a ME inserts within an exon, it may change the ORF, coding for an abnormal peptide or causing missense/nonsense mutations (Lee, Ayarpadikannan, & Kim, 2015). This was exemplified by the alternative splicing of the *fukutin* gene induced by the insertion of a SVA element, generating a protein that was mislocalized from the Golgi to the endoplasmic reticulum (Kobayashi et al., 1998; Taniguchi-Ikeda et al., 2011). In contrast, ME insertions into intronic regions can result in alternative splicing or cause mRNA destabilization, thereby reducing gene expression (Ayarpadikannan et al., 2015).

Among the various mechanisms associated with alternative splicing, an exon can be lost from one of the mRNA variants, which is an event defined as exon skipping (Ayarpadikannan et al., 2015). Similarly, a process known as exonization, where intron sequences are retained as exons, may occur as a consequence of alternative splicing. ME insertions are known to incorporate 5’ or 3’ splicing sites that cause a full or partial ME to be retained as an exon (Abascal, Tress, & Valencia, 2015). Such changes following a splicing process may result in the production of abnormal proteins, causing alternations in gene expression and therefore phenotypic differences between individuals (Ayarpadikannan et al., 2015).

MEs, mainly non-LTR retrotransposons, are capable of introducing novel splice sites following integration in the genome (Cowley & Oakey, 2013). When introduced into a gene, Alu elements are known to provide weak acceptor and donor splice sites, with the potential of accumulating mutations over time that can activate these sites and enable the creation of new exons (Deininger, 2011). A study in 2007 confirmed that the donor splice site of one of the exons of the s*urvivin* gene, belonging to family of apoptosis inhibitors, was Alu-derived (Ayarpadikannan et al., 2015; Mola et al., 2007). Overall, Alu elements contribute to a significant portion of alternatively spliced exons (Cowley & Oakey, 2013). More specifically, Alus are responsible for most of the non-constitutively expressed alternative splicing variants in the human genome, suggesting the relevance of these MEs to transcriptome differences (Hasler & Strub, 2006).

1.3.4 Impact of MEs in post-transcriptional regulation

RNA interference (RNAi) pathways are suggested to be key to ME evolution. The RNAi process uses short RNA sequences to recognize and prevent the annealing of complementary nucleic acids. Therefore, this process functions to regulate gene expression and epigenetic modifications (Obbard et al., 2009). MiRNAs are defined as one of the classes of RNAi and are made up of short non-coding RNA sequences that are approximately 22 bp in length (Qin et al., 2015). They are initially transcribed in the nucleus and are processed to form short stem-loops, known as precursor-miRNAs (pre-miRNA). The pre-miRNA then becomes cleaved to form double-stranded miRNA sequences (Qin et al., 2015).

In humans, miRNAs regulate gene expression by targeting complementary mRNA-specific regions, known as miRNA-target sites (Qin et al., 2015). These miRNA-target sites are primarily located in the 3’ UTR regions (Qin et al., 2015). Recent reports claim that MEs are responsible for the origin of new miRNAs and miRNA-target sites. For instance, LINE and SINE elements have been found to activate the function of miRNAs by acting as promoters for miRNA synthesis or as miRNA-binding sites in miRNA-target regions (Elbarbary et al., 2016). More specifically, mir-28, mir-95, and mir-151 are functionally validated LINE element-derived miRNAs (Spengler, Oakley, & Davidson, 2014). Other findings have also revealed that MEs can generate miRNA target sites in the 3’ UTR of protein-coding transcripts. For instance, it was demonstrated that Alu elements embedded in the 3’ UTR regions of *EIF2S3* and *MAP3K9* genes were the source of miR-24 and miR-122 target sites (Spengler et al., 2014).

In addition, MEs have the potential to serve as miRNA decoys to inhibit the function of miRNAs on their real targets through competition. As single-stranded non-coding RNAs, miRNAs target the 3’ UTR of mRNAs initiating mRNA silencing (Li et al., 2013). PPSGs approach this action by sharing sequence similarities with miRNA target sites (3’ UTR of the parent gene transcripts) and competing to bind to the miRNA to further inhibit its function (Li et al., 2013). An example of this is seen in *PTENP1*, which is a PPSG derived from a tumour suppressor gene with sequence resembling its parental gene *PTEN* (Muro, Mah, & Andrade-Navarro, 2011). *PTENP1* contains a 3’ UTR that serves as the miRNA target site and functions as the decoy by disrupting the interaction between the miRNA and its real target site (Li et al., 2013).

MiRNAs can also function to suppress the expression of MEs. In somatic cells, the miRNA induced silencing complex (miRISC) is directed by miR-128 to bind directly to a target site residing in the ORF2 RNA of a L1 element (Pedersen & Zisoulis, 2016). This interaction results in the destabilization of the L1 transcript, leading to the translational suppression of the L1 proteins (Szitenberg et al., 2016). Overall, the miRNA pathway is an essential component of post-transcriptional control of gene expression, in which MEs make significant contribution (Pedersen & Zisoulis, 2016).

1.3.5 Mechanisms of regulating ME expression

Networks of diverse mechanisms have evolved in humans to regulate the expression and activity of MEs. DNA methylation patterns and chromatin structures appear to be recognized as some of the key developed processes to control ME expression in the human genome (Lee et al., 2015). ME expression is regulated by chromatin-modifying agents including histone alterations and modifications in chromatin packaging and condensation (Macia et al., 2011). A histone molecule contains N- and C-terminal tails that protrude from the body of the nucleosome, which is the structural unit of the packaged DNA (Strachan & Read, 2011). Precluding histone tail modifications are known to stimulate heterochromatin formation, which is where nucleosomes are densely packed, thereby altering the binding of protein factors and levels of methylation (Huda, Marino-Ramirez, & Jordan, 2010). MEs are involved in instigating this behaviour of repressive histone modification. An example of this was seen following an L1 insertion in teratorcarcinoma cell lines where the insertion resulted in the deacetylation of histone tails following the recruitment of a histone-deacetylase enzyme (Garcia-Perez et al., 2010).

The insertion of MEs can also result in the recruitment of other chromatin modulating factors, typically through the methylation of DNA at CpG dinucleotides (Kazazian & Moran, 2017). Cytosines can become methylated to form 5-methylcytosine that frequently leads to the conversion of methylcytosine to thymidine during DNA replication (Batzer & Deininger, 2002). This is classified as genomic modifications that can cause changes in gene expression. Changes in DNA methylation patterns play a crucial role in suppressing transcription, thereby restricting ME expression in both germline and somatic cells (Kazazian & Moran, 2017). The methylation of LINE- and SINE-embedded CpG islands also has the potential to silence the expression of nearby genes (Elbarbary et al., 2016). Furthermore, somatic cells display a defense mechanism against ongoing retrotransposition through the methylation of MEs (Bestor & Bourc'his, 2004; Scott & Devine, 2017). Alternatively, there seems to be a correlation between hypomethylation and ME activation (Konkel & Batzer, 2010). In particular, demethylation of ME promoters may result in their activation, which could have implications on increasing levels of MEs and function of transcription factors (Lee et al., 2015). Cancer-associated chimeric transcripts can also be generated from the activation of the L1 antisense promoter as a result of demethylation

(Kazazian & Moran, 2017; Konkel & Batzer, 2010). This indicates that the methylated state of DNA relating to epigenetic changes can adjust the levels of ME activity, leading to alternations in gene function or overall gene expression (Ayarpadikannan et al., 2015).

1.3.6 ME association to human disease

ME-driven genetic modifications as a result of retrotransposition events in the genome or inappropriate expression levels of MEs may have significant implications on health and disease. Several studies have explored the connection between ME-mediated genome regulation and disease related phenotypic effects and revealed ME polymorphisms that are likely to be associated with disease phenotypes. For example, an SVA insertion was located in the enhancer of the *B4GALT1* gene (Wang, Norris, & Jordan, 2017). This gene acts to convert the immunoglobulin G (IgG) antibody from a pro-inflammatory to an anti-inflammatory form. The insertion of the SVA element resulted in the down-regulation of *B4GALT1*, thereby influencing increased inflammation (Wang & Jordan, 2018). This event has been linked to inflammatory conditions known as Crohn’s disease and systemic lupus erythematosus. Other diseases that have been demonstrated to be caused by ME insertions include cystic fibrosis, haemophilia, and X-linked genetic disorders (Kazazian et al., 1988; Wang et al., 2017).

Both L1 and Alu insertions are instigators of human genetic disorders, as it has been estimated that about 0.3% of all human disease are caused by insertions (Ayarpadikannan et al., 2015). As of 2009, 65 reported cases of *de novo* insertions were shown to cause various heritable diseases such as cystic fibrosis and haemophilia (Cordaux & Batzer, 2009). Furthermore, polymorphic Alu elements are known to be highly associated with disease risk. Reports have indicated polymorphic Alu elements to be potential causative candidates for various health conditions, including multiple sclerosis, obesity, acute lymphoblastic leukemia, psoriasis, and breast cancer (Payer et al., 2017). More specifically, ectopic recombination of Alu elements has led to the genomic rearrangement of the breast cancer gene (*BRCA1*) (Peixoto et al., 2013). Modifications to the *BRCA1* gene through Alu insertions have also been associated with ovarian and breast cancer in women (Peixoto et al., 2013). Taken together, MEs have a considerable impact on human genome evolution through processes influencing genetic instability, gene regulation, and disease occurrence.

1.4 ME insertion polymorphisms

The continuous activity of retrotransposition can create novel ME insertions in the germline cells that are capable of being inherited, thereby generating genetic diversity among human populations, as well as within individuals. ME polymorphisms are defined as the presence or absence of an insertion at a specific location in the genome, and it is not the sequencing variations within the MEs. ME polymorphisms can be detected either as insertions or as deletions in samples relative to the reference genome. Therefore, ME insertions distinguished as insertions to the reference are labeled as non-reference MEs, while ME insertions detected as deletions to the reference are known as reference MEs. Such polymorphic MEs represent superior genetic markers for studies of human ancestry and evolution because of their nature of being identical by descent (IBD) instead of being identical by state (IBS) like single nucleotide polymorphisms (SNPs) (Rishishwar, Villa, & Jordan, 2015). ME activity has had a major impact on the development and structure of the genome throughout the evolutionary history of humans (Wang et al., 2017). Therefore, polymorphic ME insertions exemplify valuable genetic variations for the study of human genetic diversity and structure of present human populations.

1.4.1 Identification of polymorphic MEs in the human genome

Various methodologies have been applied over a number of years in identifying the currently known polymorphic ME insertions. Many of the earlier methods revolved around PCR-based techniques. For instance, genomic library screening with primers specific for young Alu elements was used in earlier studies to discover a small number of polymorphic Alu insertions (Arcot et al., 1995; Batzer et al., 1996; Batzer et al., 1991; Roy et al., 1999). The utilization of library screening combined with sequencing strategies successfully identified new polymorphic insertions, including L1 and LTR elements. More specifically, the analysis of 17 genomes with the use of high-throughput pair-end Sanger sequencing led to the discovery of 198 reference L1 insertions and 1 HERV-K non-reference insertion (Kidd et al., 2010). Despite the progress of polymorphic ME discovery, challenges still existed to find these polymorphisms due to the high sequence similarity between the copies found in the genome. The first large-scale comprehensive study was presented when human genome sequences became available (Lander et al., 2001; Venter, 2001). Utilizing human genome sequences by computational sequence analysis in addition to PCR allowed for ascertaining the polymorphic status of the ME candidates by screening DNA samples from diverse human samples. Taken together, this approach used by many studies are responsible for the identification of over 400 polymorphic Alu insertions (Carroll et al., 2001; Otieno et al., 2004; Roy-Engel et al., 2001). However, this selection of discoveries was limited to MEs covered in the public version of the human genome sequence and the candidates were biased towards young ME subfamilies. To identify polymorphic MEs that are absent in the reference genome, genomic DNA sequences from more human individuals representing different populations are needed. The first attempt at this approach used partial human trace genomic sequences representing 36 diverse humans to compare with the reference genome, which identified over 600 Alu, L1, and SVA polymorphisms (Bennettt et al., 2004). Another large-scale computational study identified over 800 Alu and 150 L1 polymorphisms by comparing the public and Celera versions of the human genome sequences (Konkel et al., 2007; Wang et al., 2006).

With the advancement of next generation sequencing (NGS) technologies, it has become possible to characterize genome-wide patterns of polymorphic insertions using computational analysis of whole genome sequence data. The use of NGS to selectively sequence the junction areas between ME insertions and their flanking genomic sequences, in addition to mapping sequences to the reference genome, have identified a large number of novel ME polymorphisms in the last few years (Ewing & Kazazian, 2010; Witherspoon et al., 2010). As time progresses, a large number of personal genome sequences are becoming available, such as those generated by the 1000 Genomes Project, which contains a comprehensive genome-wide data set of over 16,000 ME polymorphic sites cataloged among 2504 individuals from 26 human populations (Phase 3 data November 2014) (Altshuler et al., 2012; Auton et al., 2015). With the growing collection of data from various sequencing projects, a number of computational tools and pipelines have been developed for the detection of ME polymorphisms in human genomes. For instance, a tool called AluMine, has recently been developed to analyze polymorphic Alu insertions in the human genome through the analysis of personal genomes (Puurand et al., 2019). This approach has been tested on 2241 individuals from the Estonian Genome Project and has identified more than 28,000 potential polymorphic Alu insertions (Metspalu, 2004; Puurand et al., 2019). Another recent study has applied *de novo* whole-genome sequencing of one Korean individual, and through computational analysis has resulted in the identification of over 19,000

ME polymorphisms. Further analysis was performed to experimentally validate 271 of these polymorphic candidates (personal communication).

Other initiatives of large-scale whole genome sequencing are currently in progress, including the Sanger institute in the United Kingdom to sequence 100,000 human genomes (Turnbull et al., 2018). With increasing numbers of sequenced individual genomes becoming available, we can expect the discovery of novel ME polymorphisms to increase at a very fast rate. Overall, only a limited number of polymorphic MEs with complete sequences have been generated from these analyses, demonstrating the need of improving the quality of analyses to obtain a better compilation of polymorphic ME data.

1.4.2 Documentation of ME polymorphisms in the Database of Retrotransposon Insertion Polymorphisms (dbRIP)

Due to the large number of polymorphic MEs with the potential of many more to be identified, dbRIP offers the compilation of this data in its own database (http://dbrip.org/) (Wang et al., 2006). dbRIP provides detailed sequence information associated with polymorphic ME insertions, which include the ME insertion sequence, the TSD sequences, and the flanking regions of each ME insertion. Additional information that can be found for each entry in the database includes the presence of IMDs and 5’ or 3’ transduced sequences, as well as selected genotype analysis and insertion allele frequencies in the examined human populations completed through polymerase chain reaction (PCR). Full components of information such as this can provide sufficient data about the potential impact of the insertion. Therefore, dbRIP is a valuable resource for the study of ME polymorphisms present in humans. Prior to this study, dbRIP contained a total of 3133 unique ME insertion polymorphisms, among which are 2539 Alus, 492 L1s, 9 LTRs, and 93 SVAs.

1.5 Research Objectives

The overall goal of this study is to identify and characterize polymorphic ME insertion sequences and their distribution pattern in the human genome, as well as to examine human population structure with these ME sequences through genotype analysis. Novel polymorphic ME identification and complete sequence characterization will provide the opportunity to update dbRIP with detailed documentation of new entries and make changes to current dbRIP data. To accomplish this, a non-redundant list must first be generated from a collection of data representing human polymorphic MEs from the literatures, collaborators, and new data from our research group. Further processing of all polymorphic MEs will be done to ensure conformance to the new standardized formatting. All processed polymorphic MEs will require sequences to be fully characterized, validated using computational and experimental methods, and organized in proper format for future use and amendment of dbRIP. The advancement of genome sequencing and the availability of whole genome sequences allows for the opportunity to replace traditional methods of genotyping via PCR with new strategies such as *in silico* genotyping. With the standardized and validated data set of polymorphic MEs, *in silico* genotyping will be performed using a diverse selection of human populations to assess the patterns of polymorphic MEs. Genotype data collected across all polymorphic MEs will be used to examine the distribution of human genetic diversity by measuring ME insertion allele frequencies. Lastly, a functional assessment of polymorphic MEs will be performed based on the context of the genomic location of the insertions. Furthermore, these polymorphic MEs will also allow to evaluate the activity of DNA transposition in current human genomes and provide an insight of their potential functional impact on genome evolution and function.

Chapter 2: Materials and Methods

The methods for examining the genetic polymorphism in humans involved computational analysis of personal genomes to characterize polymorphic ME insertions and performing an allele frequency survey through *in silico* genotyping.

2.1 Generation of a non-redundant list of polymorphic MEs

The preliminary list of MEs is those reported in the database of retrotransposon insertion polymorphisms in humans (dbRIP) (http://dbrip.org/) (Wang et al., 2006). New polymorphic ME candidates were accumulated by a combination of in-house data, 1000 Genomes Project (Auton et al., 2015; Sudmant et al., 2015), and unpublished data from our collaborator, Kyudong Han (personal communication). The collection of in-house data contains polymorphic MEs from our research group as well as a Brock computer science graduate student, Yaroslava Girilishena (Girilishena, 2017). Many ME candidates from our group were subjected to experimental validation and a limited survey of allele frequency through PCR. The genomic DNA samples used in validation include 6 samples for two trio families, and a 24-sample panel of Polymorphism Discovery Resource, all purchased from the Coriell Biorepository (http://catalog.coriell.org/). These samples cover the major ethnic population groups and were used for PCR genotyping to confirm and validate the candidates from computational analysis. A set of data with full sequence characterization was collected through a computational approach (Girilishena, 2017). This was performed to generate complete sequences for polymorphic MEs previously reported from our research group with incomplete sequences. This approach was developed through a set of computer algorithms to provide complete genome sequence characterization for polymorphic MEs via local *de novo* sequence assembly or progressive assembly using discordant and concordant read pairs and split-reads associated with the ME insertion loci. Personal genome data that is available to the public was utilized in this method (Girilishena, 2017). Taken together, this exhaustive list combining all sources of data contained a redundant total of 6434 polymorphic MEs (Table 1). A non-redundant polymorphic ME list was generated by performing a position overlap using BEDtools window to ensure there were no overlapping between the MEs (Quinlan & Hall, 2010). A total of 4400 polymorphic MEs, with 3133 entries documented in dbRIP, plus an additional 1267 newly identified entries, were included in the final list of polymorphic MEs (Table 1). These final numbers were obtained following various filtering processes that were applied to the entire list of polymorphic MEs. Entries with insertion sequences less than 50 bp in length were not considered to be ideal candidates, resulting in the elimination of 34 entries. Downstream analysis using genotype data gathered for each polymorphic ME was used to further process and filter the list of MEs. A total of 41 entries were eliminated with this process. (Details are provided in section 2.6).

Table 1. Collection of polymorphic ME data from various sources

Data Sources	Redundant	Non-Redundant
dbRIP	3185	3133
Korea	182	72
1000 Genomes Project	2176	433
In-house	891	762
Total	6434	4400

2.2 Sequence format and organization of polymorphic MEs

Following the generation of collected data representing polymorphic MEs, the next steps were designated to the organization and standardization of the ME entries. Each polymorphic ME contained pre-integration and insertion allele sequences which were assembled in a uniformal format to include 400 bp of upstream and downstream flanking sequences with full ME insertion and applicable TSD sequences. The description line for each ME entry consists of a dbRIP ID and other related specific information regarding the locus. The dbRIP ID system are unique numbers at a fixed length of 7 digits with the first digit indicative of the ME type. Detailed information regarding the dbRIP ID designation for each ME type can be found in Table 2.

Table 2. Unique dbRIP ID system used to label individual polymorphic MEs

ME Type	dbRIP ID System
Alu	1000000~1999999
L1	2000000~2999999
SVA	3000000~3999999
LTR	4000000~4999999
PPSG	5000000~5999999

The description line for each ME entry also presents the ME type, family and subfamily, strand orientation (plus or minus), human genome build (hg19), genome position of the ME with flanking regions included, genome positions of the ME insertion only, allele type (reference/non-reference), and specific genomic sequences and rearrangements MEs may contribute to including TSDs, transductions, and IMDs. This standardized organization and format was computationally applied to all polymorphic ME entries using in-house Perl scripts and complete sequences for each locus are organized as a pair (for dimorphic) or three (for tri-morphic LTR insertions) sequences in FASTA format. The sequence layout of an example ME entry is shown in Figure 2.

```
>dbRIP|1003228|ME|SINE:Alu:AluYb8|Strand|+|Genome|hg19|Pos|chr13:67725457-
67726249;chr13:67725857-67725858|Allele|pre|Insertion|non-
ref|TSD|7:AAAACTG|IMD|0|5TR|0|3TR|0

ATTTGTTCCTCTTTAGGGATATTAACGTTTAATGCAATGAAAGAACATGACCTTATTTCCTACAGCCTTTGCAC
ATCTCTTCATGGGTTTCTGGTTGTGGGTGTCTGTGTCTGAGATGCAAAAAGCAGAATTGGGTTCTACTGGTAAA
CTCAGAGCTCCTACTTCTTTTATTGTTTGCCTGAAAAAAAATAGGCATAGTTTTGACTTAGGTAGGGGATCTAA
GAGTACTGCATCATTCCCAAATCCAACACTGTTAGAGTTAGAAGATCCTGTATGATTCAAGCACCTGGTTTCAC
ACAGTGACTAAGGATGGGGAATATATCCCTTCTACAAGCTGGCAAAGCTTTTGCTCTATACTTAGAGGAAGAAA
AAAATTCATCTTTTCAATCTATC

AAAACTG

AGGCACATAATAATTGAGAGGAGAAACTTATCATTATGCTCAGGGACTTTATGCTTTTGAGGCCTCATAAAGGA
CTCAGGAAGTATCAGCCCTTGTATAGCTGCAGGCAACATCTCCACCCAATTCTAGGCCACACTCATTTTTAAGG
CACCAAAGCTCAAATGCACTCAGGGTCCTTGCATCATCATTTTGCCTAAAGAGAGATAATTTTTTTTTTGTTAC
AAGACTATATGAATGAACTAAAAGACTCTTAAGGATACTTTCTCATCTCATAAAACTCTCATCCCAAGTTACAT
GAGATAAACTGAAATAGTATTTTACAAATTCCTCCAAATTGTATCTTCTTATAGCTGAAGAACTTACAAGACTT
TTTGTGTCCTATTTGTTCCAAGA

>dbRIP|1003228|ME|SINE:Alu:AluYb8|Strand|+|Genome|hg19|Pos|chr13:67725457-
67726249;chr13:67725857-67725858|Allele|ins|Insertion|non-
ref|TSD|7:AAAACTG|IMD|0|5TR|0|3TR|0

ATTTGTTCCTCTTTAGGGATATTAACGTTTAATGCAATGAAAGAACATGACCTTATTTCCTACAGCCTTTGCAC
ATCTCTTCATGGGTTTCTGGTTGTGGGTGTCTGTGTCTGAGATGCAAAAAGCAGAATTGGGTTCTACTGGTAAA
CTCAGAGCTCCTACTTCTTTTATTGTTTGCCTGAAAAAAAATAGGCATAGTTTTGACTTAGGTAGGGGATCTAA
GAGTACTGCATCATTCCCAAATCCAACACTGTTAGAGTTAGAAGATCCTGTATGATTCAAGCACCTGGTTTCAC
ACAGTGACTAAGGATGGGGAATATATCCCTTCTACAAGCTGGCAAAGCTTTTGCTCTATACTTAGAGGAAGAAA
AAAATTCATCTTTTCAATCTATC

AAAACTG

GCGCGGTGGCCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCG
AGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGCGGTGGCGGG
CGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGA
GCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGCGAGACTCCGTCTCAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAACAA

AAAACTG

AGGCACATAATAATTGAGAGGAGAAACTTATCATTATGCTCAGGGACTTTATGCTTTTGAGGCCTCATAAAGGA
CTCAGGAAGTATCAGCCCTTGTATAGCTGCAGGCAACATCTCCACCCAATTCTAGGCCACACTCATTTTTAAGG
CACCAAAGCTCAAATGCACTCAGGGTCCTTGCATCATCATTTTGCCTAAAGAGAGATAATTTTTTTTTTGTTAC
AAGACTATATGAATGAACTAAAAGACTCTTAAGGATACTTTCTCATCTCATAAAACTCTCATCCCAAGTTACAT
GAGATAAACTGAAATAGTATTTTACAAATTCCTCCAAATTGTATCTTCTTATAGCTGAAGAACTTACAAGACTT
TTTGTGTCCTATTTGTTCCAAGA

//
```

Figure 2. A single polymorphic ME entry with complete sequence characterization organized in a standardized format. Top definition line presents the pre-integration allele followed by 400 bp of 5’ flanking sequence, TSD sequence, and 400 bp of 3’ flanking sequence. The bottom definition line presents the insertion allele sequence and differs by the presence of the ME insertion sequence (yellow) and an additional TSD sequence (red).

2.3 Sequence characterization and validation of polymorphic MEs

The University of California at Santa Cruz (UCSC) genomic website (http://genome.ucsc.ca) is an interactive resource offering sequence and annotation data downloads for various genome assemblies, including human and other primate genomes. The available annotation database for the current version of the *Pan troglodytes* (chimpanzee) genome (Jan.2018 Clint_PTRv2/panTro6), was used to ensure the TSD sequences for each polymorphic ME were correct in sequence and length. The pre-integration allele of each ME was aligned to the *Pan troglodyte* genome with a complete alignment indicative of the valid presence or absence of the TSD sequence. This was completed by using BLAST-like-alignment tool (BLAT), which is a tool used to perform sequence alignments at a high sequence similarity level (Kent, 2002).

Further sequence characterization of polymorphic MEs was necessary to identify ME contribution to potential genomic sequence rearrangements such as transductions and IMDs. This was accomplished using various in-house Perl scripts. To identify potential IMDs caused by ME insertions, flanking sequences on each side of the insertion was extracted and joined together to form a total of 100 bp of sequence for each ME. Using BLAT, these sequences were aligned to the human reference genome (non-reference MEs) or panTRo6 genome (reference MEs). From the alignment analysis, entries containing a gap region (extra sequences) in the reference genome next to the insertion position were considered cases of IMDs. To identify potential transduction events, the insertion sequence for each ME was extracted and analyzed using RepeatMasker, which is a program used to screen DNA sequences for repeats and low complexity DNA sequences (http://www.repeatmasker.org/). The RepeatMasker run was performed using a customized repeat library consisting of only the known active types of MEs as a way to speed up the process. With the known type of repeat sequences masked to lower cases, additional sequences not belonging to the ME insertion could be identified. Sequences in upper cases on the 5' end of the sequence were considered as 5' transductions while the presence of extra sequences on the 3' end were labelled as 3' transductions. These candidate cases were further manually verified using the UCSC Genome Browser, which also allows to track the source of the transduced sequences for a portion of the entries.

2.4 Sources for personal genome sequences

An approach for characterizing polymorphism is to computationally analyze the personal genome sequence data that have become available for a large number of human individuals through large genome sequencing efforts, such as the 1000 Genomes Project (Auton et al., 2015). The personal genome sequence data used in this study included mainly the phase three data of the 1000 Genomes Project, plus several additional small data sets available at the NCBI sequence read archive (SRA), including one set of 2 genomes for Khoisan and Bantu from Southern Africa (Schuster et al., 2010) and another set of 32 genomes for Native Americans (Raghavan & Eriksson, 2015). The phase three 1000 Genomes Project data provides sequences of genomes for more than 2500 individuals representing 26 commonly recognized human populations from five continents. The final data set contains genome sequences of individuals sequenced at an average of 2-4x coverage (Auton et al., 2015). All genome sequences were downloaded onto our local servers for in-house analyses. Table 3 lists detailed information of the personal genome data used to represent various human populations.

Table 3. Human populations analyzed in study

Continental Group	Population ID	Full Description	*n*
African	ACB	African Caribbean in Barbados	96
	ASW	African Ancestry in Southwest US	61
	ESN	Esan in Nigeria	99
	GWD	Gambian in Western Divisions in the Gambia	113
	LWK	Luhya in Webuye, Kenya	99
	MSL	Mende in Sierra Leone	85
	YRI	Yoruba in Ibaden, Nigeria	108
East Asian	CDX	Chinese Dai in Xishuangbanna, China	93
	CHB	Han Chinese in Beijing, China	103
	CHS	Southern Han Chinese	105
	JPT	Japanese in Tokyo, Japan	104
	KHV	Kinh in Ho Chi Minh City, Vietnam	99
South Asian	BEB	Bengali from Bangladesh	86
	GIH	Gujarati Indian from Houston, Texas	103
	ITU	Indian Telugu from the UK	102
	PJL	Punjabi from Lahore, Pakistan	96
	STU	Sri Lankan Tamil from the UK	102
European	CEU	Utah Residents with Northern and Western European Ancestry	99
	FIN	Finnish in Finland	99
	GBR	British in England and Scotland	91
	IBS	Iberian Population in Spain	107
	TSI	Toscani in Italia	107
Ad Mixed American	CLM	Colombians from Medellin, Colombia	94
	MXL	Mexican Ancestry from Los Angeles, USA	64
	PEL	Peruvians from Lima, Peru	85
	PUR	Puerto Ricans from Puerto Rico	104
Bushmen	BUSHMEN	Khoisan and Bantu from Southern Africa	2
Native American	NATIVEA	Native American	32

2.5 Density and distribution of polymorphic MEs across the human genome

The distribution of polymorphic MEs in the human genome was first analyzed based on their density among chromosomes. The total length of each chromosome was measured in bp of non-gap sequences and used to calculate the density of polymorphic MEs. To provide a visual representation of ME distribution in the genome, an in-house Perl script was used to generate a density plot of Alus, L1s, SVAs, LTRs, and PPSGs along the chromosome ideogram. Monte

Carlo simulation was performed to judge whether the density of polymorphic MEs are statistically above or below the expected density based on random distribution. To do this, an in-house Perl script was used to compose a simulated list of genomic locations mimicking the size of polymorphic ME insertions. The entire genome sequence was processed to generate a list of window sizes equivalent to each ME and this generation of randomly picked regions was repeated for 100 times. With this analysis, the observed and expected values of polymorphic MEs were used to perform statistical calculations for each chromosome to assess the pattern of distribution, such as mean, standard deviation, Z-scores, and P-values.

2.6 *In silico* genotyping of polymorphic MEs

Publicly available personal genome sequences have allowed the possibility of performing *in silico* genotyping of all polymorphic MEs through the use of in-house Perl scripts. A genotyping pipeline implemented in Perl was assembled and used to obtain the specific genotype of a ME insertion in individual genomes, ME insertion allele frequencies, and distribution of polymorphic MEs in a population.

For genotyping a ME in a genome, junction sequences representing the pre-integration allele and the 5’ and 3’ insertion alleles for each ME locus were processed using another in-house Perl script and stored in FASTA format for use in the genotyping pipeline. Specifically, the pre-integration allele consists of 50 bp sequences from each side of the ME insertion joined together to represent the 5’ and 3’ flanking regions. The 5’ insertion allele consists of 50 bp from the 5’ flanking sequence and 50 bp from the 5’ insertion sequence, whereas the 3’ insertion allele consists of 50 bp from the 3’ insertion sequence and 50 bp from the 3’ flanking sequence. Figure

2 illustrates these three forms of junction allele sequences, with total lengths of 100 bp, generated for each polymorphic ME.

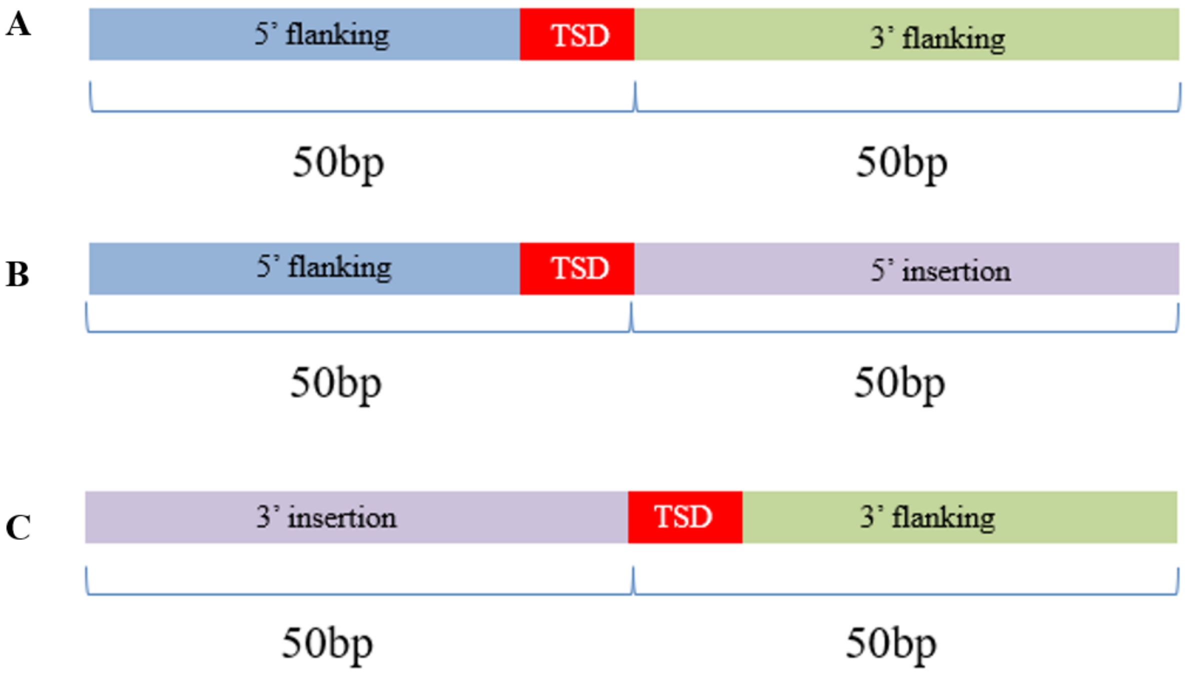

Figure 3. Three forms of junction allele sequences generated for each ME in preparation for *in silico* genotyping. A) Junction sequence representing the pre-integration allele. B) Junction sequence representing the 5' insertion allele. C) Junction sequence representing the 3' insertion allele.

Personal genome sequences were stored as binary (BAM) representations of the Sequence Alignment/Map (SAM) format and processed using SAMtools (Li et al., 2009). The use of this tool extracts all sequences (reads) aligned to a specified region, which in our case is to the start and end positions of each ME locus, including 150 bp of upstream and downstream flanking sequence. All BAM files were processed for each ME locus and the reads mapping to the specified position were collected and stored into a FASTA file. The collected reads were further aligned to the pre-integration allele and insertion allele sequences of the ME insertion as described in section 2.2 and filtered to ensure that each read was properly aligned to the pre-integration or insertion allele. The final number of reads mapped to the pre-integration and insertion alleles were used to determine the genotype of the ME insertion as one of the three possible genotypes -/-, -/+, and +/+, plus -/NA, and +/NA, for cases where the presence or absence of an allele is not determinable due to insufficient number of reads available for the locus from a subject. In this case, a "-" represents an absence of the ME insertion, while "+" represents the presence of the ME insertion.

On a smaller scale, individual level genotyping was obtained by organizing collected reads for each individual through the arrangement of sample IDs within the populations analyzed. Genotype calls for each ME were made by using BLAT to perform an alignment between the ME junction allele sequences and the collected reads representing individual populations or individuals within a population. The optimal criteria for determining an appropriate alignment between sequences were determined by the span of the alignment and the read count per junction allele type. A proper alignment between the read sequence and ME junction allele sequence should span the junction point with a read count of 2 or more to confirm the true presence of the specific allele type. Table 4 illustrates the various genotype calls determined by the BLAT output. ME entries with only 1 read supporting either the pre-integration allele or insertion allele were not considered as reliable candidates for the list of polymorphic MEs and were therefore eliminated.

Table 4. Genotype calls generated from ME allele read counts processed in the *in silico* genotyping pipeline

Number of reads aligned to the ME pre-integration allele	Number of reads aligned to the ME insertion allele	Genotype
>=2	>=2	-/+
0	>=2	+/+
>=2	0	-/-
<2	0	-/NA
0	<2	+/NA

More specifically, allele frequencies were calculated for all polymorphic MEs either with samples pooled by population or by the sum of individual genotypes. The allele frequency calculation was based on the ratio between the average number of reads mapping to the insertion allele to the overall number of reads mapped to both the insertion allele and pre-integration allele of each ME. The formula used to calculate the insertion allele frequencies for each polymorphic ME locus is shown below:

$$\frac{Total\ Number\ of\ Reads\ for\ 5'and\ 3'Insertion\ Junction\ Alleles/2}{Total\ Number\ of\ Reads\ for\ both\ alleles\ of\ the\ ME}$$

2.7 Distribution of polymorphic ME insertion allele frequencies

To evaluate the spectrum of ME insertion allele frequencies across all examined human populations, the frequencies were binned into 20 groups of 0.05 increments ranging from 0 to1. The allele frequency binning was processed using an in-house Perl script and completed using two different approaches. The first method examined the distribution of ME allele frequencies for each population individually and the second method combined the insertion allele frequencies across all populations to view the distribution as one sample. A similar process was used to examine the distribution of insertion allele frequencies comparing reference and non-reference polymorphic MEs, ME types, and continental groups.

2.8 Principle component analysis of human population clustering based on the genotype data of polymorphic MEs

Principle components analysis (PCA) is regularly used in population genetics as a common statistical method to identify structure in historical demographic processes (McVean, 2009). This technique was used to analyze the relationship of human populations based on polymorphic ME genotype data obtained through *in silico* genotyping. PCA algorithms and plots were programmed in R using various packages including 3D PCA imaging.

2.9 Phylogenetic analysis of polymorphic MEs

Genotype data was used to illustrate and construct the phylogenetic relationship between individuals based on the presence or absence of polymorphic MEs. All genotype data was concatenated to generate single pseudo-sequences per sample to represent each individual analyzed in the study. Genotype calls revealed by the *in silico* genotyping pipeline were converted to nucleotide letters, where the letter “C” represents the presence of the insertion allele either as “-/+”, “+/+”, or “+/NA” in genotype, the letter “A” represents the absence of the insertion allele (i.e., with a genotype of “-/-”), and the letter “N” represents an unknown presence of the insertion allele (with the genotype being “-/NA” or “NA/NA”. All generated pseudo-sequences with more than 1800 bp of “N” were filtered out using an in-house Perl script, resulting in 869 individual samples in the phylogenetic analysis. The samples filtered with this cut-off still contain a good representation of individuals from all 28 human populations. A sample representing the ancestral lineage of polymorphic MEs was generated to have all 4400 MEs with a “-/-” genotype, indicating a complete absence of the insertion allele. The pseudo-sequence of the ancestor sample was therefore represented by 4400 bp of "A's". All pseudo-sequences were aligned with ClustalW-MPI, which is a parallel application of ClustalW, a tool for performing multiple sequence alignments (Li, 2003). This alignment was achieved through pair-wise alignment followed by the generation of a neighbour-joining phylogenetic tree to represent the evolutionary relationship among the sequences. FigTree was used to graphically view the generated phylogenetic tree (http://tree.bio.ed.ac.uk/software/figtree/).

2.10 Heatmap clustering of polymorphic MEs

With the calculated insertion allele frequencies, a heatmap was created to illustrate the distribution of polymorphic MEs across 28 human populations. Multiple experiment viewer (MeV) is an application that allows to analyze and visualize large genomic data through heatmaps and clustering of data (Howe et al., 2010). Population-specific polymorphic ME entries, being those with the insertion allele absent (allele frequency of 0) in 2 or more populations, were processed using an in-house Perl script and included in this analysis, resulting in a total of 703 loci. Hierarchical clustering was applied to both the samples and polymorphic MEs to identify the cluster relationships.

2.11 Functional assessment of polymorphic MEs

A functional assessment of polymorphic MEs was performed based on the context of the genomic location of the insertions. This was completed by incorporating the utility of UCSC liftOver tools (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert all hg19 genomic positions to hg38 prior to analysis to utilize its better functional annotation information.

2.11.1 Analysis of polymorphic MEs in exon regions

A non-redundant human gene exon list from GENCODE was used to identify the number of MEs in these regions (https://www.gencodegenes.org/). Based on the detailed content in the

GENCODE file, each exon is annotated with a feature among CDS, 5' UTR, 3' UTR, NR, representing coding DNA sequences, 5' untranslated regions, 3' untranslated regions, and non-coding RNA. A position overlap was performed between the MEs and the non-redundant exons using BEDtools.

2.11.2 Analysis of polymorphic MEs in regulatory regions

The Ensembl Regulatory Build containing regions that are predicted to regulate gene expression was downloaded onto our local server and used to identify the number of MEs in these regions. The different types of regulatory features annotated in the Ensembl file include promoters, promoter flanking regions, enhancers, CTCF binding sites, transcription factor (TF) binding sites, and open chromatin regions (Zerbino et al., 2015). To identify the impact of MEs on regulatory features, a position overlap was performed between the MEs and the regulatory regions using BEDtools.

2.11.3 Analysis of polymorphic MEs in CpG islands

The available annotation database for the current version of the human genome (Dec.2013 hg38, GRChg38) at http://genome.ucsc.edu was used to obtain the positions of all identified CpG islands in humans. Using BEDtools, a position overlap was performed between the MEs and the CpG islands to identify MEs' involvement in these regions.

2.11.4 Analysis of polymorphic MEs' contribution to miRNA target sites

The Ensembl human gene annotations include updated data sets that can be retrieved and downloaded for use. The human miRNA target sites were extracted and downloaded directly from https://useast.ensembl.org onto our local servers. To identify the contribution of miRNA target sites by MEs present in the human genome, a position overlap was performed using BEDtools.

2.11.5 Analysis of polymorphic MEs' distribution pattern in the human genome among different genomic regions

The distribution pattern of polymorphic MEs throughout the human genome was examined by analyzing the relationship between the length of various genomic regions and the number of MEs present in those specific regions. Specifically, the regions examined include CDS exons, non-CDS exons (includes an accumulation of 5' UTR, 3' UTR, and NR exons), intron splice sites, intron non-splice sites, and intergenic regions. Additionally, the distribution of MEs in regulatory regions were also observed, including specific features of promoters, promoter flanking regions, enhancers, CTCF binding sites, TF binding sites, and open chromatin regions. The length of each genomic region was collected by totaling the lengths of each identified feature in each region. The density of MEs in each genomic region was calculated for every million bp based on the formula below:

$$\frac{Number\ of\ MEs\ Observed\ in\ Genomic\ Region}{Total\ Length\ of\ Genomic\ Region}\ x\ 1000000$$

2.12 Computational facilities

Data generation and analysis was made possible by SHARCNET computer systems, which are part of the Compute Canada high performance computing facilities (https://www.sharcnet.ca; http://www.computecanada.ca).

Chapter 3: Results

3.1 Overview of polymorphic ME identification, characterization, and documentation

3.1.1 A summary of identified polymorphic MEs in the human genome

To examine the composition of polymorphic MEs in the human genome, we first generated a list of all identified polymorphic MEs by combining multiple data sets documented by independent research groups (Table 1). These data sets were merged and processed to form a non-redundant set of polymorphic MEs. Along with specific chromosomal positions, each polymorphic entry contains full pre-integration and insertion allele sequences that were computationally formatted in a uniform FASTA format with the definition line containing detailed information about the locus and the insertion sequences as described in the Materials and Methods section. This complete data set only includes polymorphic MEs with full sequences for both alleles available. There are additional candidates that do not have full allele sequences and are therefore not included in this study.

The non-redundant list of polymorphic MEs consist of entries representing reference and non-reference cases for five different types of MEs including Alus, L1s, SVAs, LTRs, and PPSGs. These entries are categorized into two main groups, dbRIP and non-dbRIP. Non-dbRIP polymorphic MEs are collected entries from Korea, 1000 Genomes Project, and in-house data (Table 1). Prior to accumulating data from various data sets, dbRIP contained a total of 3133 (71.2%) entries, with 2170 representing reference entries and 963 non-reference entries (Table 4). From the 3133 entries, the majority of the MEs belonged to the family of Alus with 2539 entries, followed by L1s with 492 entries, SVAs with 93 entries, and LTRs with 9 entries (Table 5). Prior to this updated list of polymorphic MEs, dbRIP did not cover any reported cases of polymorphic PPSGs.

The accumulation of new data led to an increase of polymorphic MEs by approximately 29% from our original set in dbRIP, with 1267 newly identified non-dbRIP entries. From the 1267 ME entries, 433 (9.8%) entries were from the 1000 Genomes Project, 762 (17.3%) entries belonged to in-house data obtained from our research group and Yaroslava, and 72 (1.6%) entries were derived from the Korean data set from the Kyudong Han Laboratory at Seoul National University (Figure 4A). From the total 1267 non-dbRIP entries, 453 represented reference entries and 814 represented non-reference entries (Table 5). Similar to dbRIP, the majority of MEs among the 1267 entries belonged to Alus with 1113 entries, followed by L1s with 75 entries, SVAs with 29 entries, PPSGs with 28 entries, and LTRs represented by the smallest number of 22 entries. The updated list of non-dbRIP entries also contain 28 cases of polymorphic PPSGs (Table 5). Overall, the non-redundant list of polymorphic MEs contain a total of 4400 entries, with 2623 entries representing reference MEs and 1777 entries representing non-reference MEs. When organized by ME type, the total 4400 entries include 3652 Alus (83%), 567 L1s (12.9%), 122 SVAs (2.8%), 31 LTRs (0.7%), and 28 PPSGs (0.6%) (Figure 4B). When comparing the total number of polymorphic MEs between reference and non-reference entries, there seems to be a greater number of reference entries for each ME type (Figure 5).

Table 5. Summary of polymorphic MEs reported in the non-redundant list

ME type	Ref		Non-Ref		Total
	dbRIP	non-dbRIP	dbRIP	non-dbRIP	
Alu	1786	353	753	760	3652
L1	300	50	192	25	567
SVA	77	10	16	19	122
LTR	7	20	2	2	31
PPSG	0	20	0	8	28
Sub-total	2170	453	963	814	4400
Total (dbRIP/non-dbRIP)	3133/1267				4400
Total (Ref/Non-Ref)	2623/1777				4400

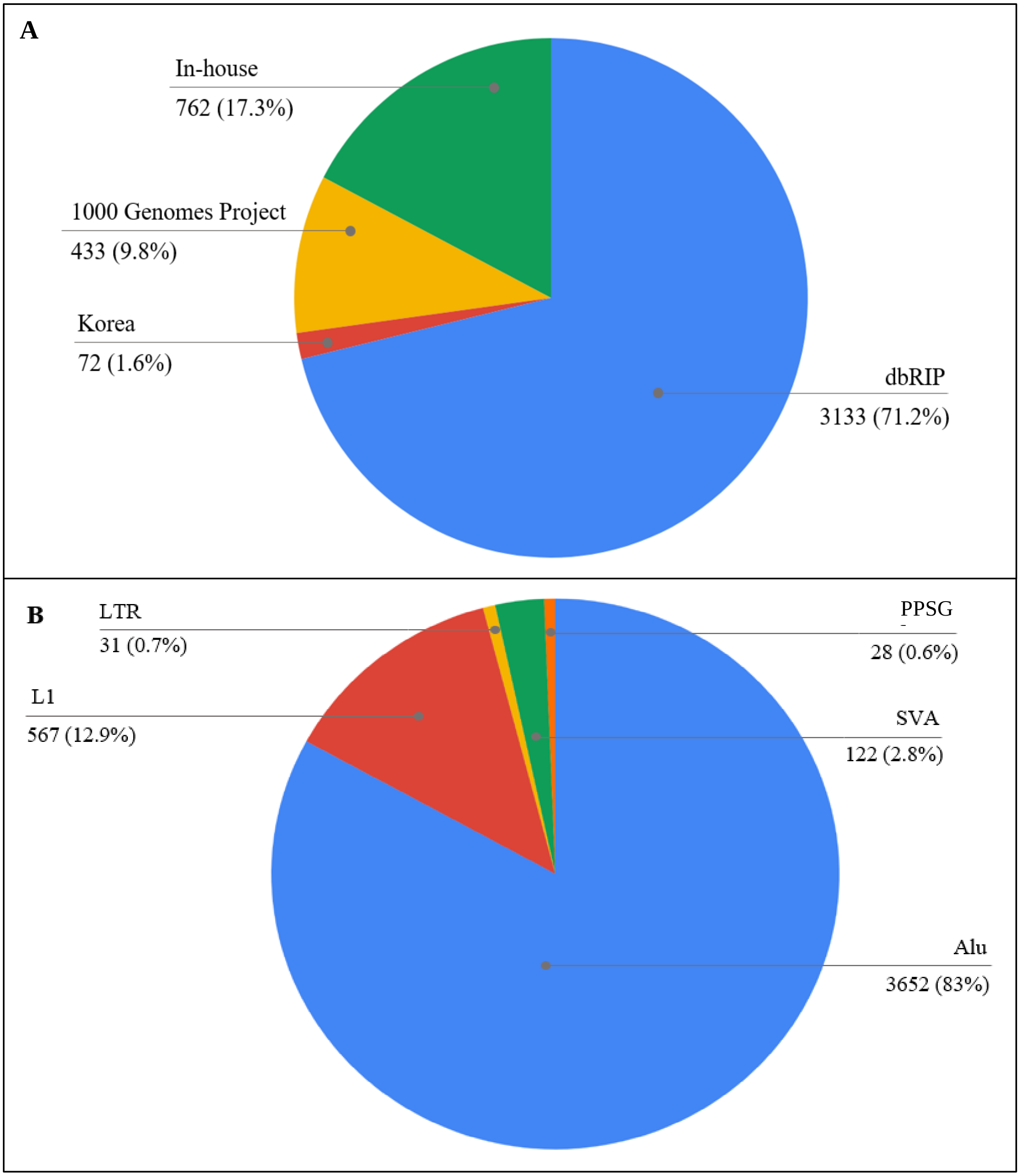

Figure 4. Outline of identified polymorphic MEs in the human genome. A) A breakdown of data set sources that were processed to generate the non-redundant list of polymorphic MEs B) Total number and percentages of identified polymorphic MEs organized by ME type

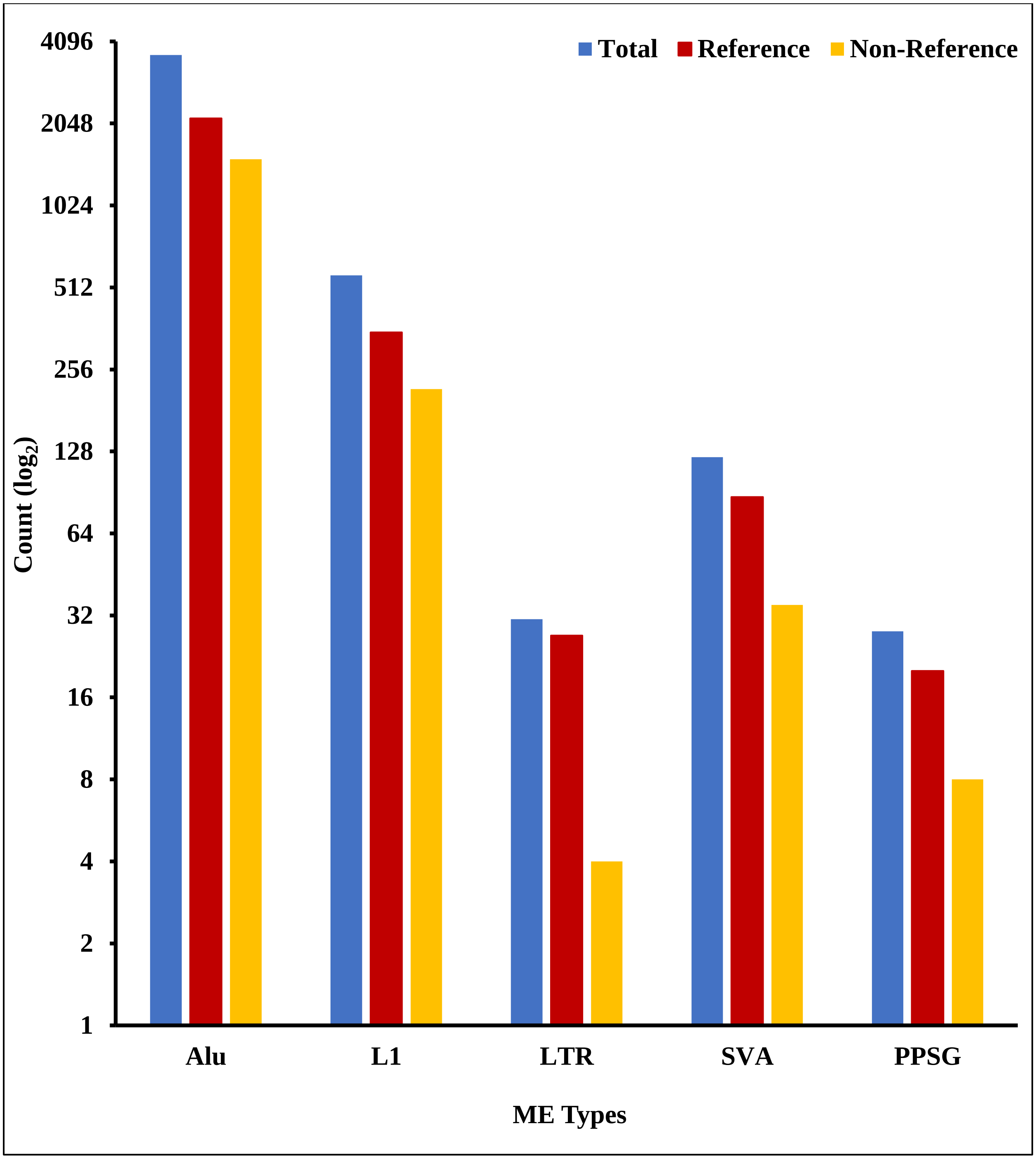

Figure 5. Comparisons of the total number of reference and non-reference entries across ME types. The ME count among each type of ME is represented in a log2 scale. The largest difference between the number of reference and non-reference entries is present among LTRs while Alus have the smallest difference between the reference and non-reference groups.

3.1.2 Validation of polymorphic MEs

Due to the various tasks including computational methods associated in identifying polymorphic MEs, it is necessary to validate the accuracy of the entries. The validation efforts focused on evaluating sequences and the polymorphism status of the MEs. Sequences and chromosomal positions of each polymorphic ME were validated through sequence alignments to the human reference genome in addition to the manual checking of entries on the UCSC Genome Browser. Data collected from various sources involved independent processes of validation. Polymorphic MEs of each type from dbRIP were validated using PCR, including a total of 686 loci with genotype information (Wang et al., 2006). Detailed information regarding each genotyped locus is available in the dbRIP database. ME entries collected from our personal communication in Korea was undergone computational and partial experimental verification. The combination of polymorphic ME data from the 1000 Genomes Project (Auton et al., 2015) and our in-house data were subjected to PCR validation using human DNA samples, completed through the efforts of multiple students from our research group (unpublished data).

Aside from direct sequence and polymorphism validation, ME entries with an insertion sequence of less than 50 bp were removed from the list of polymorphic MEs, since these are also covered in dbSNP as INDELs. Downstream analysis using genotype data obtained through this study was also used in this elimination process. ME entries with only 1 read supporting either the pre-integration allele or insertion allele in the entire whole genome data sets were not considered as ideal candidates and were therefore removed out from the non-redundant list. From the first filtering step completed by measuring the length of the insertion sequence, a total of 34 MEs were eliminated, including 27 Alus and 7 L1s. All 27 Alus were reported in dbRIP, where 3 were reference entries and 24 were non-reference entries (Table 6). All 7 removed L1s were also dbRIP entries and were non-reference MEs. From the second filtering step using genotype data, a total of 41 MEs were eliminated. More specifically, the total of 41 removed entries contained 32 Alus, 4 L1s, 1 SVA, 2 LTRs, and 2 PPSGs with a combined total of 17 reference and 24 non-reference MEs. Overall, the entire filtration process led to the deletion of 75 polymorphic MEs from the original list, resulting in the total of 4400 polymorphic MEs present in the current non-redundant list.

Table 6. Eliminated polymorphic MEs based on the filtering process of entries

ME type	Ref		Non-Ref		Total
	dbRIP	non-dbRIP	dbRIP	non-dbRIP	
Alu	8	6	43	2	59
L1	1	1	9	0	11
SVA	0	0	1	0	1
LTR	0	2	0	0	2
PPSG	0	2	0	0	2
Sub-total	9	11	53	2	75
Total (dbRIP/non-dbRIP)	62/13				75
Total (Ref/Non-Ref)	20/55				75

3.1.3 Polymorphic ME sequence characterization and contribution to genomic rearrangements

The insertional events of polymorphic MEs are known to cause alterations in the human genome through the generation of TSDs, transductions, and IMDs. Genomic modifications such as these may lead to an increase in genomic size through TSDs and transductions but may also result in reducing the size of the genome through IMDs.

A TSD length survey was performed for all polymorphic MEs and compared among the five ME types. As seen in Figure 6, the pattern of TSD length among Alus, L1s, SVAs, and

PPSGs showed a similar distribution pattern with the TSD length peaking around 15 bp. This analysis is similar to what has been previously reported for ME TSD lengths (Stewart et al., 2011; Tang et al., 2018; Wu et al., 2014). Among all non-LTR retrotransposons, majority of the entries have a TSD length between 10 to 20 bp, which has been reflected in previous studies (Szak et al., 2002). Minor differences in the complete pattern of TSD length distribution are also noticeable. Comparing to non-LTR retrotransposons, LTRs displayed a disparate pattern with the TSD length peaking around 6 bp, which is comparable to results from other studies reporting LTR TSD lengths between 5-6 bp (Dewannieux et al., 2006). A minor peak is also apparent between lengths of 2 to 4 bp for LTRs (Figure 6). PPSGs show a similar trend in TSD lengths seen in non-LTR retrotransposons with the TSD length peaking around 15 bp, however minor peaks are also present between lengths of 6 to 10 bp (Figure 6).

As shown in Table 7, the events of polymorphic ME insertions and the subsequent occurrence of genomic rearrangements do have impact on human genome size. A net genome sequence increase was observed for each ME type. Among the type of genomic rearrangements, there is a total of 26 transduction events mediated by ME insertions, with 2 cases of 5’ transductions, 12 cases of 3’ transductions, and another 12 cases of both 5’ and 3’ transductions. Between each type of transduction analyzed, the transduction of 3’ flanking sequences were more common and contributed the most in sequence length (Table 7). This trend seems to be similar among L1-mediated retrotransposition events as seen in previous studies (Goodier et al., 2000; Pickeral et al., 2000). The length of each transduction event varies, however the shortest case among the 26 transductions is 1 bp, whereas the longest was a 3’ transduction of 433 bp of flanking sequence, which involved both an Alu and LTR element (Appendix Table S1). None of the reported polymorphic MEs impacted the genome through IMDs (Table 7).

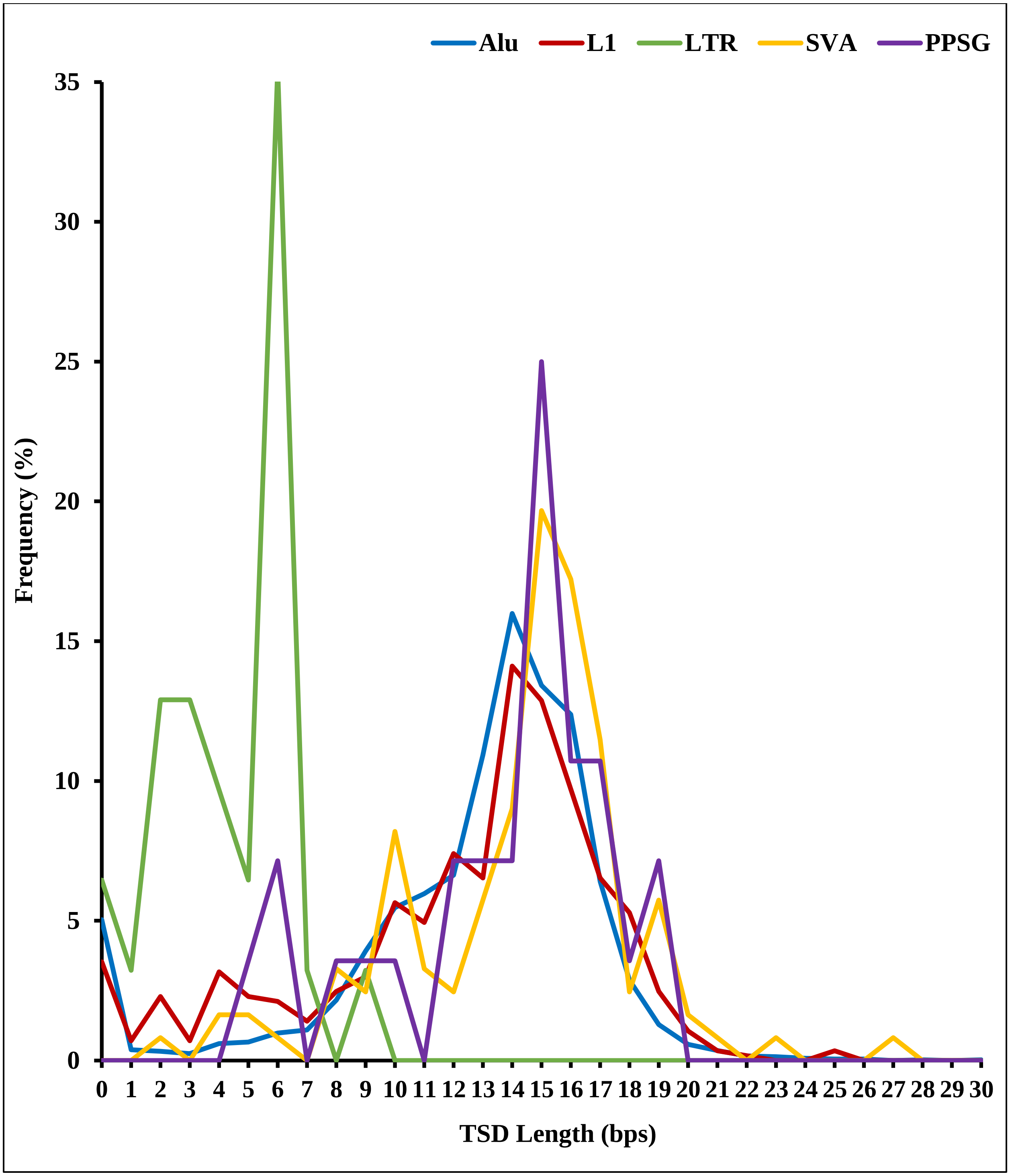

Figure 6. Survey of TSD lengths of polymorphic MEs. Line plots showing the frequencies of TSDs at each length for each type of polymorphic ME including Alus, L1s, SVAs, LTRs, and PPSGs.

Table 7. Impact of polymorphic MEs on genome size (bp)

ME Type	TSD		IMD		5' Transduction only		3' Transduction only		5' and 3' Transduction		Total	
	n	bp	*n*	bp	*n*	bp	*n*	bp	*n*	bp	*n*	bp
Alu	3468	59,912	0	0	0	0	0	0	0	0	3468	59,912
L1	547	8,235	0	0	2	6	12	733	11	253	572	9,227
SVA	122	1,742	0	0	0	0	0	0	0	0	122	1,742
LTR	29	232	0	0	0	0	0	0	1	59	30	291
PPSG	28	382	0	0	0	0	0	0	0	0	28	382
Total	4,194	70,503	0	0	2	6	12	733	12	312	4,220	71,554

3.1.4 Documentation and access of polymorphic ME data

The updated polymorphic ME data set will be deposited into the dbRIP database and be available to the public. In dbRIP, polymorphic ME entries can be visualized in the same approach as in the UCSC genome browser. The polymorphic ME data in FASTA format with the sequences organized to represent the insertion allele of each ME (left flanking, TSD1, ME insertion sequence, TSD2, right flanking), along with all related information provided in the definition lines will be available for downloading from the dbRIP download page (http://dbrip.org/). This update will also include *in silico* genotyping data for all entries for the populations included in this study to replace the limited genotype data for a small number of entries using a variable number of small samples.

3.2 Genome distribution patterns of polymorphic MEs

3.2.1 Retrotransposition activity level of polymorphic MEs

The comprehensive list of polymorphic MEs offered an opportunity to assess the activities of MEs and the patterns of polymorphism in the human genome. Similar to other studies (Mills et al., 2007; Mills et al., 2006), AluYa5 and AluYb8/9, L1HS, SVA_E/F, and HERV_K are shown to be the most active subfamilies for Alu, L1, SVA, and LTRs, respectively (Appendix Table S2). The activity level of a ME subfamily is assessed as the ratio of polymorphic MEs in relation to all MEs in the subfamilies. Based on this ratio analysis, the highest activity level for polymorphic Alus was represented by subfamilies AluYa5 (34.1%), AluYb8 (29.1%), and AluYb9 (33.9%). For polymorphic L1s, the L1HS subfamily was the most active, with an activity level of 33.2%. SVA_E and SVA_F subfamilies exhibited similar ratios of 7.1% and 6.8% respectively, and therefore represent the highest activity level for polymorphic SVAs. Polymorphic LTRs and PPSGs showed the lowest activity in the genome, with the highest level of LTR activity represented by the HERV_K subfamily (0.1%), whereas PPSGs have an activity level of 0.2% (Appendix Table S2).

3.2.2 Sex chromosomes have much lower rates of polymorphic ME insertions

The density of polymorphic MEs was measured to examine the distribution pattern exhibited in the human genome. This was completed by obtaining the total number of polymorphic MEs per 10 million bp (Mbp) of non-gapped chromosome sequences, as well as measuring the ratio of each type of polymorphic ME among all MEs in each chromosome. As shown in Appendix Figure S1 and Appendix Table S3, with all ME types combined, the polymorphic ME density varies across each chromosome, with chromosome X and Y showing the lowest density (8.7 copies/10Mbp and 5.3 copies/10Mbp, respectively), which is approximately 2 to 3 times lower than the genome average (~15 copies/10Mbp). The pattern among polymorphic MEs are seen to be the complete opposite for human-specific (HS) MEs, where the Y chromosome appears to be a hot target for all types of MEs except for SVAs (Figure 7) (Tang et al., 2018).

Among the individual types of MEs, the pattern of distribution is variable (Figure 8). Polymorphic Alus showed a more or less homogenous density among autosomal chromosomes, with chromosome 22 having the lowest density (~8 copies/10Mbp), which is approximately 1.5 times lower than the autosomal average. Both X and Y chromosomes showed the lowest density being about 2 to 3 times lower than the genome average (12.4 copies/10Mbp). Unlike what is seen in polymorphic Alus, L1s showed a variable distribution among chromosomes, primarily chromosome Y, which has no polymorphic L1 copies present. Polymorphic SVAs showed relatively similar densities among all chromosomes, with chromosome 17, 19, 20, and 22 having the highest density, more than 3 times greater than the genome average (0.42 copies/10Mbp). The lowest densities for polymorphic SVAs are seen among chromosome X and Y. Polymorphic LTRs have very low densities among all chromosomes, but unlike Alus, L1s, and SVAs, chromosome Y has one of the highest densities, approximately 4 times greater than the genome average (0.11 copies/10Mbp). Similar trends in chromosome Y are seen for the family of LTRs among the group of HS-MEs (Tang et al., 2018). Among the autosomes, chromosome 22 shows the highest density of LTRs which is 5 times greater than the genome average. Polymorphic PPSGs show a variable distribution among all chromosomes, with no copies present in either chromosome X or Y. The highest density of polymorphic PPSGs are seen in chromosome 11, 15, and 17, which are about 4 times greater than the genome average (Figure 8). Overall, chromosome X and Y are the least preferred target site for polymorphic Alus, L1s, SVAs, and

PPSGs, although chromosome Y seems to be a target for polymorphic LTRs. Among autosomes, chromosome 6 seems to have the highest density of polymorphic MEs, while other chromosomes have relatively similar density. The levels of polymorphic ME density among chromosomes seem to correlate with gene densities in some specific cases. For instance, chromosome 13 displays a very low gene density (14.94 genes/Mbp), yet a very high polymorphic ME density (18.57 copies/10Mbp). In contrast, chromosome 22 displays a very high gene density (36.19 genes/Mbp) but exhibits a low polymorphic ME density (11.49 copies/10Mbp) (Appendix Table S3). However, this phenomenon cannot be used to explain the ME density patterns among all chromosomes as there are some cases, such as chromosome X, where both the gene and polymorphic ME densities are measured to be low (Appendix Table S3).

A map representing the distribution of all 4400 polymorphic MEs within individual chromosomes showing that all polymorphic MEs were characterized in all autosomal and sex chromosomes was generated (Figure 9). As previously observed, chromosome Y contained the least number of polymorphic MEs, however it contains the highest density of HS-MEs. In contrast, chromosome 21 is the only chromosome that lacks the presence of SVAs, LTRs, or PPSGs. Unequal distribution of polymorphic MEs is seen among various chromosomes displayed by cluster formation of specific ME types (Figure 9). As a result, polymorphic MEs do not exhibit even distribution within individual chromosomes and across the chromosomes. More specifically, based on Monte Carlo simulation, the observed number of polymorphic MEs across chromosomes 1, X, and Y were significantly lower than expected based on random distribution, whereas for chromosomes 4, 5, 6, 7, 13, the observed numbers were significantly higher than the expected number (Appendix Table S4).

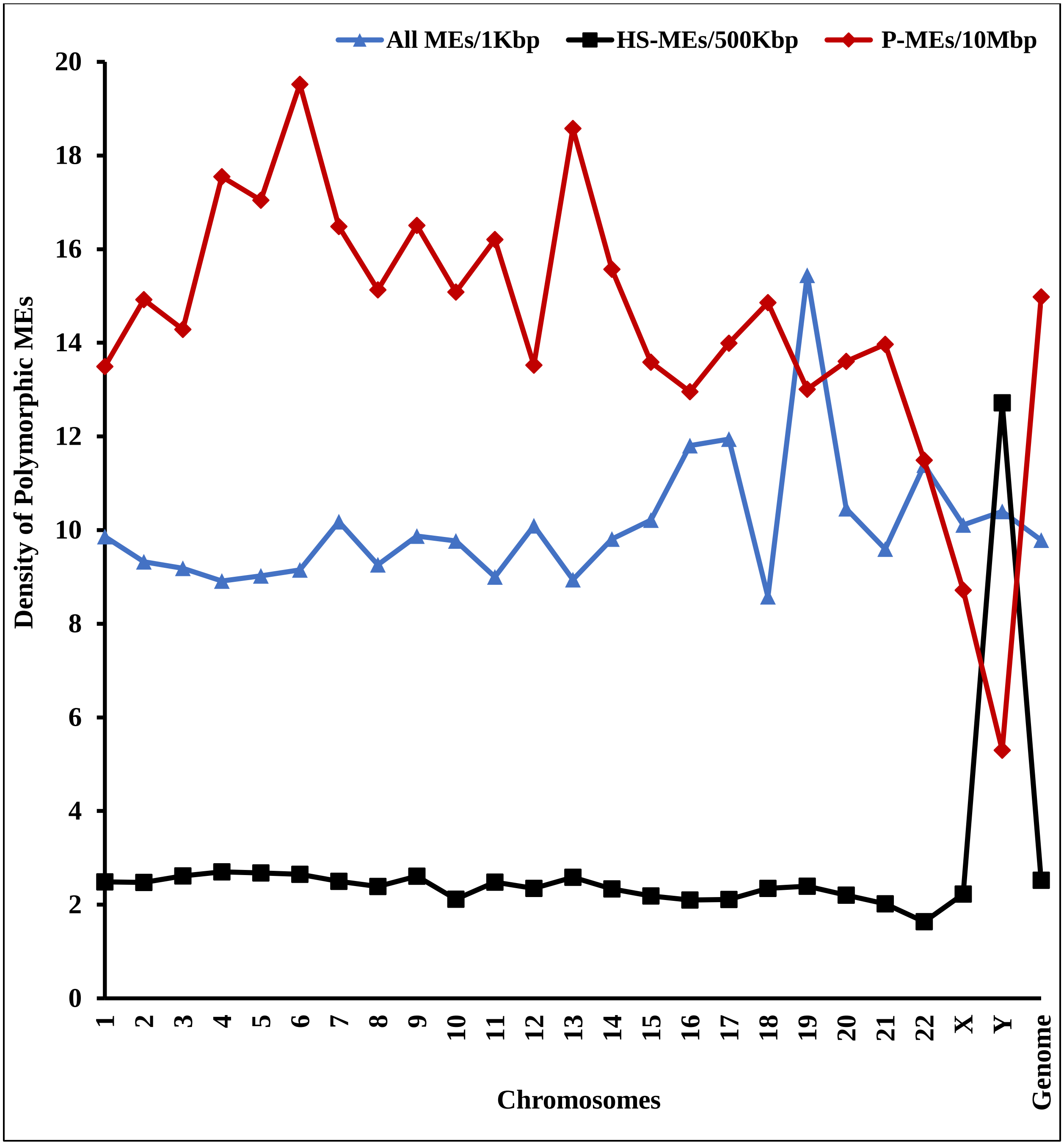

Figure 7. Densities of polymorphic MEs (P-MEs) among chromosomes. Line plots showing the differences in distribution among all MEs, human-specific MEs (HS-MEs), and all polymorphic MEs (P-MEs) across all human chromosomes.

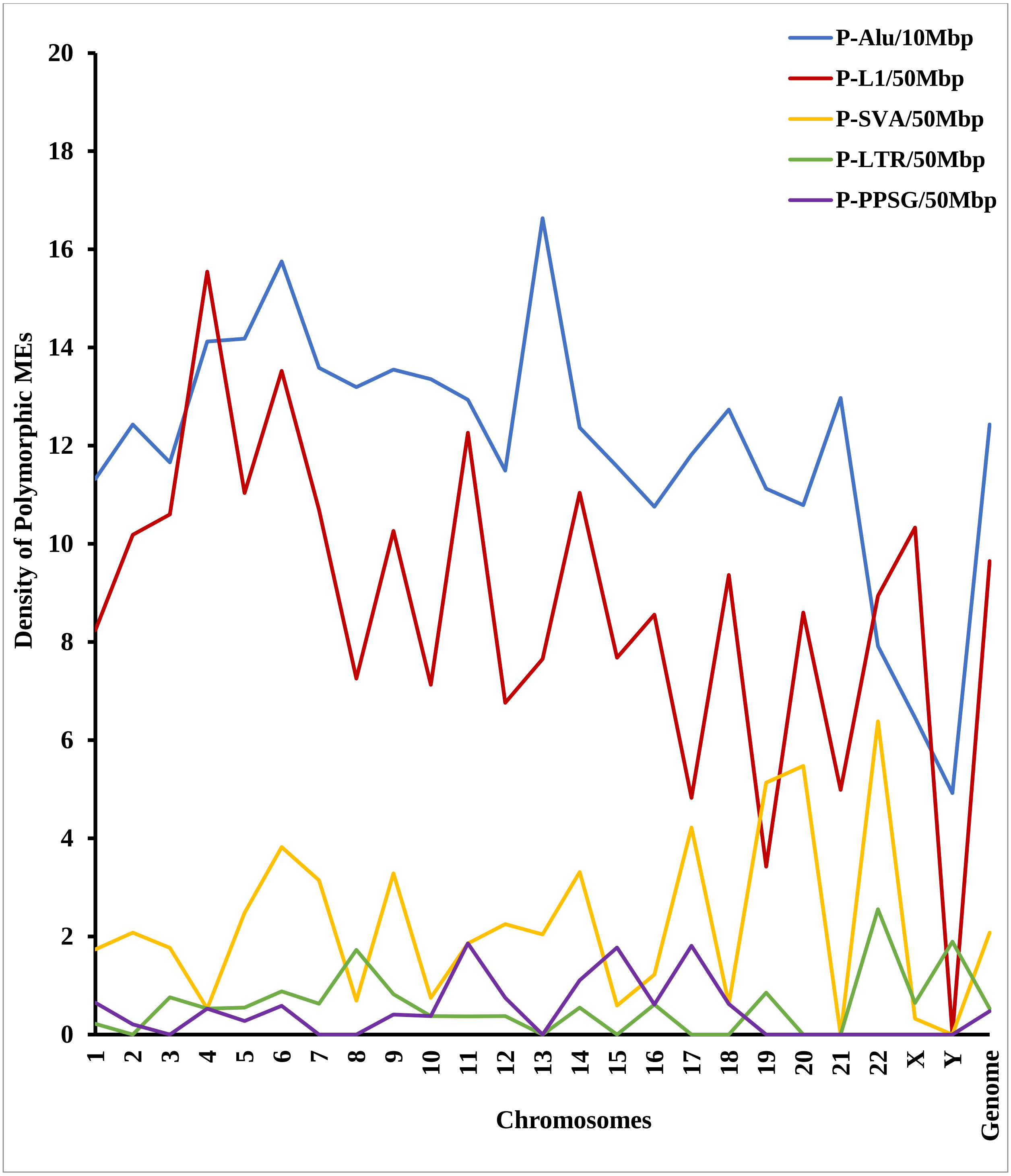

Figure 8. Densities of polymorphic MEs by ME class among chromosomes. Line plots showing the density for each type of polymorphic MEs across all human chromosomes.

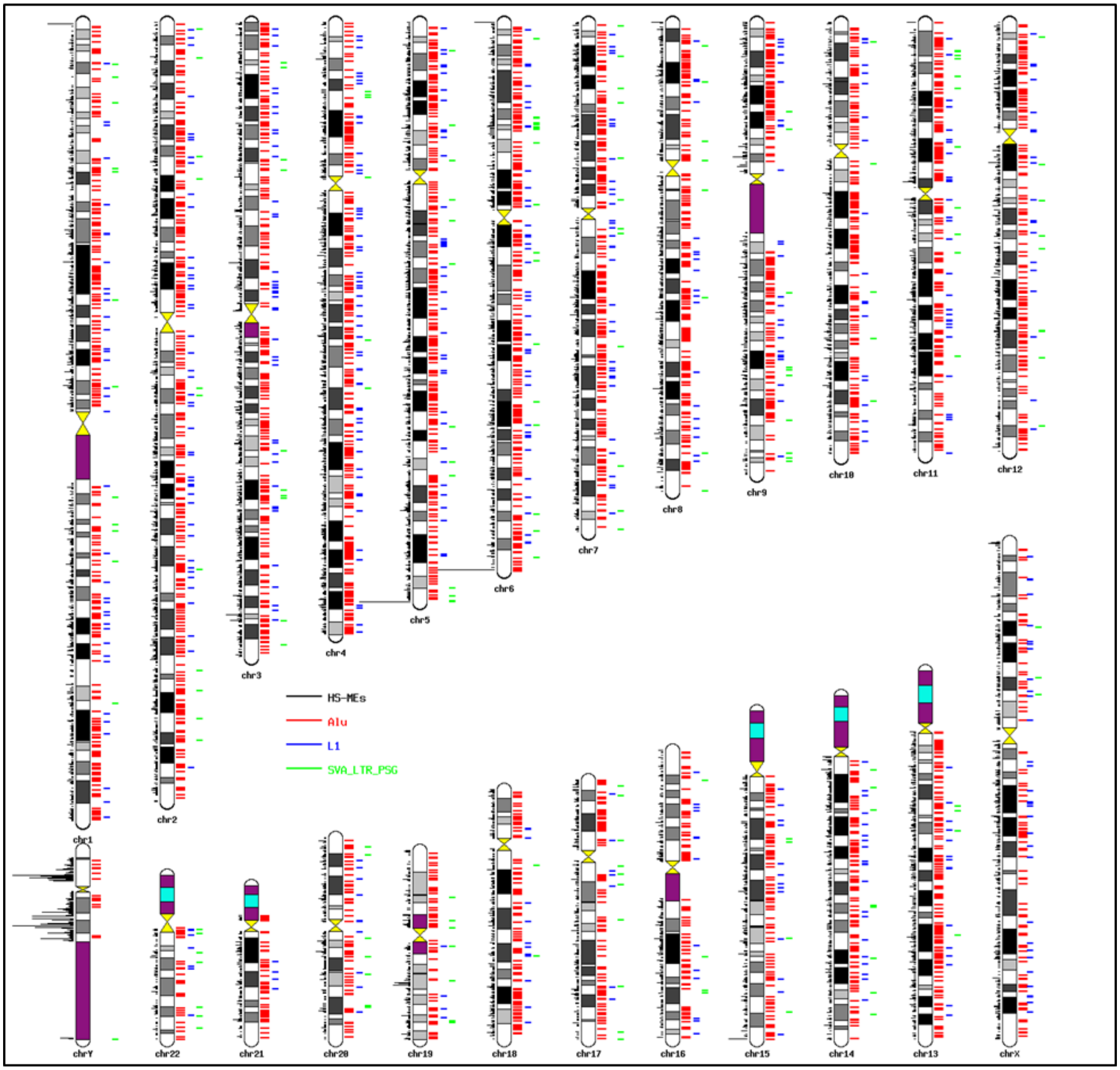

Figure 9. Human chromosome ideograms showing the distribution patterns of polymorphic MEs. Each coloured horizontal tick on the sides of the chromosomes represent different types of MEs. Positions of human-specific MEs (HS-MEs) are represented by the bar plots on the left side of each chromosome, positions of polymorphic Alus are represented by red lines, polymorphic L1s are represented as blue lines, and the collected combination of polymorphic SVAs, LTRs, and PPSGs are represented as green lines, all on the right side of each chromosome.

3.3 Patterns of polymorphic ME distribution based on *in silico* genotyping of polymorphic MEs

3.3.1 Distribution patterns of polymorphic MEs in human populations

To examine the distribution patterns of all polymorphic MEs across human populations, the insertion allele frequencies and distribution patterns were analyzed for 28 populations from 5 continental groups, including Africa, East Asia, South Asia, Europe, and Admixed America. As seen in Figure 10A, when the insertion allele frequencies are analyzed, both at the individual population level and all populations as a whole, the largest percentage of polymorphic MEs have an allele frequency of 50%. This correlates with the majority of individuals (67%) representing a heterozygous genotype for polymorphic MEs (Figure 10B). About 15% of polymorphic MEs exhibited very low insertion allele frequencies (<5%), which may represent very new ME insertions that may even be population-specific. Due to the polymorphic nature of these MEs, it is expected that these entries are relatively recent and therefore have not yet reached a high frequency within and between populations. This is enforced by our results where we see a minimal number of entries with insertion allele frequencies greater than 90% (Figure 10A).

As a further analysis, allele frequencies were organized into five arbitrary frequency categories as very low (<=1%), low (>1 to 9%), intermediate (>9 to 49%), high (>49 to 74%), and very high (>=75%). Among all polymorphic MEs, majority of the entries fall at intermediate allele frequencies (44.06%), followed by high (35.10%), low (11.06%), very low (7.38%), and very high (2.40%) (Figure 10C). The reason for the high frequency of MEs is currently unknown, however the pattern is consistent across ME types and populations.

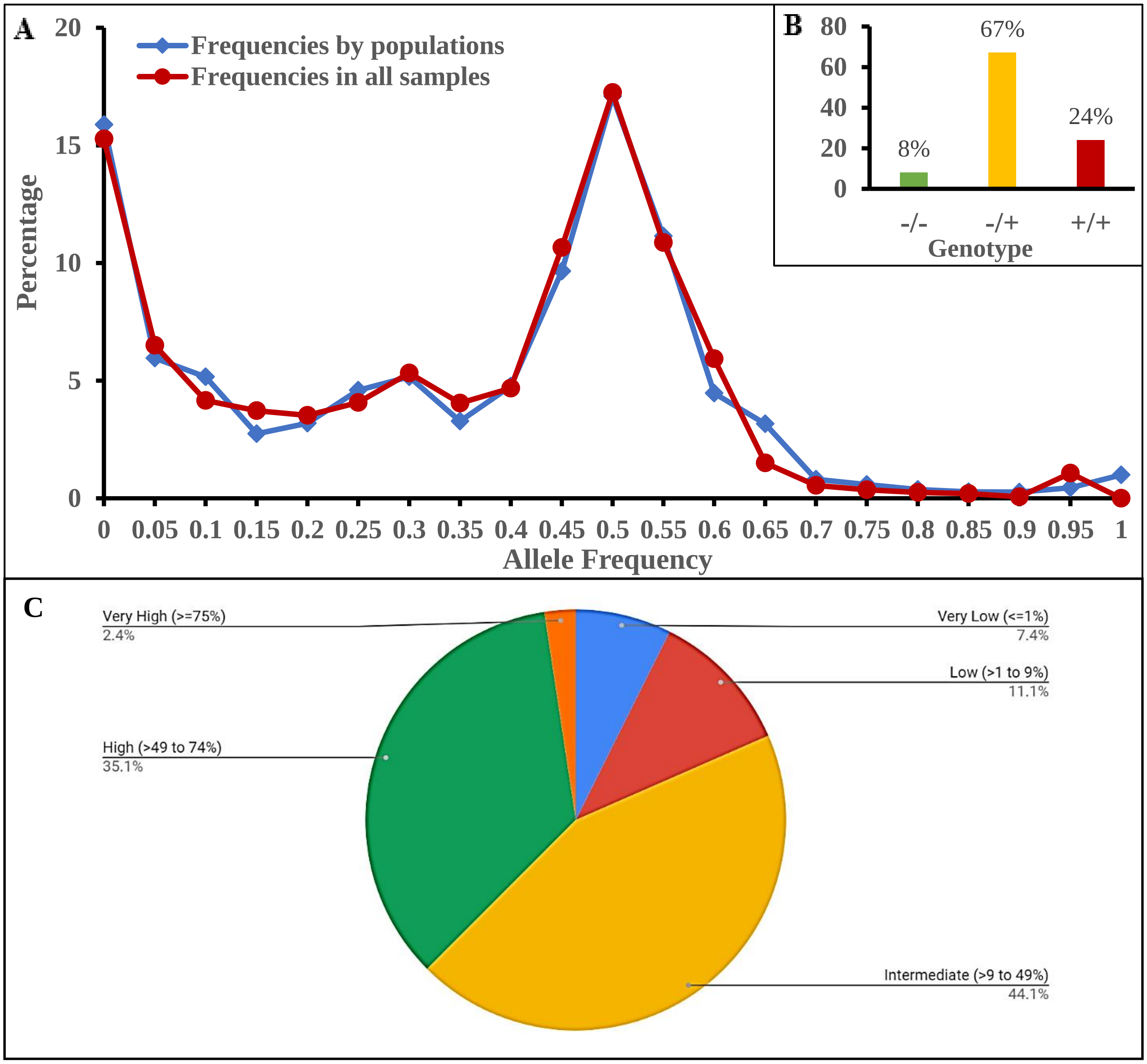

Figure 10. Allele frequency distribution of polymorphic MEs. A) Line plot representing the full continuous allele frequency spectrum for all polymorphic MEs. The spectrum was measured with two methods: Blue – ME insertion allele frequencies analyzed by population; Red – ME insertion allele frequencies combined across all populations and analyzed as one sample. B) Bar plot representing the percentage of individuals displaying each of the 3 genotypes: -/- (homozygous for the pre-integration allele); -/+ (heterozygous for both the pre-integration and insertion allele); +/+ (homozygous for the insertion allele). C) Pie chart displaying the percentage of MEs classified into five allele frequency groups.

The distribution pattern was analyzed and compared between reference and non-reference MEs. As seen in Figure 11, the vast majority of polymorphic reference MEs are found at high frequencies while there are a higher number of polymorphic non-reference MEs with low allele frequencies. More specifically, reference MEs display the highest peak at an insertion allele frequency of 50% compared to non-reference MEs, which are peaking at an insertion allele frequency less than 5% (Figure 11). The strong bias of non-reference polymorphic MEs towards the low allele frequencies compared to the reference entries is expected, as the former group are more likely to be rare variants for being absent in the reference genomes. The high similarity of patterns for the reference and all polymorphic MEs at the high peak are due to a smaller percentage of non-reference polymorphic MEs than the reference entries and the latter being concentrated at the 50% allele frequency.

Allele frequency pattern was also examined and compared across ME types. Among the five ME types, Alus, L1s, and SVAs illustrated similar patterns across the allele frequency spectrum, while the trends seen in LTRs and PPSGs were more alike (Figure 12). All ME types, except for PPSGs, have the highest percentage of entries at an insertion allele frequency of 50%. The largest percentage of PPSGs are seen at an allele frequency of 45%. At the very low allele frequencies (<5%), Alus have the greatest percentage of entries, followed by L1s, SVAs, LTRs, and PPSGs, indicating Alus having the highest activity leading to more recent insertions. It is a bit unexpected to see that SVA has a lower percentage at this frequency range than L1, as SVAs are very young and shown to be very active. Among all ME types, LTRs exhibit the largest percentage of entries with very high allele frequency (>95%) , followed by L1s, SVAs, Alus, and PPSGs (Figure 11), indicating that LTRs have the lowest activity leading to a high percentage of old insertions that are near being fixed.

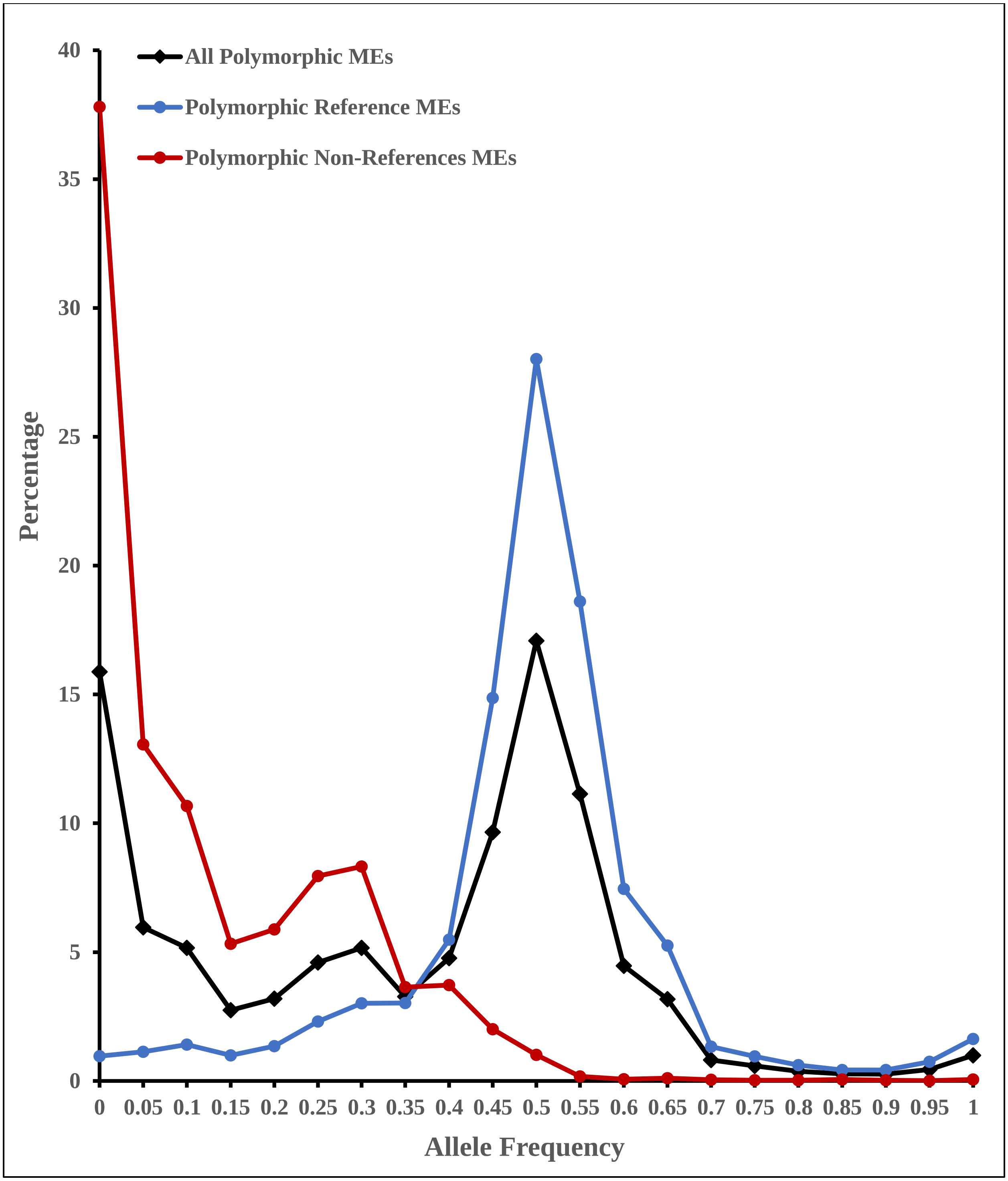

Figure 11. Allele frequency distribution between reference and non-reference polymorphic MEs. Line plots show the percentage of MEs across different insertion allele frequencies for the reference, non-reference, and all polymorphic MEs.

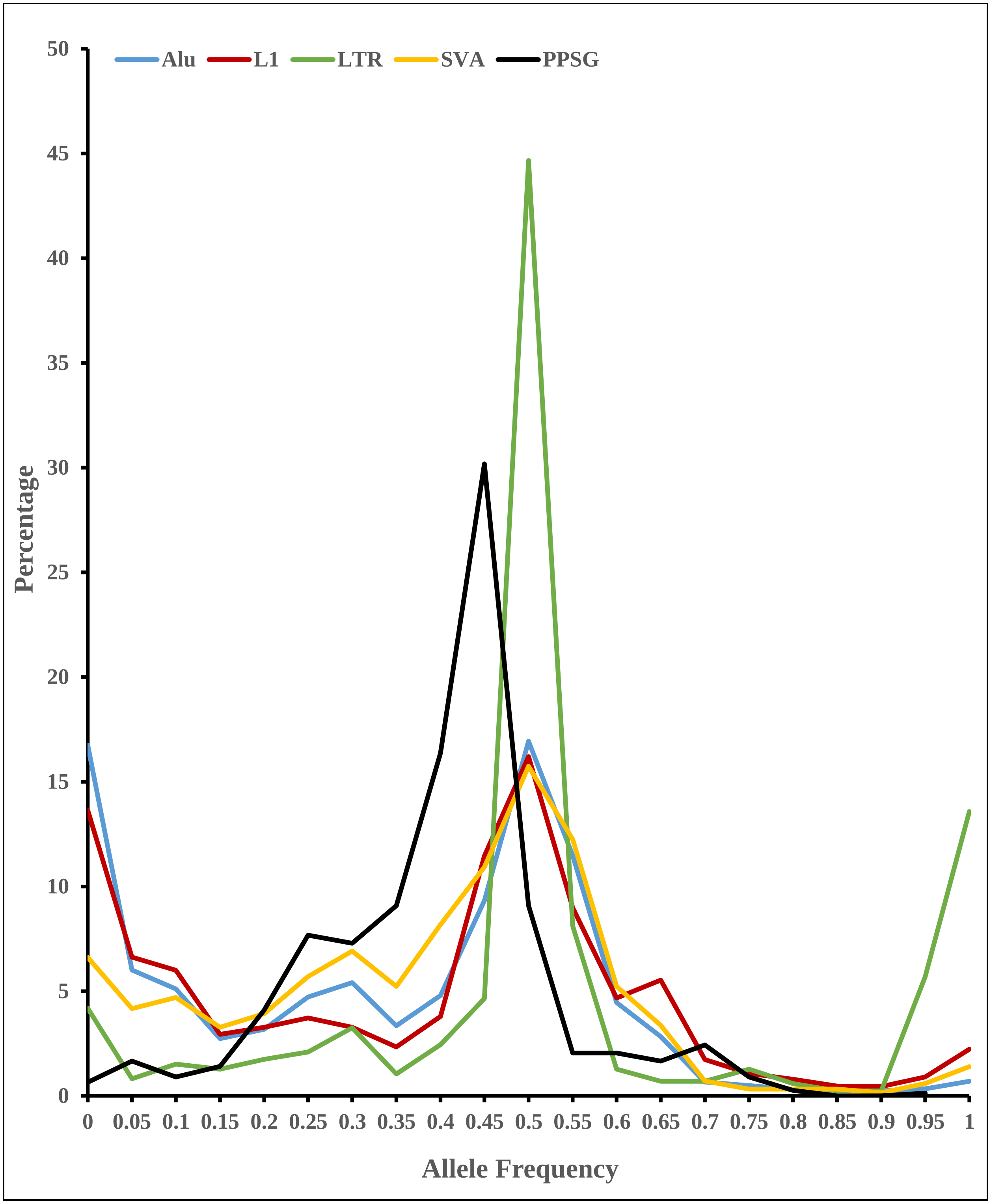

Figure 12. Allele frequency distribution across all five types of polymorphic MEs. Line plots show the percentage of each type of ME across different insertion allele frequencies.

The allele frequency distribution was also analyzed and compared among different continental groups. The overall distribution of the insertion allele frequencies showed a consistent pattern among all five continental groups (Figure 13). As seen among all polymorphic MEs, the highest peak across all continental groups is at an insertion allele frequency of 50%. Although South Asian populations also peak at 50%, a large percent of entries are seen between 45% to 55% insertion allele frequencies. A minor peak is also seen at an insertion allele frequency less than 5% across all continental groups (Figure 13). Overall, the levels of genetic diversity associated with polymorphic MEs are relatively similar for all five continental groups.

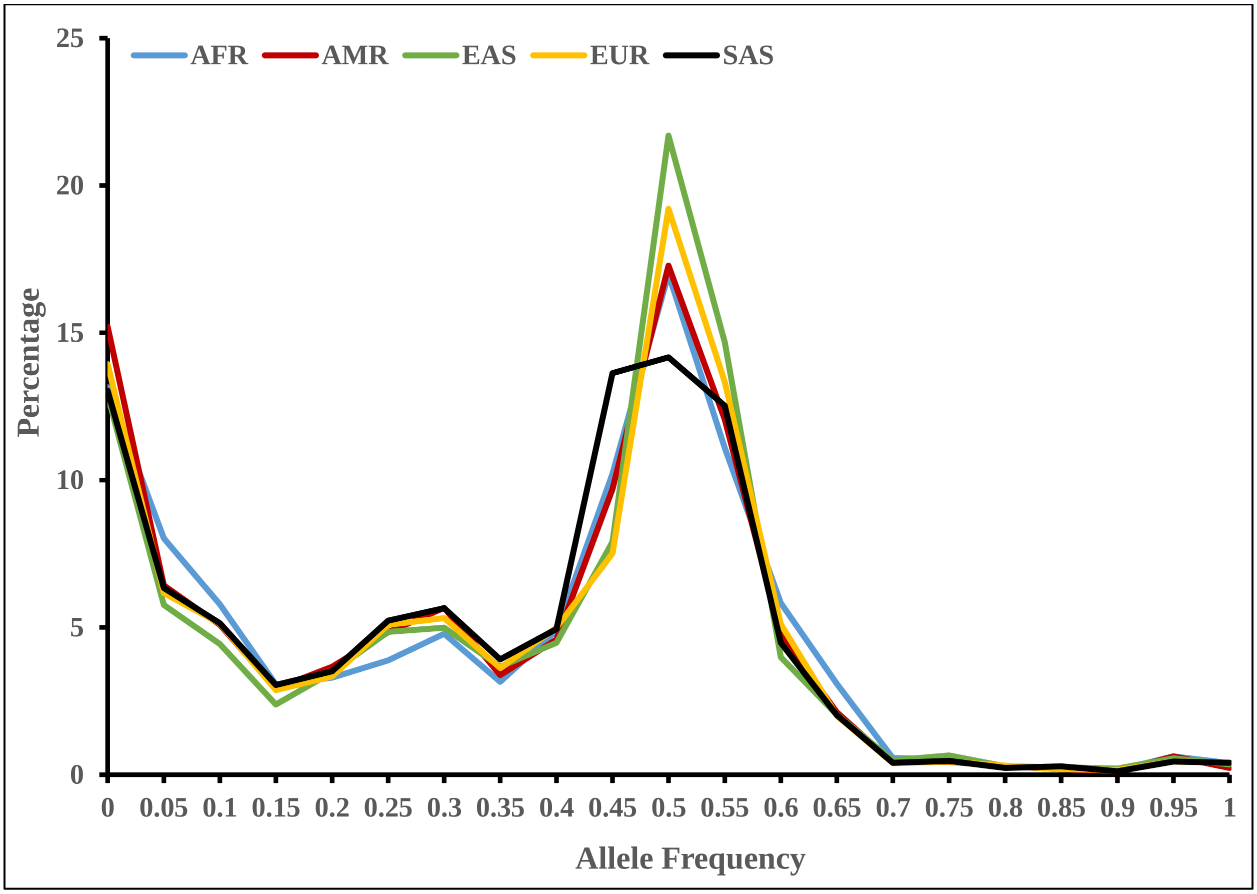

Figure 13. Allele frequency distribution of polymorphic MEs across all five continental groups. Line plots show the percentage of MEs in each continental group across different insertion allele frequencies.

Among the 4400 polymorphic MEs, one Alu entry was located on an alternative chromosome assembly, specifically on chromosome 6 cox haplotype 2, which is a more discriminative state of the chromosomal region. The sequences flanking the Alu insertion are not seen in the reference genome, indicating an uncommon haplotype region. As seen in Figure 14, the allele frequency distribution for this locus exhibits a more restricted distribution. This polymorphic ME seems to be present at a greater level in African, European, and Admixed American populations, while being virtually absent in East Asian and South Asian populations (Figure 14A).

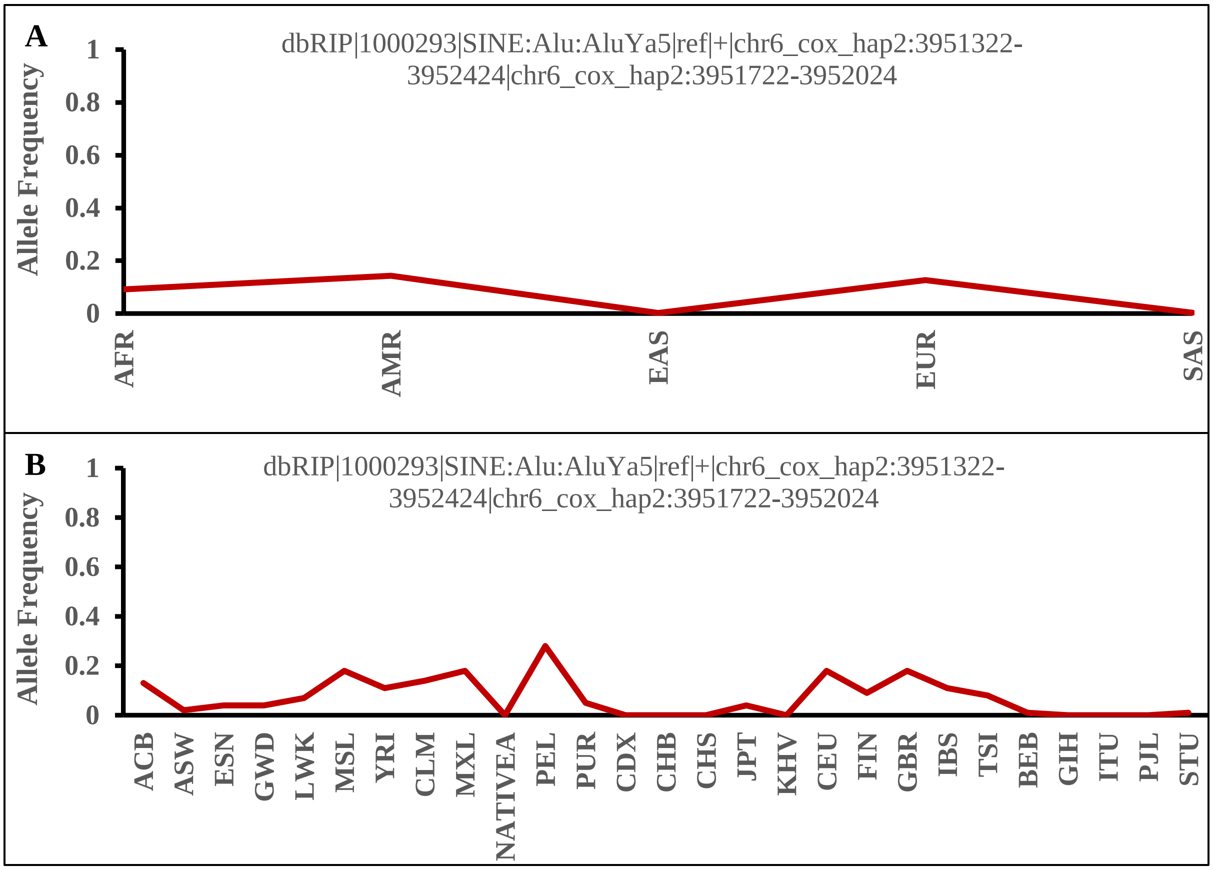

Figure 14. Allele frequency distribution of polymorphic ME on alternative chromosome 6. A) Line plot representing the differences of allele frequencies across continental groups. B) Line plot representing the differences of allele frequencies across individual human populations.

3.3.2 Some polymorphism of MEs are not human-specific, but shared with other primate genomes

It is generally understood that MEs which are polymorphic in humans represent very young insertions, which are most likely to happen after the separation of human and chimpanzee from their common ancestor. In other words, MEs that are shared between human and chimpanzee are much older and are very likely to be fixed in the populations of the two species. From the list of 4400 polymorphic MEs, 11 MEs were found to be also present in the chimpanzee genome, one of which was also present in the orangutan genome. Compared to all polymorphic MEs, 9 of these ME loci have greater insertion allele frequencies (Table 8). One explanation for this observation is that these ME insertions never reached fixation state in the populations. Very unexpectedly, the single remaining ME insertion on this list (chrX:31098368-31098369) has a very low insertion allele frequency (<1%) across all continental groups, and we do not have a good explanation for its presence.

Table 8. Polymorphic ME loci present in primate genomes

	Continental Groups					All Populations
Locus	AFR	AMR	EAS	EUR	SAS	Average
dbRIP|1003101|Alu|non-ref|chr10:53209407-53209408	0.421	0.229	0.207	0.351	0.268	0.295
dbRIP|5000029|PPSG|non-ref|chr17:32673415-32673416	0.508	0.512	0.513	0.510	0.510	0.511
dbRIP|4000017|LTR|ref|chr4:99096289-99096684	0.508	0.514	0.514	0.511	0.516	0.513
dbRIP|2000646|L1|ref|chr15:87639296-87639682	0.518	0.525	0.531	0.529	0.513	0.523
dbRIP|1002750|Alu|ref|chr9:99064831-99065205	0.508	0.513	0.508	0.512	0.506	0.510
dbRIP|2000347|L1|non-ref|chrX:31098368-31098369	0.006	0.004	0.003	0.003	0.003	0.004
dbRIP|1000167|Alu|ref|chr5:133762623-133762924	0.522	0.547	0.559	0.560	0.554	0.548
dbRIP|1000023|Alu|ref|chr17:29662701-29663008	0.536	0.542	0.553	0.549	0.542	0.545
dbRIP|1000838|Alu|ref|chr2:122056860-122057157	0.990	0.995	0.979	0.992	0.981	0.987
dbRIP|1001346|Alu|ref|chr2:122056860-122057157	0.987	0.969	0.973	0.984	0.988	0.980
dbRIP|1002824|Alu|ref|chr15:44037656-44037972	0.495	0.510	0.521	0.531	0.492	0.510
All other polymorphic MEs	0.366	0.356	0.373	0.360	0.355	0.362

3.4 Genetic relationships of human populations based on polymorphic MEs

Due to MEs' identical-by-decent nature, they are considered to be better genetic markers for evolutionary studies. The genotypes of the 4400 polymorphic MEs were used as markers to analyze the evolutionary relationships among human individuals and populations. The genotype data for all polymorphic ME loci were used to cluster individuals or populations using several methods.

3.4.1 Group relationship of individuals

To obtain an illustration of the evolutionary relationship among humans and populations, a principle component analysis (PCA) was performed. The PCA plot revealed a strong clustering of individuals by continent and population. On a continental level, a large cluster representing individuals from Africa is clearly distinct (bottom cluster in Figure 15). Within the African cluster, individuals from the Bushmen population can be found which demonstrates the relationship between the African and Bushmen population that is currently known. A second large cluster representing individuals from a combination of East Asia, South Asia, Europe, and Admixed America is also visible (top clusters in Figure 15). The area of the African cluster is more than that for all other continental groups combined, indicating that African populations are more genetically diverse than the other populations analyzed. Two small clusters of individuals from East Asia and South Asia are also present (Figure 15).

On an individual population level, groups of specific populations within the clusters can be identified (Figure 16). It appears that individuals from single populations are more or less grouped together within the clusters. This includes the two small clusters representing East Asia and South Asia, which appear to be clusters of single populations, involving Han Chinese from Beijing, China (CHB) and Indian Telugu from the UK (ITU), respectively (Figure 16).

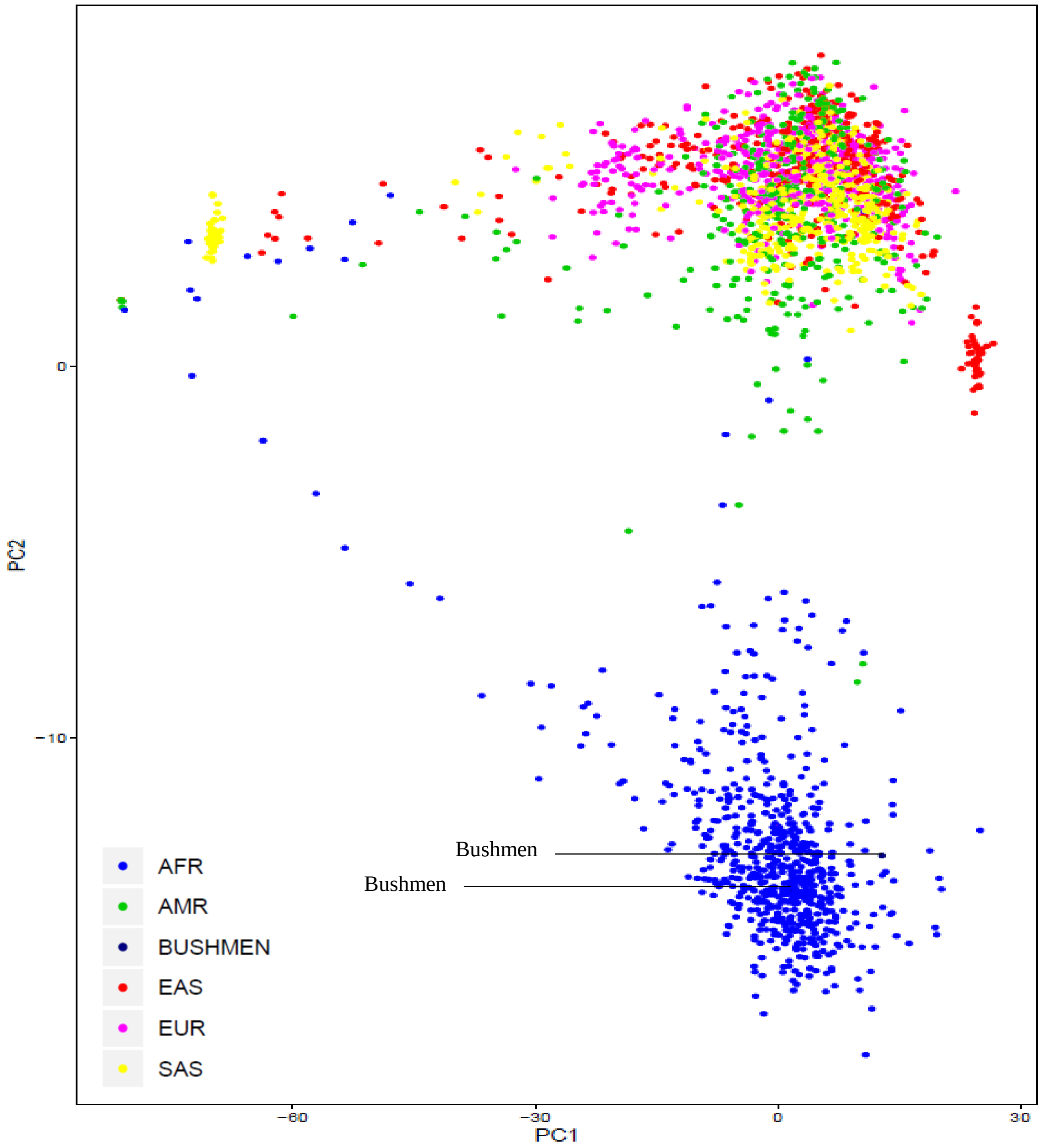

Figure 15. Evolutionary relationships among human continental groups based on polymorphic ME genotypes. PCA plot showing polymorphic ME genotyped-based cluster formation of individuals from five continental groups, as well as Bushmen individuals. Continental groups are color coded as shown in the figure legend; Blue – Africa (AFR); Green – Admixed America (AMR); Black – Bushmen; Red – East Asia (EAS); Purple - Europe (EUR); Yellow – South Asia (SAS).

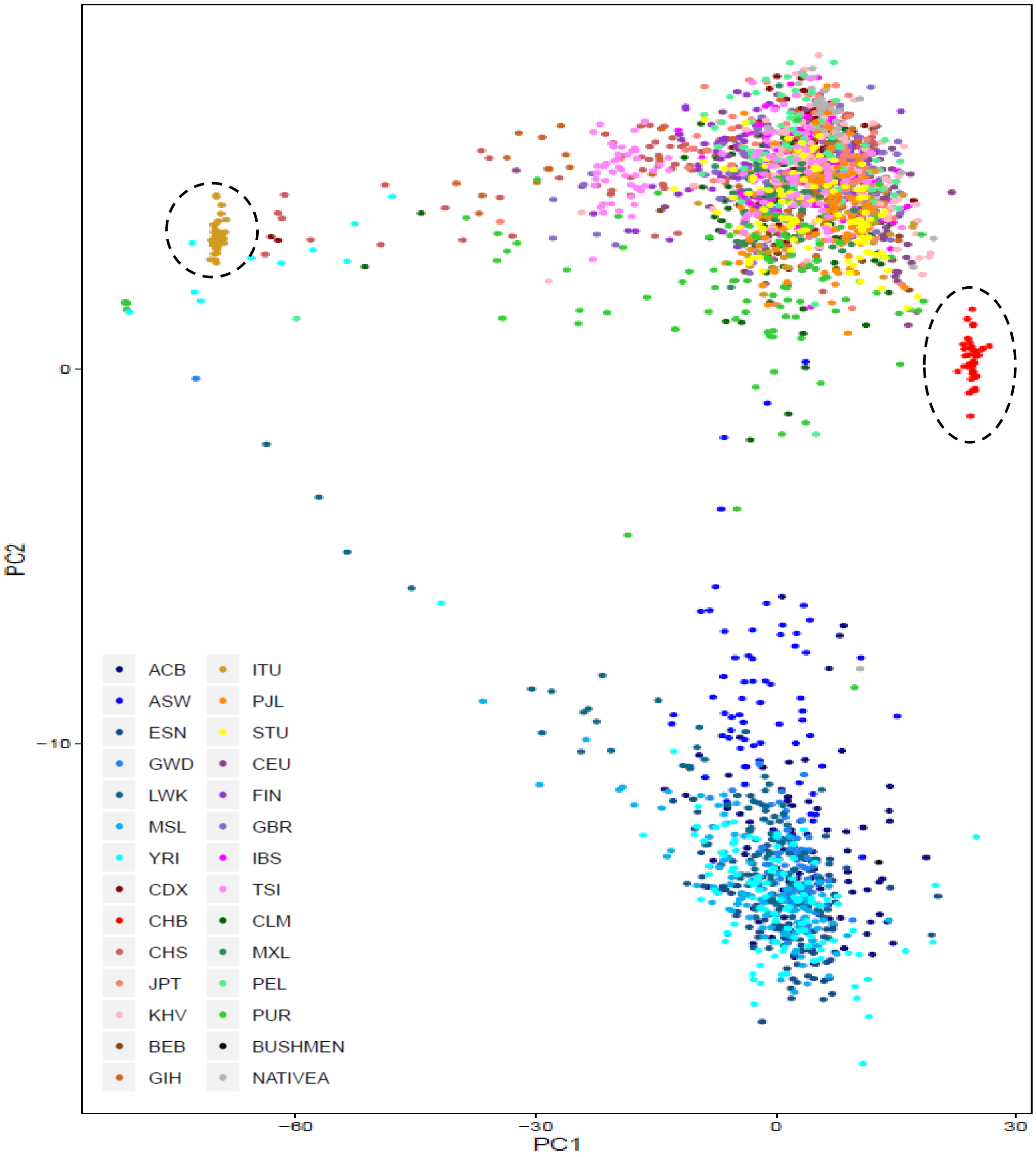

Figure 16. Genetic relationships among human populations based on polymorphic ME genotypes. PCA plot showing polymorphic ME genotyped-based cluster formation of individuals from 28 human populations, including Bushmen and Native American populations. Population groups are color coded as shown in the figure legend. Two small clusters representing single populations, involving Han Chinese from Beijing, China (CHB) and Indian Telugu from the UK (ITU) are highlighted by circles in the figure.

To better illustrate the genetic relationships among individuals, a 3-dimensional PCA plot was generated (Figure 17A). The evolutionary relationships revealed by this analysis specify the groups of individuals within the mixture cluster. Within this large cluster exists groups of miniature clusters representing East Asia, South Asia, Europe, and Admixed America (Figure 17B).

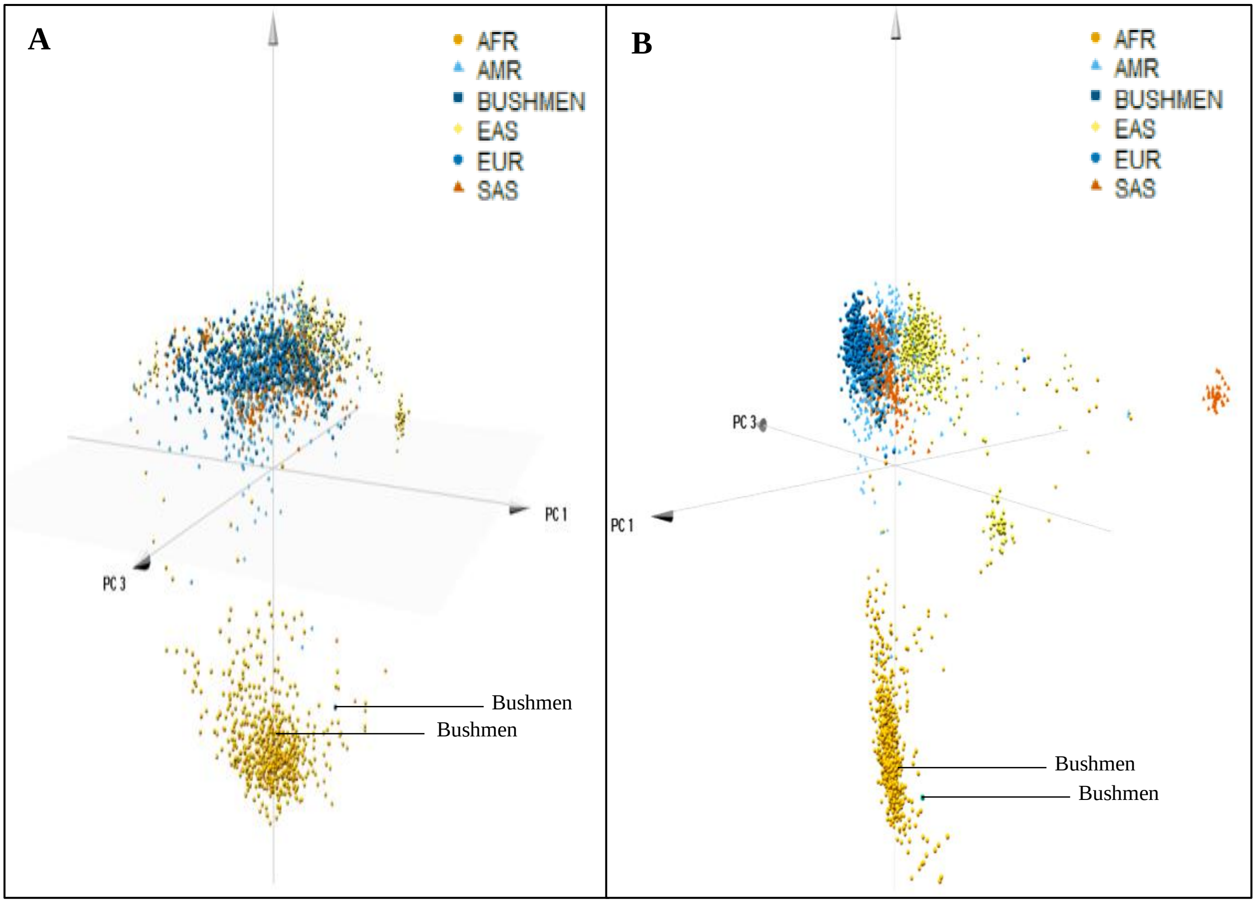

Figure 17. An expanded view of the evolutionary relationships among human continental groups based on polymorphic ME genotypes. A) 3D PCA plot showing polymorphic ME genotyped-based cluster formation of individuals from five continental groups, as well as Bushmen individuals. B) Rotated snapshot of 3D PCA plot displaying the groups of individuals within the mixture cluster. Continental groups are color coded as shown in the figure legend.

A phylogenetic tree was generated to further illustrate and reconstruct the relationships of individuals between human populations based on ME polymorphisms (Figure 18). This was completed by transforming individual genotype data into pseudo-sequences and performing a sequence alignment to outline the evolutionary relationships based on the presence or absence patterns of the polymorphic MEs. From the phylogenetic analysis, all individual populations were shown to cluster within their continental groups, which implies that individuals within similar geographic locations share closer genetic relationships (Figure 18).

The basal part of the tree represents the ancestral lineage, which we consider to be a sample containing the pre-integration state of all polymorphic MEs. As we move from the root to the tips, the evolutionary relationship can be distinguished based on the genetic distance between populations. The transition from the ancestor to African populations, including Bushmen individuals, illustrate the close genetic relationship that is known among these groups. Among all 7 individual African populations, the Yoruba group from Ibadan, Nigeria (YRI) located in West Africa, seems to be the closest to the ancestor sample. Based on the cluster formation of individual groups within the evolutionary tree, South Asian populations seem to be the closest to the African populations, whereas East Asian populations are seen to be the furthest from the root (Figure 18). As clusters are formed on the basis of similarities among populations, it is apparent that there are populations within continental groups that have a closer relationship among each other than others. As expected, due to the timeline of human population divergence, European and Admixed American populations share a close evolutionary relationship (Figure 18).

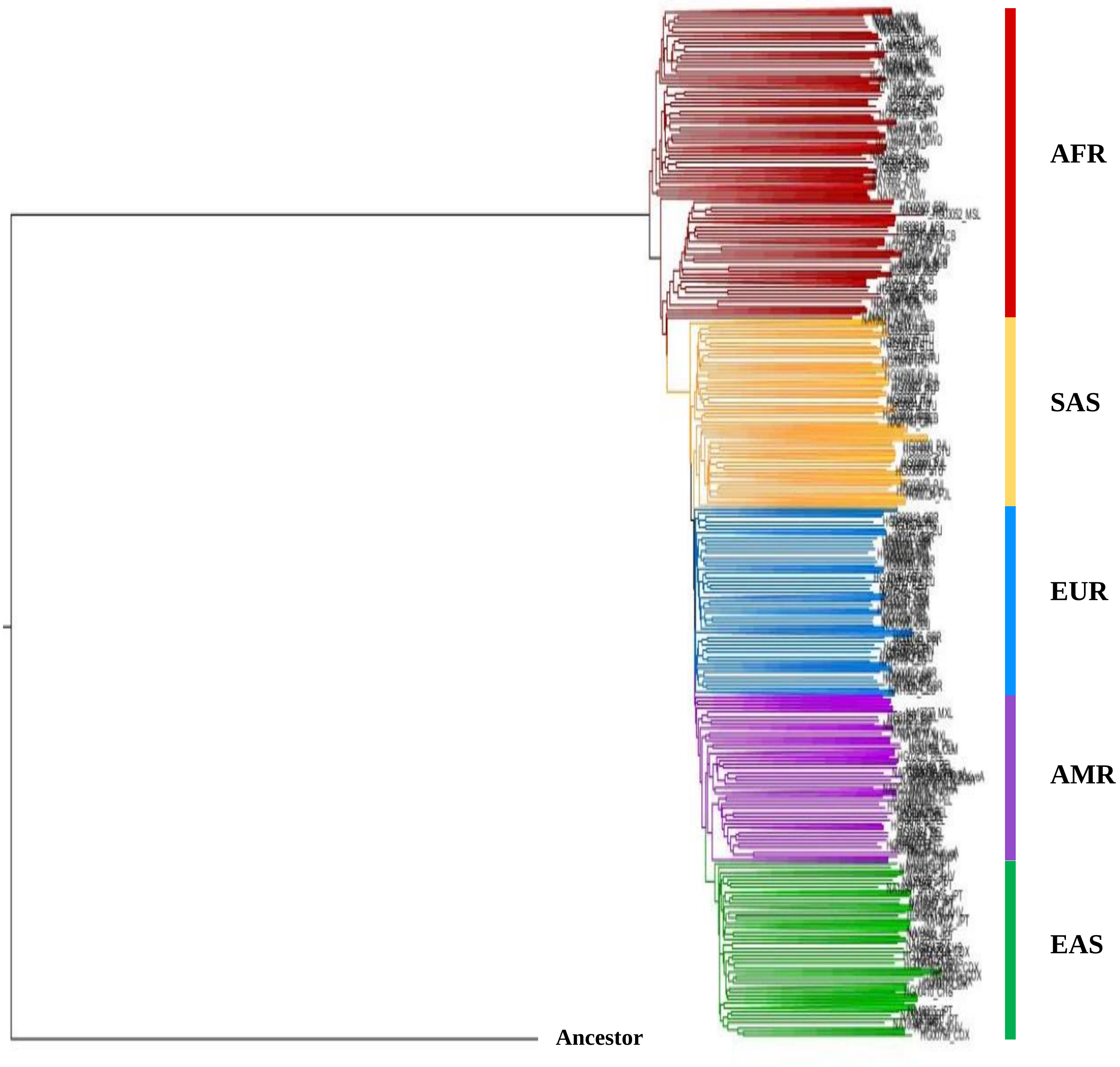

Figure 18. Evolutionary relationships among individuals based on ME polymorphisms. A phylogenetic tree rooted with the ancestor sample, illustrating human genomic diversity through similarities and differences among ME variations. Individual lines are represented by sample ID and population ID. Clusters are colour-coded and represent the five continental groups. Yoruba group from Ibadan, Nigeria (YRI) can be seen closest to the ancestor in the transition from the root to African populations.

3.4.2 Population-specific MEs

We examined geographic differential distribution of polymorphic MEs to identify population-specific MEs, which occurred at certain time points of human population migrations. Despite the similar shapes of allele frequency distributions, some individual polymorphic MEs displayed high levels of geographic differentiation when ME loci were examined individually. Although the loci were not present exclusively to an individual population, many loci were more commonly present or absent in specific groups of populations, relating directly to the population's continental group. For instance, some polymorphic loci were found at low allele frequencies for all populations except all 7 African populations (Figure 19A) while another group of MEs were found at high allele frequencies across all populations except the 7 African populations (Figure 19B). MEs at greater frequencies in both East Asian and European populations, in only South Asian populations, or in only European populations were also present (Figure 19C, D, E).

To obtain a complete visual illustration of potential continental-specific polymorphic ME insertions, a heatmap of ME insertion allele frequencies across 28 human populations was generated (Figure 20A). Hierarchical clustering (HCL) of both populations and ME loci was also performed. This unsupervised clustering algorithm constructs a hierarchical tree which displays clusters of similar populations based on the frequency profiles of all polymorphic MEs and of similar ME loci based on their frequency profile across populations. From this analysis, human populations were clustered together based on continental groups in good agreement with what we know based on other data for these populations. The Bushmen population is closely related to the African populations, and the Native American population are close in distance to the Admixed American group (Figure 20B). Groups of polymorphic MEs more commonly present or absent in specific continental groups were found (Figure 20B). Based on the clustering patterns observed, there are cases of polymorphic MEs in high frequencies shared among either African populations or non-African populations only. There are also clusters of ME polymorphisms that are seen more frequently in both East and South Asian populations compared to others, while there are some entries that seem to be rarely present or completely absent in either East or South Asian populations. There are also several polymorphic MEs that are only present in higher frequencies in European or Admixed American populations (Figure 20B).

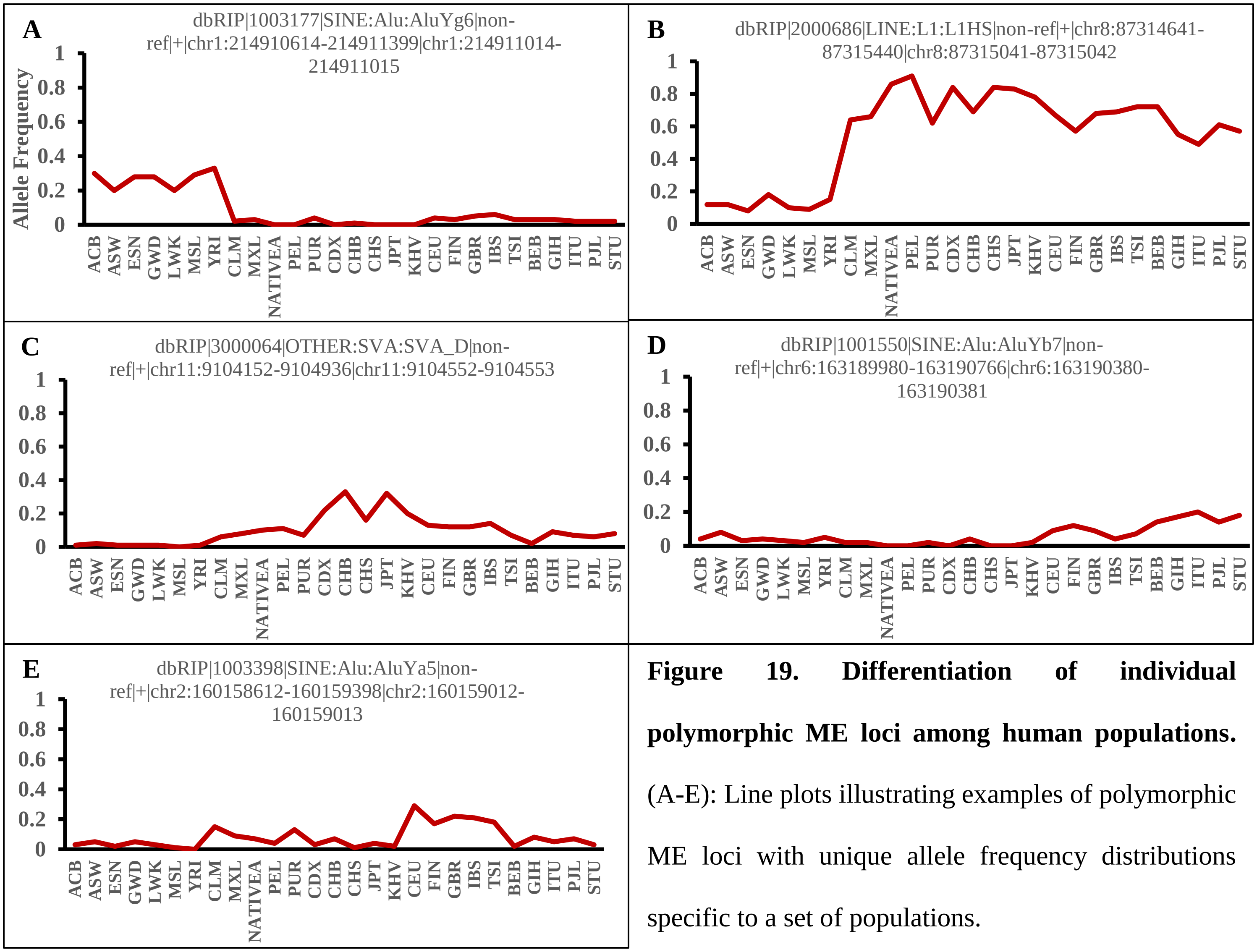

Figure 19. Differentiation of individual polymorphic ME loci among human populations. (A-E): Line plots illustrating examples of polymorphic ME loci with unique allele frequency distributions specific to a set of populations.

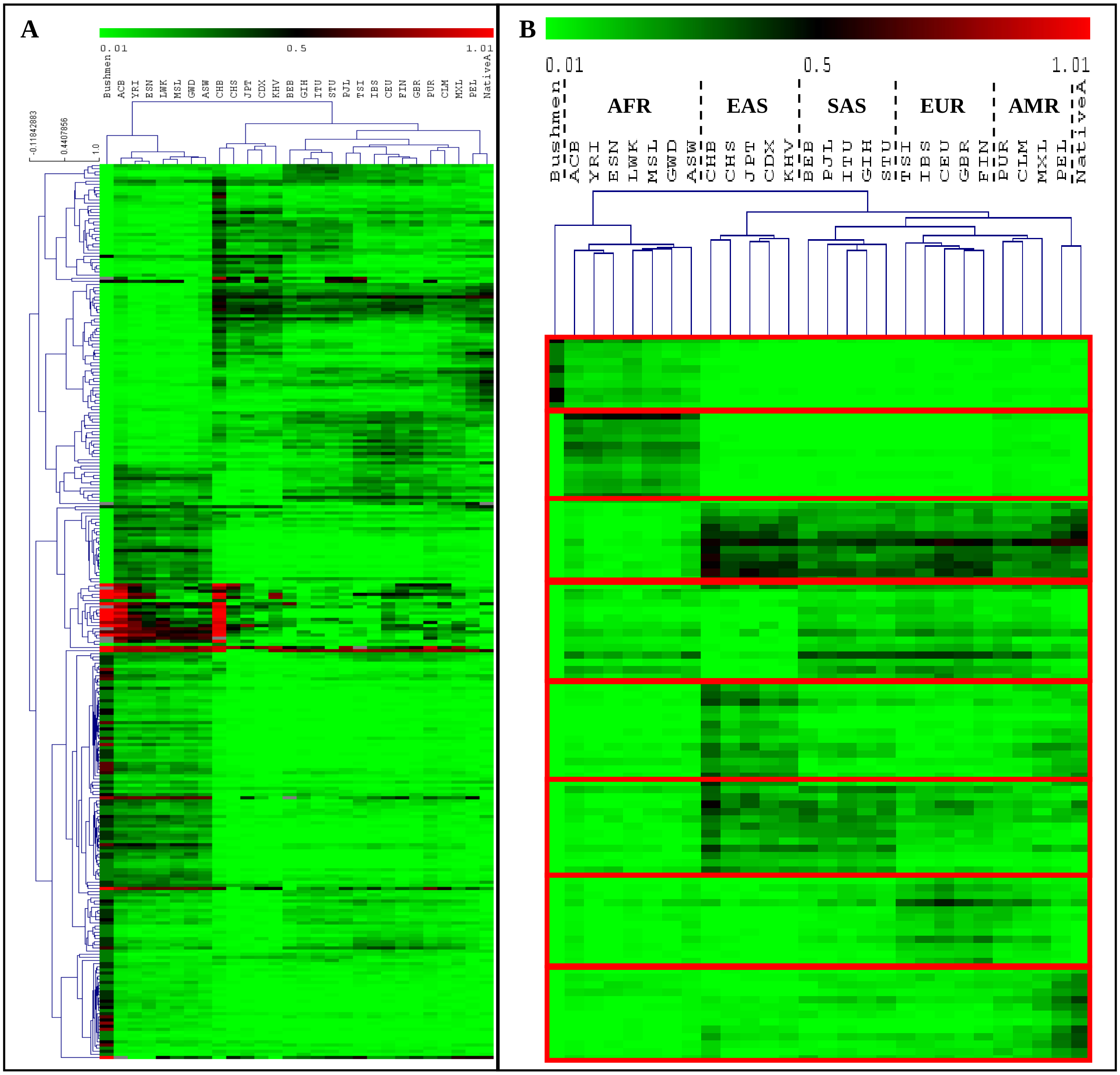

Figure 20. Heatmap representation of polymorphic ME allele frequencies across 28 human populations. A) A complete representation of HCL clustering of populations and polymorphic MEs. Green = low allele frequency, Black = intermediate allele frequency, Red = high allele frequency. B) Specific groups of polymorphic MEs outlined in red boxes to display various patterns of continental specific polymorphic MEs.

Overall, the analysis of allele frequency distributions indicates that there are patterns associated with the insertion of individual polymorphic MEs across human populations. To further characterize the genetic diversity represented by ME polymorphisms, allele frequency distributions were examined between African and non-African populations. The group of non-African populations consist of individual populations representing East Asia, South Asia, Europe, and Admixed America. A direct comparison of the allele frequency spectrums generated for the two population groups displays the large levels of variation between the presence or absence of polymorphic MEs (Figure 21A). When organized by ascending allele frequencies in the African populations, there are occurrences of polymorphic ME entries with higher or lower allele frequencies in the non-African populations, which can be used to identify specific MEs that are commonly found in one population group compared to the other (Figure 21A). To further classify polymorphic MEs that were more commonly present in African populations compared to non-African populations, loci with insertion allele frequency differences of 10%, 20%, and 30% between the two groups were examined (Figure 21B, C, D). The larger the difference in allele frequency between the two groups, the more prominent the insertion is to that population. Overall, 286 polymorphic MEs had 10% greater allele frequencies in non-African populations, 54 MEs with 20% greater allele frequencies, and 11 entries with differences of 30%. In contrast, 333 polymorphic MEs had 10% greater allele frequencies in African populations, 67 MEs with 20% greater allele frequencies, and 13 entries with differences of 30% (Figure 22). The remaining 3636 polymorphic MEs had allele frequency differences of less than 10%. This indicates that African populations have more occurrences of polymorphic ME insertions compared to non-African human populations. However, the vast majority of polymorphic MEs are found at relatively similar frequencies among African and non-African populations, suggesting that population-specific MEs represent a relative smaller proportion among the 4400 polymorphic MEs and thus the majority of ME polymorphism came from the populations before the exodus out of Africa.

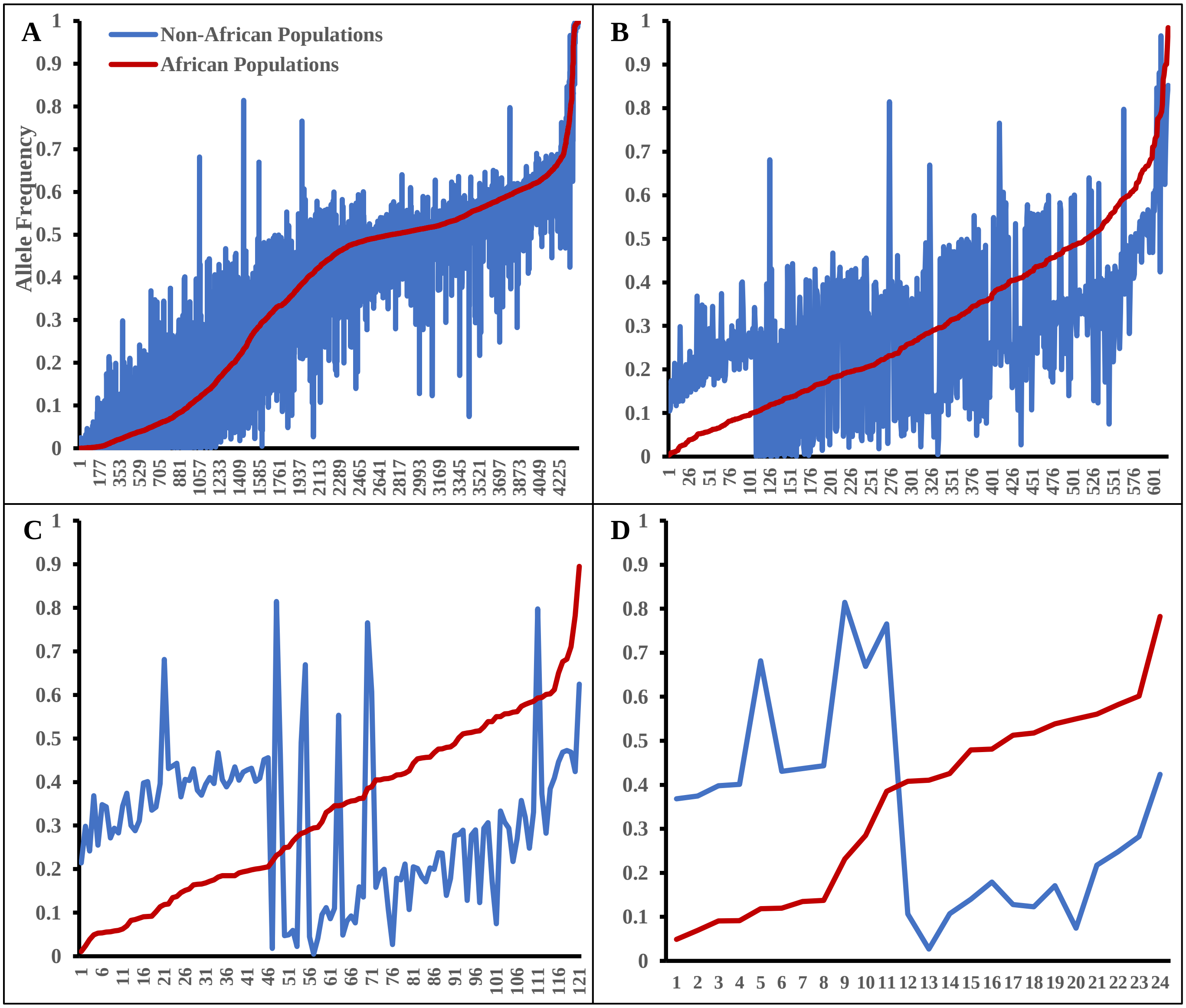

Figure 21. Allele frequency distributions comparing non-African populations to African populations. A) Overview of all polymorphic ME allele frequencies between non-African populations and African populations. B-D) Distribution for polymorphic loci with insertion allele frequency differences of minimally 10%, 20%, 30%, respectively.

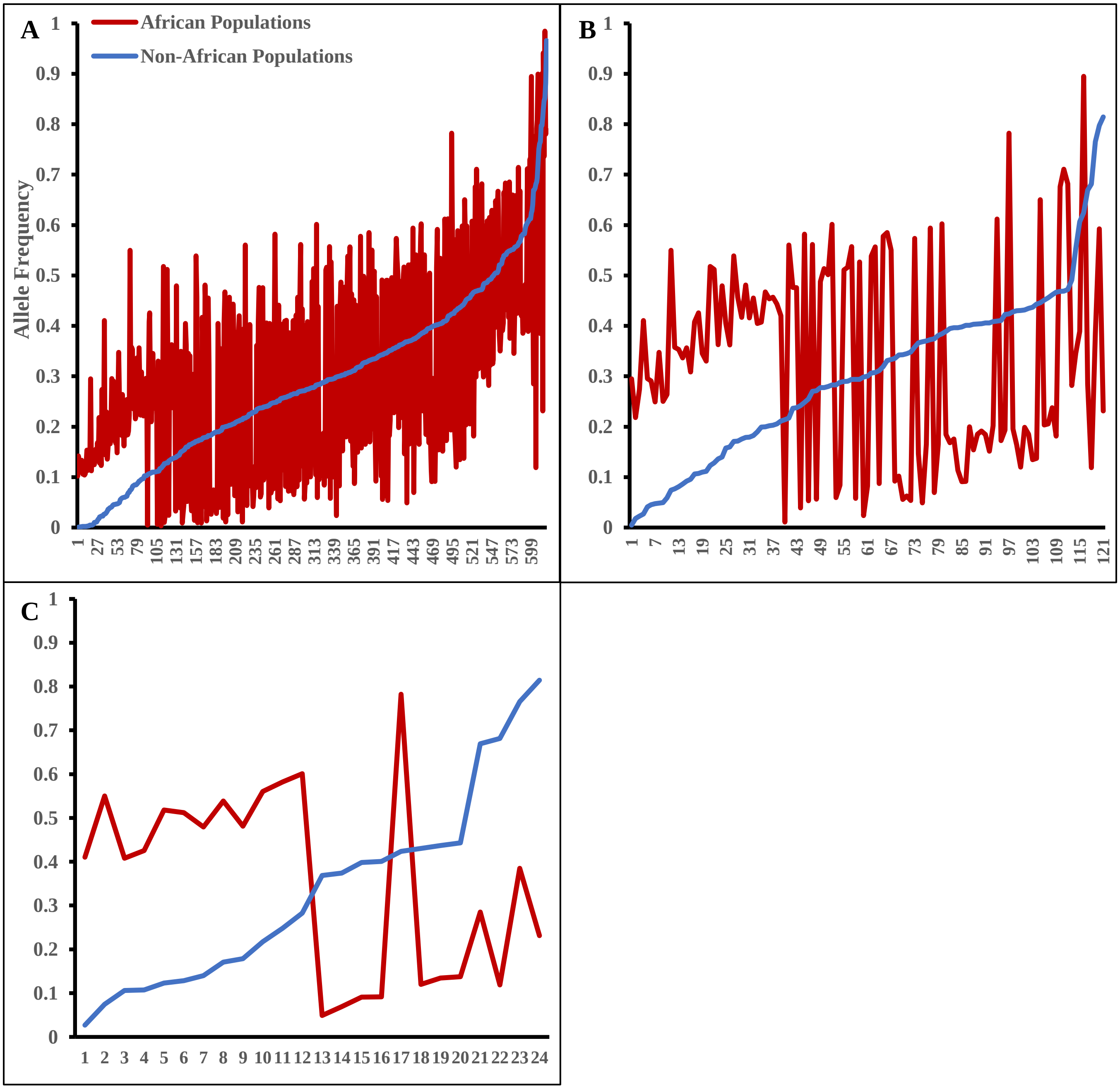

Figure 22. Allele frequency distributions comparing African populations to non-African populations. A-C) Distribution for polymorphic loci with insertion allele frequency differences of minimally 10%, 20%, 30%, respectively.

3.4.3 Continental-specific polymorphic MEs

The patterns of polymorphic ME allelic diversity present among human populations suggest the possibility that there may be ME loci specific to exclusive groups of individuals. To identify such entries, specific criteria on the allele frequency distributions were set among the five continental groups. The first condition involved finding ME entries that were present in only 1 out of 5 continental groups while the insertion allele was completely absent in the other 4 continental groups. The second condition was applied to locate entries that was found at an allele frequency greater than or equal to 10% in one continental group only while the other groups had an allele frequency less than or equal to 1% for the same entry. From this analysis, a total of 16 unique polymorphic ME entries were identified (Table 9). More specifically, from the list of 16 loci, all were found to be specific to African populations, indicating a higher level of genetic diversity in African than in non-African as seen in the form of ME insertion polymorphism.

Table 9: Continental-specific polymorphic MEs based on specific criteria to identify unique loci

	Continental Group (CG)					
Criteria	**AFR**	**AMR**	**EAS**	**EUR**	**SAS**	**Total**
1 CG >0, 4 CG =0	2	0	0	0	0	2
1 CG >=10%, 4 CG <=1%	14	0	0	0	0	14
Sub-total	16	0	0	0	0	16
Unique	16	0	0	0	0	16

3.5 Polymorphic MEs contribute to the coding and regulatory regions in the human genome

3.5.1 Overview of the gene context and functional assessment of polymorphic MEs

To examine the potential functional impact of polymorphic MEs to the human genome, UCSC liftOver tools were used to first convert the hg19 chromosomal positions of polymorphic MEs to the hg38 version, in order to compare with the various gene annotation files including GENCODE and Ensembl. The liftOver conversion was able to successfully modify the positions of 4389 polymorphic MEs, while 11 ME entries failed to convert due to issues reported in the hg38 build of the human genome (Appendix Table S5). The total 4389 polymorphic MEs is represented by 3642 Alus, 567 L1s, 121 SVAs, 31 LTRs, and 28 PPSGs. While most of the polymorphic MEs have inserted into intergenic regions, there are many MEs (13%) inserted into various coding and regulatory regions in the genome. Among the different types of MEs, Alus contribute the most in number, followed by L1s, SVAs, PPSGs, and LTRs (Table 10). Interestingly, by the ratio among all polymorphic MEs in the class, PPSGs are found to have the highest percentage (75%) inserted into these genomic regions, followed by SVAs (56%), L1s (38%), LTRs (16%), and Alus (7%).

3.5.2 Polymorphic MEs' contribute to protein coding regions in the human genome

In examining the different exonic regions which are divided into protein-coding (CDS, 5' UTR, 3' UTR) and non-coding (NR) transcripts, polymorphic MEs were found to contribute to all parts. A total of 66 polymorphic MEs are found to contribute to NR transcripts, while a total of 51 entries contribute to protein-coding transcripts (Table 10). Among those inserted into protein coding exons, most polymorphic MEs participated in the 3' UTR (31), followed by CDS (14) and 5' UTR (6). There are more polymorphic MEs in the 3' UTR than in 5' UTR, and this might be attributed to the fact that 3’ UTR regions are much longer than 5’ UTRs in general (Table 10).

The collected data shows that polymorphic MEs have the potential to impact human genes by directly participating in protein coding regions, which can lead to gene function differences among populations and even individuals. In total, polymorphic MEs have contributed to the CDS regions for a total of 14 genes. In the analysis of the genes involving polymorphic MEs, it was observed that the main molecular functions of these genes included protein-binding, ATP-binding, and ion-binding (Appendix Table S6). The specific mechanisms and related biological significance of these polymorphic MEs in proteins remain to be examined in future studies. In the examples shown in Figure 23, a L1 element contributes to the CDS portion of RAS And EF-Hand Domain Containing (*RASEF*) gene, which is a protein coding gene associated with calcium ion binding and GTPase activity (Figure 23A) (Shintani et al., 2007). Two SVA elements also contribute to protein coding genes, Heparan-Alpha-Glucosaminide N-Acetyltransferase (*HGSNAT*) and Zinc Finger Protein 83 (*ZNF83*) (Figure 23B,C). The functional annotation related to these genes include transferase activity, nucleic acid binding and DNA-binding transcription factor activity, respectively (Fedele et al., 2007; Laity, Lee, & Wright, 2001). The distribution pattern of these 3 examples involve intermediate to high insertion allele frequencies (45-65%) that are relatively similar among all 28 populations.

Aside from directly impacting protein-coding, polymorphic MEs contribute to 3’ UTRs of several genes. As seen in Figure 24, an entire Alu element contributes to the 3’ UTR of the Nucleic Acid Binding Protein 1 (*NABP1*) gene, which encodes a protein for nucleic acid binding and single-stranded DNA binding (Figure 24A). Another entire Alu element contributes to the 3’ UTR of the Solute Carrier Family 35 Member E3 (*SLC35E3*) gene (Figure 24B). In Figure 24C, two individual Alu elements contribute to the 3’ UTRs of two different protein transcripts, including *RP11-745010.4* and *LLPH-AS1* genes. The presence of these polymorphic MEs may impact the post-transcriptional regulation of these genes through mechanisms including miRNA interference and RNA-editing (Athanasiadis, Rich, & Maas, 2004; Chen, DeCerbo, & Carmichael, 2008).

Table 10. Summary of polymorphic MEs’ contribution to coding and regulatory regions in the human genome based on ME type

Genomic Regions	Alu	L1	SVA	LTR	PPSG	Total
Total ME Count	3642	567	121	31	28	4389
CDS	5	3	5	0	1	14
5’ UTR	1	2	2	0	1	6
3’ UTR	21	2	7	0	1	31
Intron-Exon Splice Sites	17	10	10	0	1	38
NR	31	9	5	0	21	66
Promoter	15	1	0	0	12	28
Promoter Flanking Region	38	17	0	1	0	56
Enhancer	9	1	1	0	0	11
TFBS	9	2	1	0	0	12
CTCF	58	203	57	3	1	322
Open Chromatin	67	10	5	1	0	83
CpG Islands	1	0	2	0	4	7
miRNA Target Sites	1	0	0	0	1	2
Non-redundant Total	252	216	68	5	21	562
Percentage	7%	38%	56%	16%	75%	13%

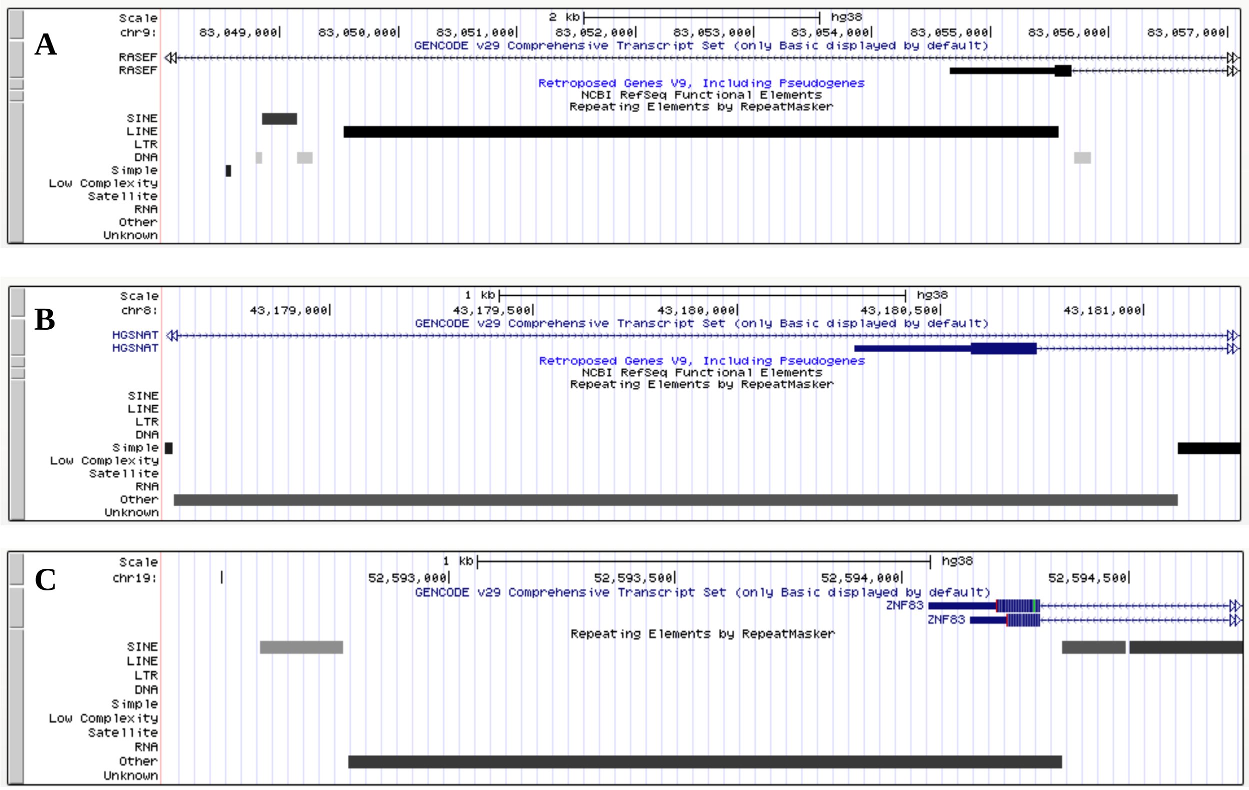
A
Scale
chr9:
2 kb
hg38
83,049,000
83,050,000
83,051,000
83,052,000
83,053,000
83,054,000
83,055,000
83,056,000
83,057,000
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
RASEF
RASEF
Retroposed Genes V9, Including Pseudogenes
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
B
Scale
chr8:
1 kb
hg38
43,179,000
43,179,500
43,180,000
43,180,500
43,181,000
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
HGSNAT
HGSNAT
Retroposed Genes V9, Including Pseudogenes
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
C
Scale
chr19:
1 kb
hg38
52,593,000
52,593,500
52,594,000
52,594,500
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
ZNF83
ZNF83
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 23. Examples of polymorphic MEs contributing to CDS of genes. A screenshot of polymorphic MEs in CDS regions were shown in the UCSC Genome Browsers with the gene structure (GENCODE) and RepeatMasker tracks displayed. A) A L1 element contributes to the CDS exon of the *RASEF* gene. B) An SVA element contributes to the CDS exon of the HGSNAT gene. C) An SVA element contributes to the CDS exon of the *ZFP83* gene.

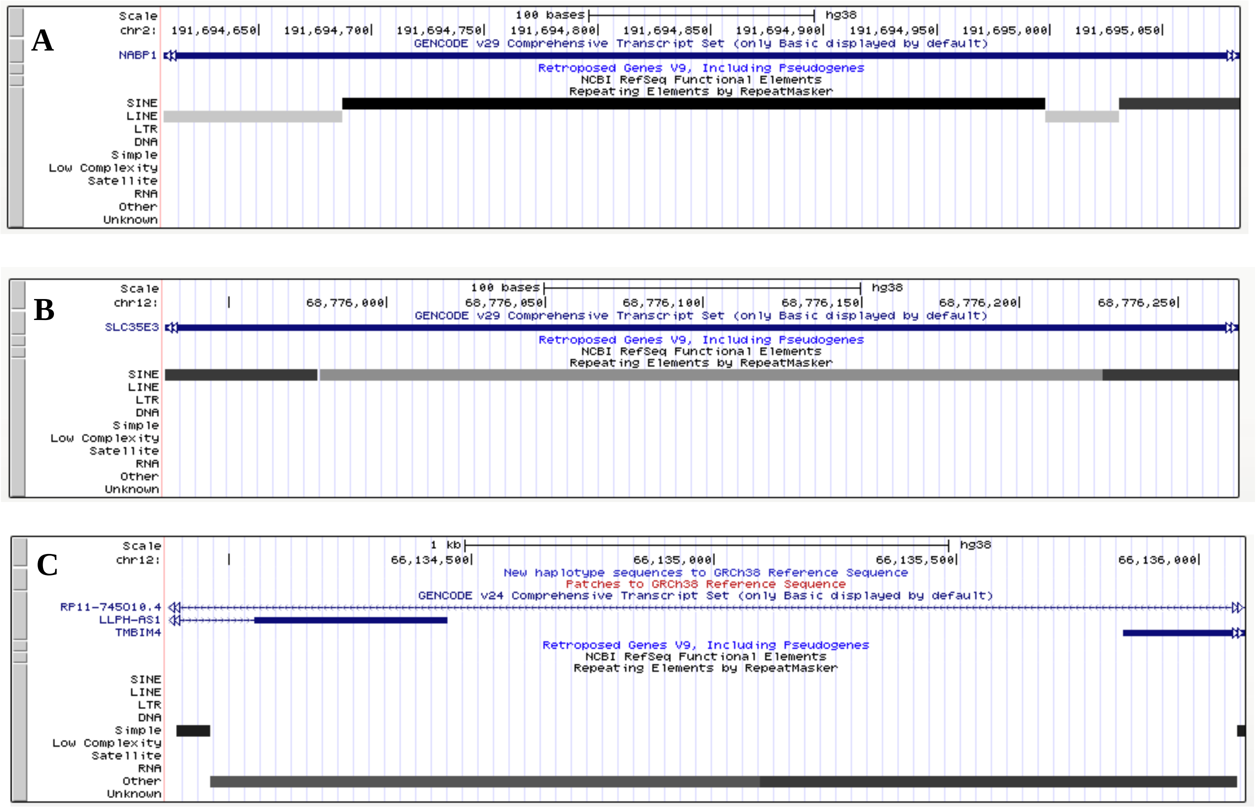
A
Scale
chr2:
100 bases
hg38
191,694,650
191,694,700
191,694,750
191,694,800
191,694,850
191,694,900
191,694,950
191,695,000
191,695,050
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
NABP1
Retroposed Genes V9, Including Pseudogenes
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
B
Scale
chr12:
100 bases
hg38
68,776,000
68,776,050
68,776,100
68,776,150
68,776,200
68,776,250
GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)
SLC35E3
Retroposed Genes V9, Including Pseudogenes
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
C
Scale
chr12:
1 kb
hg38
66,134,500
66,135,000
66,135,500
66,136,000
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
RP11-745O10.4
LLPH-AS1
TMBIM4
Retroposed Genes V9, Including Pseudogenes
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 24. Examples of polymorphic MEs contributing to 3' UTR regions of genes. A screenshot of polymorphic MEs in 3' UTR regions were shown in the UCSC Genome Browsers with the gene structure (GENCODE) and RepeatMasker tracks displayed. A) An entire Alu element contributes to the 3' UTR of the *NABP*1 gene. B) An entire Alu element contributes to the 3' UTR of the *SLC35E3* gene. C) Alu elements contribute to 3' UTRs of two different protein transcripts, *RP11-745O10.4* and *LLPH-AS1*.

3.5.3 Polymorphic MEs' participate in regulating gene expression

In the analysis of regulatory features of the human genome, regions including promoters, promoter flanking regions, enhancers, transcription-factor binding sites (TFBS), CTCF binding sites, and open chromatin regions were examined. A total of 95 polymorphic MEs were identified in promoter and enhancer regions, which suggests that these MEs have a potential impact on gene expression in the human genome (Table 10). A large number of polymorphic MEs were also found to contain the binding sites of transcriptional factors, particularly 322 entries for CTCF binding sites and 12 entries for TFBS (Table 10). This proposes the impact polymorphic MEs can have in altering gene expression aside from interrupting existing binding sites by insertion, but also through the possible creation of new sites for transcriptional factors.

Apart from direct interactions with transcriptional binding sites, polymorphic MEs can influence gene expression by inserting into open chromatin regions and CpG islands. A total of 83 polymorphic MEs were found inserted in the open chromatin regions, as well as 7 MEs found within CpG islands (Table 10). Examples of polymorphic MEs in CpG islands of various lengths can be seen in Figure 25. Polymorphic ME elements were also reported to have the ability to impact the human genome through intron splicing mediated by insertions within splice sites. A total of 38 entries were inserted within intron-exon splice sites which implies the impact MEs may have on the genome splicing process (Table 10). The example event of polymorphic ME insertions within intron-exon splice sites can be seen in Figure 26.

Since gene expression can also be regulated post-transcriptionally, the contribution of polymorphic MEs to miRNA target sites were examined. As seen in Table 10, a total of 2 polymorphic MEs were found to be inserted within miRNA target sites, indicating a small proportion of MEs contributing to these regions. MEs have also been seen to participate in the generation of miRNA sequences, however no cases of polymorphic MEs were identified in this study.

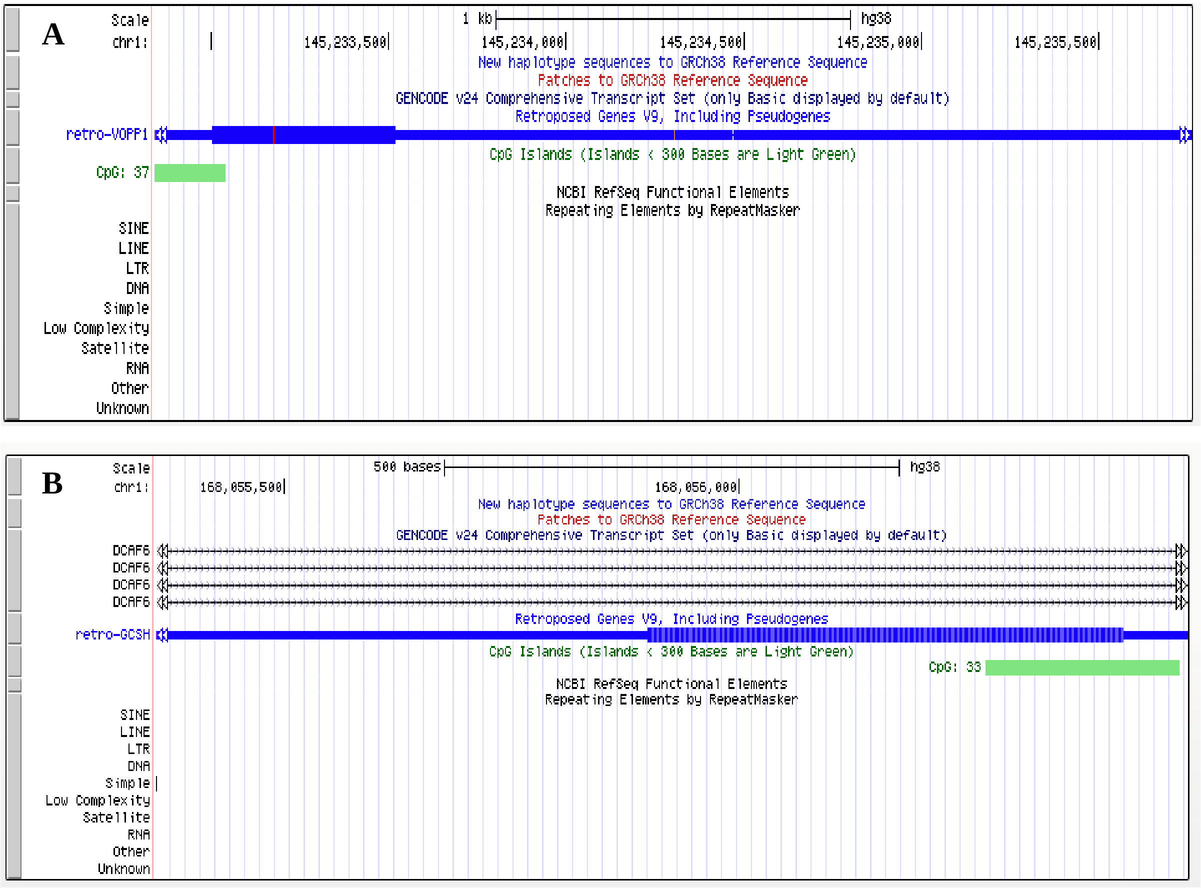
A
Scale
chr1:
1 kb
hg38
145,233,500
145,234,000
145,234,500
145,235,000
145,235,500
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
Retroposed Genes V9, Including Pseudogenes
retro-VOPP1
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 37
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
B
Scale
chr1:
500 bases
hg38
168,055,500
168,056,000
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
DCAF6
DCAF6
DCAF6
DCAF6
Retroposed Genes V9, Including Pseudogenes
retro-GCSH
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 33
NCBI RefSeq Functional Elements
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 25. Examples of polymorphic MEs contributing to CpG islands. A screenshot of polymorphic MEs in CpG islands were shown in the UCSC Genome Browsers with the gene structure (GENCODE), RepeatMasker, and CpG island tracks displayed. A-B) Two individual PPSG elements contribute to CpG islands of various lengths.

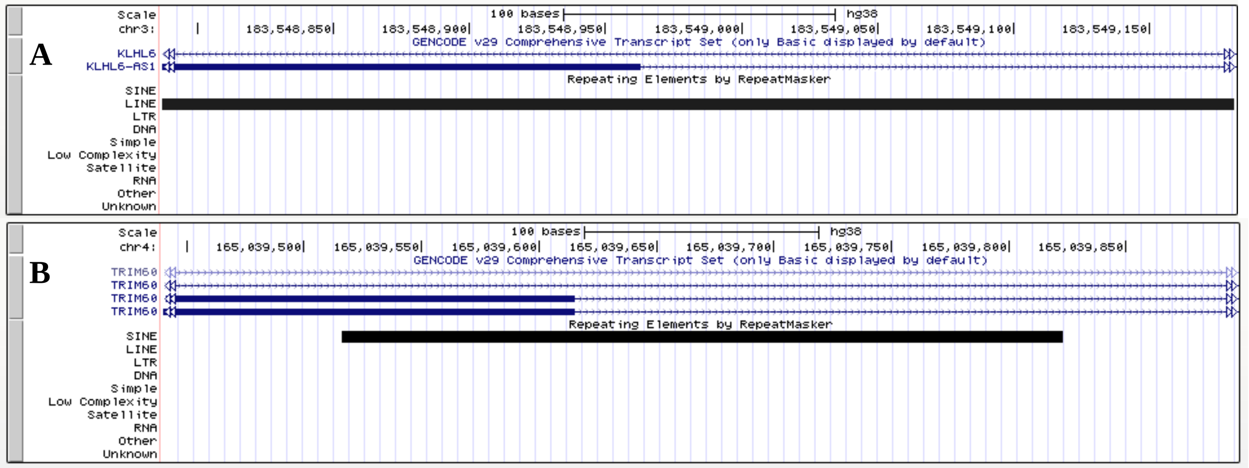

Figure 26. Examples of polymorphic MEs contributing to intron-exon splice sites. A screenshot of polymorphic MEs in CpG islands were shown in the UCSC Genome Browsers with the gene structure (GENCODE) and RepeatMasker tracks displayed. A) A reference polymorphic L1 element contributes to the intron-exon splice site of *KLHL6-AS1*. B) A reference polymorphic Alu element contributes to the same exon of two different protein transcripts of *TRIM5*.

3.5.4 Functional assessment of polymorphic MEs based on allele frequency groups

For further analysis, polymorphic MEs were categorized into 3 main allele frequency groups including low (<10%), intermediate (>=10-49%), and high (>=50%), based on their overall distribution across 28 populations. This assortment included 935 polymorphic MEs in the low frequency group, 1689 entries as intermediate frequency, and 1765 entries in the high allele frequency groups (Table 11). Among all functional genomic regions, MEs with higher frequencies were more abundant in number. Overall, the percentage of polymorphic MEs in each allele frequency group, from low to high, that contributed to specific genic regions was 8%, 13%, and 15%, respectively (Table 11). However, when comparing the ratio between the total number of MEs in each allele frequency group to the number of MEs seen across each genic region, CDS and 3’ UTR regions had a larger percentage of low frequency MEs (Table 12). This implies that the insertion of MEs into genic regions are subject to selection according to the level of functional constrain or importance.

Table 11. Summary of polymorphic MEs’ contribution to coding and regulatory regions in the human genome based on allele frequency groups

ME Class	Low	Med	High	Total
Total ME Count	935	1689	1765	4389
CDS	6	4	4	14
5’ UTR	0	3	3	6
3’ UTR	8	14	9	31
Intron-Exon Splice Sites	0	16	22	38
NR	10	23	33	66
Promoter	4	11	13	28
Promoter Flanking Region	10	22	24	56
Enhancer	0	6	5	11
TFBS	1	5	6	12
CTCF	19	127	176	322
Open Chromatin	19	37	27	83
CpG Islands	1	0	6	7
miRNA Target Sites	0	1	1	2
Non-redundant Total	76	221	265	562
Percentage	8%	13%	15%	13%

Table 12. Percentage of polymorphic MEs' contribution to coding and regulatory regions in the human genome based on allele frequency groups

ME Class	Low	Med	High	Total
Total ME Count	935	1689	1765	4389
CDS	0.64%	0.24%	0.23%	0.32%
5' UTR	0.00%	0.18%	0.17%	0.14%
3' UTR	0.86%	0.83%	0.51%	0.71%
Intron-Exon Splice Sites	0.00%	0.95%	1.25%	0.87%
NR	1.07%	1.36%	1.87%	1.50%
Promoter	0.43%	0.65%	0.74%	0.64%
Promoter Flanking Region	1.07%	1.30%	1.36%	1.28%
Enhancer	0.00%	0.36%	0.28%	0.25%
TFBS	0.11%	0.30%	0.34%	0.27%
CTCF	2.03%	7.52%	9.97%	7.34%
Open Chromatin	2.03%	2.19%	1.53%	1.89%
CpG Islands	0.11%	0.00%	0.34%	0.16%
miRNA Target Sites	0.00%	0.06%	0.06%	0.05%

3.5.5 Distribution of polymorphic MEs among coding and regulatory regions

To examine the distribution of polymorphic MEs in coding and regulatory regions in the human genome, the total non-gapped lengths of each region was measured. The human genome was divided into sections including CDS exon, non-CDS exon (5' UTR, 3' UTR, NR), intron-exon splice sites, non-splicing introns, promoter, promoter flanking region, enhancer, TFBS, CTCF binding sites, open chromatin, and intergenic regions (Table 13). Overall, the vast majority of polymorphic MEs were found in intergenic regions, which is correlated to the total length of this region being the greatest among all genomic regions analyzed. The lowest density of polymorphic MEs was found in the CDS and enhancer regions with 0.40 copies/1Mbp and 0.47 copies/1Mbp, respectively. In contrast, the highest density was present among intron-exon splice sites with 9.88 copies/1Mbp (Table 13), indicating a high potential for impacting splicing.

Interestingly, the density of polymorphic MEs is much higher (4.79 copies/1Mbp) in the CTCF sites than others and the genome average with the exact functional implication unknown. The distribution pattern seems to be quite similar among the remaining genomic regions as being ~1 copy/1Mbp.

Table 13. Density of polymorphic MEs in regulatory and coding regions across the human genome

Categories	Total Length (bp)	Observed Number of MEs	Density/1Mbp
CDS exon	35011234	14	0.40
Non-CDS exon	97704305	103	1.05
Intron Splice	3844470	38	9.88
Intron Non-Splice	458606380	679	1.48
Intergenic	2291477617	3827	1.33
Promoter	27621262	28	1.01
Enhancer	23429470	11	0.47
TFBS	12102373	12	0.99
CTCF	67185000	322	4.79
Open Chromatin	69551414	83	1.19
Promoter Flanking Region	65307208	56	0.86
Genome Average	3151840733	4400	1.40

Chapter 4: Discussions

The main objective of this study was to provide a comprehensive compilation of identified polymorphic MEs with full sequence and genotype characterization among human populations. Through the compilation of multiple data sets, we assembled a non-redundant list of 4400 polymorphic MEs with complete sequences for both the insertion and pre-integration alleles characterized. This list represents the best characterized polymorphic MEs to date, which are to be used to modify and update dbRIP. The data can be used as a reference data set for many research activities related to the analysis of human ME polymorphism, for example, as training and test data sets for evaluating bioinformatics tools aiming to identify ME-derived structural variants (SVs) in personal genome analysis (Puurand et al., 2019). The availability of complete sequences for both alleles permits development of new standalone tools or new modules within existing pipelines to provide accurate identification of SVs associated with known ME variants in personal genome analysis. Furthermore, the availability of the insertional allele sequences for the non-reference MEs permits the detection of sequence polymorphism in these sequences, which is not possible without.

We then used this high-quality data set to analyze the pattern of polymorphic MEs in the human genome and assess the potential impact on genome evolution and function.

4.1 How much ME polymorphisms is there for humans?

Among the 4400 polymorphic MEs, the vast majority belong to the family of Alus, followed by L1s, SVAs, LTRs, and PPSGs (Table 5). A total of 28 polymorphic PPSGS were reported, representing the first for this ME type (Figure 4). The Alu family being the most common ME in the human genome correlates with the high degree of polymorphism exemplified by this type of ME (Stewart et al., 2011). ME insertions distinguished as insertions to the reference are labeled as non-reference MEs, while ME insertions detected as deletions to the reference are known as reference MEs. In the list of all identified polymorphic MEs, a total of 2623 entries are categorized as reference MEs while a total of 1777 are non-reference MEs. The greater levels of reference polymorphic MEs apparently attribute to most of the polymorphic MEs being at high frequencies and the limited number of analyzed personal genomes (Figure 5; Figure 10). The current resources in terms of data sets and bioinformatics tools make it easier to identify reference polymorphic MEs than the non-reference MEs due to the involvement of detecting new sequences in the latter case rather than detecting loss of existing sequences in the former case. However, the number of reference polymorphic MEs that can be identified is a finite number, which is the number of recent MEs documented in the reference genome, which should mostly be human-specific, a number determined to be around 15,000 (Tang et al., 2018). In contrast, while the current number of non-reference polymorphic MEs is smaller than the reference entries, the discovery of new non-reference candidates is only limited by the number of personal genome sequences available. Therefore, as sequencing technologies improve (e.g. the personal genome sequences from the long-read platforms) (Buermans & den Dunnen, 2014; Lee et al., 2016), with increasing coverage of genome sequencing for more diverse human populations and the development of better tools, the number of non-reference MEs is expected to gradually increase (Ewing, 2015; Keane, Wong, & Adams, 2013). As a matter of fact, a large number of non-reference polymorphic ME loci without full insertion sequences have been documented in the literatures (Ewing et al., 2013; Gardner et al., 2017; Puurand et al., 2019; Wang et al., 2006; Witherspoon et al., 2010).

Among the five ME types, the largest difference in number between reference and non-reference entries was present in the family of LTRs (Figure 5). This much lower ratio of non-reference LTRs is likely a result of two factors: the low activity of LTRs and their long sequence, which makes it even more difficult to obtain the full insertion sequences than for the shorter MEs. For the latter reason, there are more tools that focus on identifying non-LTR MEs, but very few for identifying polymorphic LTRs (Chen, 2019; Kang et al., 2016; Valencia & Girgis, 2019).

Based on the total number of human specific MEs (HS-MEs), we predict that with the ongoing personal genome analysis into the future years at an accelerated pace, the number of reference polymorphic MEs will increase slightly to a certain level around 10,000, while the number of non-reference polymorphic MEs will increase quickly to the order of tens of thousands. These new non-reference polymorphic MEs are expected to have low frequency and mostly belong to Alus, L1s, and SVAs. Based on the fact that African populations have a greater level of genetic diversity as shown in our data and the latest consensus that African population diversity is much richer than what we currently appreciate (Rishishwar et al., 2015), we can expect the most of the new ME polymorphisms will be discovered from personal genome analysis of more diverse African populations.

4.2 Polymorphic MEs' contribution to human genetic diversity

A hallmark of the integration process of ME insertions is the generation of TSD sequences (Kim et al., 2006) . It also represents a type of genomic rearrangement generated by polymorphic ME insertions in addition to the ME insertion itself, and its sequence pattern reflects the retrotransposition mechanism involved. Following a TSD length survey for all polymorphic MEs, it was found that Alus, L1s, SVAs, and PPSGs, otherwise known as non-LTR retrotransposons, have a TSD length of on average 15 bp (Figure 6). This confirms the existing model that the mobilization of non-LTR retrotransposons is mediated by the same mechanism facilitated by L1elements (Han, 2010; Tang et al., 2018). This also supports the classification of

PPSGs as non-LTR retrotransposons (Dewannieux & Heidmann, 2005; Pavlicek et al., 2006). In contrast, LTRs have significantly shorter TSD lengths than the non-LTR families which agrees with the fact that LTRs use a different retrotransposition mechanism (Figure 6) (Schorn et al., 2017). Altogether, the length and sequence characteristics endorses the mechanism of retrotransposition carried for each type of MEs.

The transduction events of MEs have also contributed to genomic rearrangements. Among the two types of transductions, the transduction of 3’ flanking sequences is more common and longer in sequence length. This trend seems to be similar among non-LTR retrotransposition due to the potential skipping of the poly-A signal and instead using a second downstream poly-A site mediated by the 3’-processing machinery (Goodier et al., 2000). From all polymorphic ME types, L1s are the most successful family in terms of size contribution through transductions. In total, 25 cases of 5’ and 3’ transduction events by L1s have brought an increase of 992 bp of sequence in addition to the ME insertion sequence itself. One case of LTR insertion generated a transduction of sequences on both 5’ and 3’ ends of the ME insertion, resulting in a total of 59 bp of additional sequence to the human genome (Table 7). Consequently, ME insertion-mediated transduction of flanking DNA is likely to be an important mechanism of genome evolution through increasing genome plasticity and facilitating new combinations of sequences (Pickeral et al., 2000). Transductions are also known to be significantly less disruptive to genome organization than larger-scale genome rearrangement such as inversions or translocations, and therefore may be used as a potential mechanism for novel sequence combinations in the human genome (Pickeral et al., 2000).

The integration of polymorphic MEs at new genomic sites can also disrupt genome through the deletion of adjacent genomic sequence, known as insertion-mediated deletions

(IMDs). No cases of these events were reported for polymorphic MEs likely due to its low rate of occurrences; however, it has been shown to occur in the human genome. Previous studies have reported multiple distinct Alu and L1 insertions that mediate distinct disease-associated deletions (Cordaux & Batzer, 2009; Hancks & Kazazian, 2016). Alterations such as these have also been linked to the loss of gene function (Jahic et al., 2016). Overall, the modifications to genome sequences associated with polymorphic MEs signifies the potential changes of this type genomic variants in the human genome, which demands for our attention in future personal genome analysis.

4.3 The ongoing activity levels of MEs in the human genome

The profiles of active MEs have been extensively studied in the human genome. Based on these analyses, studies show that Alus, L1s, SVAs, and HERV-K from the family of LTRs, are the specific ME types still undergoing retrotransposition in the human genome (Mills et al., 2007; Shin et al., 2013; Sudmant et al., 2015; Wang et al., 2006). Identifying the most active subfamilies for each polymorphic ME type will reflect the retrotransposition profiles and reveal the patterns among current human populations.

A higher polymorphic ME ratio means there are more entries relatively contributed by recent retrotransposition events and therefore, this ratio was used as an indicator for the recent or ongoing activity level of different ME families. As expected, Ya5 and Yb8/9 for the family of Alus, L1HS for L1s, F and E subfamilies of SVAs, and HERV-K for LTR retrotransposons were the most active ME subfamilies, which confirms the data identified by prior studies (Appendix Table S2) (Tang et al., 2018). Compared to all other MEs, the family of polymorphic LTRs was the least active, which resembles the low number of polymorphic LTRs identified in current human genomes.

4.4 Different preferences for sex chromosomes by different polymorphic ME classes

As shown in Figure 7 and Appendix Table S3, the overall polymorphic ME density on chromosomes X and Y is significantly lower than all autosomal chromosomes. By ME type, polymorphic MEs from Alus, SVAs, and PPSGs all showed a lower density for both sex chromosomes (Figure 8). For these 3 families of polymorphic MEs, the density in chromosome X and Y is at least 2 to 6 times lower than the genome average, while for polymorphic LTRs, the density for both X and Y chromosomes are 2 to 4 times greater than the genome average. In contrary, polymorphic L1s displayed a low density for chromosome Y, but a density greater than the genome average for chromosome X. More specifically, the observed number of polymorphic MEs across chromosomes X and Y, as well as chromosome 1, were significantly lower than the expected number. In contrast, autosomal chromosomes, including chromosomes 4, 5, 6, 7, and 13 had greater number of polymorphic MEs than the expected value (Appendix Table S4). This differential bias for chromosomes X, Y, and 1 among different ME types could be a result of gene density, as its polymorphic ME densities show a strong deviation from the genome average. Chromosome 1 exhibits a gene density that is greater than the genome average, whereas chromosomes 4,5,6,7, and 13 display very low gene densities compared to the genome average (Appendix Table S3). However, this phenomenon cannot be explained in the same manner for chromosome X. The gene density for chromosome X is lower than the genome average, yet the density of all polymorphic MEs, except for L1 and LTRs, remain low for this chromosome. The distribution patterns among polymorphic MEs was found to be partially different for HS-MEs, specifically for the density on chromosome Y. All HS-MEs except for SVAs displayed a high bias towards the Y chromosome, with LTRs exhibiting the largest density. The high preference for chromosome Y by HS-MEs was described to be partially due to the lack of homologous recombination-based deletions and lack of selection pressure in the male germline (Tang et al., 2018). The exact mechanism of ME distribution patterns still remains unknown, however the lower density of Alus, SVAs, and PPSGs in the sex chromosomes can be partially explained by their lower copy numbers in the gene pool relative to the autosomes as a target for insertions.

4.5 Polymorphism of MEs exist in humans with the older MEs present in other primate genomes

Based on previous studies, polymorphic MEs were defined as recent retrotransposition events that show a polymorphic status for its presence and absence in the human population, thus such polymorphisms were interpreted to be human specific (Wang et al., 2017). However, our results demonstrate that polymorphic sequences are not specific to one species, suggesting the existence of other mechanisms for MEs to be polymorphic. In total, 11 polymorphic ME loci was identified to be present in the chimpanzee genome, including one entry that was present in both chimpanzee and orangutan genomes (Table 8). This signifies that although a majority of polymorphic entries are human-specific, it should not be limited to human-specific MEs as candidates for identifying polymorphic MEs, as insertions shared with primates can still be polymorphic. All but one of the 11 loci had insertion allele frequencies greater in all continental groups than the average allele frequencies among other polymorphic ME loci. The genetic basis of polymorphic MEs present among humans and primates suggest the influence of balancing selection, which can be defined as a form of adaptation that leads to the perseverance of variation in a population or species due to the loss by genetic drift (Leffler et al., 2013). There is also evidence suggesting that although very rarely MEs can be precisely deleted from the genome, most likely by the recombination events between TSDs that are flanking the ME insertion (van de Lagemaat et al., 2005). The deleted locus resembles the pre-integration allele, therefore effectively reversing the insertion, which can describe events of polymorphic MEs present in primate genomes such as the chimpanzee, yet not be completely fixed in human populations (van de Lagemaat et al., 2005).

4.6 Distribution of polymorphic MEs in human populations

The analysis of human polymorphic MEs within and between populations revealed that ME insertion polymorphisms are found at intermediate allele frequencies (Figure 10). The patterns depicted from the insertion allele frequency spectrum is consistent with a role for positive selection for ME insertions in human evolution. This suggests that some ME insertions may have increased in frequency owing to the effects of positive selection, instead of purifying selection that is commonly seen with other structural variants (Lowe, Bejerano, & Haussler, 2007). ME polymorphisms with a higher allele frequency in one particular population can also be expected to have evolved earlier during the evolution timeframe of that population. For example, polymorphic MEs displaying high allele frequencies could represent a set of ME insertions that occurred before migration of humans from Africa (Rishishwar et al., 2015).

Closer inspection of the polymorphic ME allele frequency spectrum between reference and non-reference MEs revealed the differential skewed patterns between the two groups. Reference MEs were found to have insertion allele frequencies towards the higher end of the spectrum, while non-reference MEs were primarily skewed towards the lower end of the spectrum (Figure 11). The large number of non-reference polymorphic MEs with allele frequencies less than 5% indicates this set of insertions to be more recently derived among human populations, and it could potentially resemble ME polymorphisms unique to individual human populations (Stewart et al., 2011). Reference polymorphic MEs are expected to be higher in insertion allele frequency as they are more likely to represent common variants that are present in the reference genome. Among ME types, polymorphic Alus, L1s, SVAs, and PPSGs generally had similar allele frequency distributions across all populations, except for LTRs, which was the only ME type to display a greater percentage of entries with allele frequencies greater than 95% (Figure 12). The higher rate of polymorphic LTR insertions could be affiliated with the very low activity in humans (Deininger & Batzer, 2002).

4.7 Human population relationships based on polymorphic MEs

The pattern of human genetic diversity has been studied with a wide variety of genetic polymorphisms in diverse human populations. In this study, ME polymorphisms are used to assess the evolutionary relationships of current human populations. As polymorphic MEs are considered to be identical by decent, which is a better class of genetic markers for ancestral study, their patterns of sharing can reveal aspects of human population history (Rishishwar et al., 2015; Sudmant et al., 2015).

The analysis of ME polymorphism patterns can be used to review the evolutionary relations between current human populations. Our data demonstrated that individuals within same continental groups are well clustered together based on the computed analysis of polymorphic ME genotype data (Figure 15-17). This suggests that patterns of polymorphic ME insertion divergence within and between populations can illustrate the known patterns of human evolution. Similar relationships among human individuals was presented from the phylogenetic analysis of the ME genotype data (Figure 18). The transition from the ancestor sample to Bushmen and African populations illustrate the close genetic relationship between these populations. Based on the cluster formation of individuals within the evolutionary tree, South Asian populations seem to be the closest to the African populations, followed by European, Admixed American, and East Asian. Overall, these findings indicate that ME polymorphisms are a type of genetic variant that can be used to reconstruct evolutionary histories or assist in distinguishing specific lineages (Rishishwar et al., 2015).

The diversity of polymorphic MEs can also be affiliated with geographic ancestry of human populations. While examining the distribution pattern of individual polymorphic MEs, high levels of geographic differentiation was observed, solely by comparing insertion allele frequencies between individual populations. For most cases, polymorphic MEs were not present exclusively to an individual population, but instead more commonly present or absent in specific groups of populations. Patterns in allele frequencies were found to be directly relating to all individual populations within specific continental groups (Figure 19-20). Polymorphic MEs also showed substantial geographic differentiation when compared between African and non-African populations. African populations have a higher number of polymorphic MEs and therefore show the highest level of ME polymorphism (Figure 21-22). Further analysis discovered several polymorphic MEs that were completely absent in all populations, except for African populations, thus representing African-specific MEs (Table 9). This geographic differentiation suggests the possibility of a selection of polymorphic MEs to be used as markers for future evolutionary studies.

4.8 Impact of polymorphic MEs on gene function

The overall functional impact of polymorphic MEs on the human genome was also assessed by examining their location in relation to known genes and functional regulatory elements. In examining different exonic regions, polymorphic MEs have a selection preference to non-coding regions, where 66 polymorphic MEs were observed contributing to NR transcripts, versus the 51 polymorphic MEs found in protein-coding regions (Table 10). The breakdown of polymorphic MEs in the protein-coding region consisted of 14 polymorphic MEs in the CDS region, 6 in the 5’ UTR, and 31 in the 3’ UTR regions. Changes to protein coding sequences from the insertional event of polymorphic MEs can result in novel mutations that may modify or disrupt the functions of the genes. Coding regions have been identified to undergo modifications with the insertion of specific MEs, including the introduction of alternative start and stop codons, which ultimately lead to alternations in gene function (Nekrutenko & Li, 2001). Apart from the direct impact of polymorphic MEs in CDS regions, MEs can also affect gene function by influencing stability of mRNA when interacting with 3’ UTR regions of coding genes (Elbarbary et al., 2016). Such insertional polymorphisms in exon regions may have contributed to the genetic versatility and phenotypic differences present among individuals and populations. Non-coding regions have also been recognized as important regulators of gene expression associated with important biological functions (Ulitsky & Bartel, 2013). Studies have identified the role of Alu elements within long non-coding RNAs, which involved the regulation of multiple processes including gene transcription, mRNA decay, alternative splicing, and translation (Kim et al., 2016). With the event of polymorphic MEs identified in both coding and non-coding regions, it is reasonable to believe that ME polymorphisms contribute to the polymorphism in gene regulation as further detailed in the next section.

4.9 Impact of polymorphic MEs in regulating gene expression

Polymorphic MEs are shown to substantially contribute to the sequences of non-coding transcripts and with the high number of polymorphic MEs present in these regions (66 entries) (Table 10). Further, from the functional assessment, polymorphic MEs have shown to provide to a wide variety of gene regulatory sequences including promoters, enhancers, and transcription factors (Table 10). ME polymorphisms were also shown to potentially influence aspects of chromatin structure throughout the genome by interacting with open chromatin regions and

CTCF binding sites (Table 10). Previous studies have observed retrotransposons that carry CTCF binding sites, which has facilitated the prevention of DNA methylation thereby repressing chromatin modifications (Rand et al., 2004; Schmidt et al., 2012). Consequently, interactions between MEs and CTCF binding proteins might be benefitting the insertions by protecting them against repressive chromatin and DNA modifications (Schmidt et al., 2012). Moreover, ME polymorphisms located directly upstream of protein-coding genes may function as promoters or enhancers by functioning as potential binding sites for many transcription factors. ME-mediated transcription factor binding sites can act to modulate gene transcription or simply act as decoys that influence transcription factors away from their active binding sites (Elbarbary et al., 2016). Both Alus and L1s have been identified to introduce new transcription start sites and thereby enhance the increased expression of nearby genes. More specifically, 6 to 30% of human transcripts are believed to use repetitive sequence associated transcription start sites (Elbarbary et al., 2016). Alternative splicing is also a known mechanism that may be initiated by ME polymorphisms. Alternative mRNA splice forms can encode proteins with different function and can be differentially regulated, ultimately contributing to transcript diversity in humans (Cowley & Oakey, 2013).

DNA methylation can also serve as a regulatory switch for transcriptional initiation of genes. CpG islands are highly associated with the promoter region of genes. Polymorphic ME insertions within these CpG islands can potentially represent as alternative promoters and play a role in regulating gene expression (Grandi et al., 2015). In total, 7 polymorphic MEs were identified within CpG islands (Table 10). Furthermore, polymorphic MEs have been documented to contribute non-coding regulatory sequences that modulate gene expression post-transcriptionally. MiRNAs are classified as part of the non-coding region and previous studies have shown that MEs have played a role in the generation of miRNA (Pedersen & Zisoulis, 2016; Roberts, 2014). Polymorphic MEs were not observed to play a role in generating miRNA, however our results revealed that only 2 polymorphic loci interact with miRNA target sites and thereby potentially influence miRNA regulated gene expression (Table 10).

Apart from genes and functional regulatory regions, the majority of polymorphic MEs are located within intergenic regions. While their functional impact is hard to predict, they may also participate in regulatory pathways or directly alter the expression of neighbouring genes. For instance, intergenic non-coding DNA polymorphisms was shown to modulate regulatory domains, which resulted in differential gene expression of Basonuclin 2 (*BNC2*), leading to variation in common human traits, including skin colour (Visser, Palstra, & Kayser, 2014).

The distribution of polymorphic MEs based on allele frequencies demonstrated that there is a great number of high frequency loci in functional and regulatory regions (Table 11). However, when comparing the percentages of polymorphic MEs within each allele frequency group, a larger percentage of low frequency MEs can be seen within CDS and untranslated regions (Table 12). Overall, the density of polymorphic MEs were observed to be higher in non-coding regions comparing to coding regions, likely as a result of differential selection pressure (Table 13).

Together, these collected data strongly state that polymorphic MEs have indeed provided sequences to functional regulatory regions in the human genome, which can fuel regulatory innovation during human evolution. Overall, this implies ME polymorphisms can have various levels of impact in the human genome and gene function, and along with other genetic diversity, they contribute to the phenotypic diversity seen in current human populations both at the population and individual levels.

4.10 Summary and future directions

In this study, through extensive computational and limited experimental analysis, we identified a total of 4400 polymorphic MEs with complete sequence and extensive genotype characterization. This compilation of data includes full sequence characterization of both the insertion and pre-integration alleles that have undergone various validation processes. Additional detailed information regarding sequence contributions to the human genome, such as TSDs, transductions, IMDs, are included for each ME and is also fully characterized. This high-quality data set permits a major update of the dbRIP database, which is an important resource for research on ME polymorphisms in humans.

As polymorphic variants have long been used as markers in human population genetic studies, they are largely known to be suitable for discriminating between populations, and our analysis based on these polymorphic ME markers has demonstrated their value by providing new insights into the relationship of human populations. South Asian populations seem to be the closest to the African populations while East Asian populations are the furthest. Altogether, African populations are shown to be more genetically diverse than the other populations analyzed. The availability of personal genome sequence data for more than 2500 individuals across 28 human populations permitted *in silico* genotyping for these polymorphic MEs, which in turn allowed us to assess their insertion allele frequencies in human populations. All types of MEs displayed a similar distribution pattern across the allele frequency spectrum, with the majority around 50%. The application of several criteria also led to the discovery of continental-specific polymorphic MEs. These findings indicate that ME polymorphism is a type of genetic variant that can be used to assess and review the evolutionary relations between current human populations. These polymorphic MEs are shown to participate in protein-coding regions, regulating gene expression, and alternative splicing.

Overall, this data concludes that polymorphic MEs impact the evolution of humans and may contribute to phenotypic diversity among human populations and individuals. Future studies may focus on extending the identification of such ME polymorphisms by utilizing the richer and better-quality personal genome data and elucidating the specific functions of these polymorphic MEs.

References

Abascal, F., Tress, M. L., & Valencia, A. (2015). Alternative splicing and co-option of transposable elements: the case of TMPO/LAP2 alpha and ZNF451 in mammals. *Bioinformatics, 31*(14), 2257-2261. doi:10.1093/bioinformatics/btv132

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., . . . Genomes Project, C. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Anwar, S. L., Wulaningsih, W., & Lehmann, U. (2017). Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *International Journal of Molecular Sciences, 18*(5). doi:10.3390/ijms18050974

Arcot, S. S., Shaikh, T. H., Kim, J. Y., Bennett, L., Alegriahartman, M., Nelson, D. O., . . . Batzer, M. A. (1995). SEQUENCE DIVERSITY AND CHROMOSOMAL DISTRIBUTION OF YOUNG ALU REPEATS. *Gene, 163*(2), 273-278. doi:10.1016/0378-1119(95)00317-y

Athanasiadis, A., Rich, A., & Maas, S. (2004). Widespread A-to-I RNA editing of alu-containing mRNAs in the human transcriptome. *Plos Biology, 2*(12), 2144-2158. doi:10.1371/journal.pbio.0020391

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., . . . Schloss, J. A. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Ayarpadikannan, S., & Kim, H.-S. (2014). The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics, 12*(3), 98-104. doi:10.5808/gi.2014.12.3.98

Ayarpadikannan, S., Lee, H.-E., Han, K., & Kim, H.-S. (2015). Transposable element-driven transcript diversification and its relevance to genetic disorders. *Gene, 558*(2), 187-194. doi:10.1016/j.gene.2015.01.039

Bannert, N., & Kurth, R. (2004). Retroelements and the human genome: New perspectives on an old relation. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 14572-14579. doi:10.1073/pnas.0404838101

Batzer, M. A., Arcot, S. S., Phinney, J. W., AlegriaHartman, M., Kass, D. H., Milligan, S. M., . . . Stoneking, M. (1996). Genetic variation of recent Alu insertions in human populations. *Journal of Molecular Evolution, 42*(1), 22-29. doi:10.1007/bf00163207

Batzer, M. A., & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics, 3*(5), 370-379. doi:10.1038/nrg798

Batzer, M. A., Gudi, V. A., Mena, J. C., Foltz, D. W., Herrera, R. J., & Deininger, P. L. (1991). AMPLIFICATION DYNAMICS OF HUMAN-SPECIFIC (HS) ALU FAMILY MEMBERS. *Nucleic Acids Research, 19*(13), 3619-3623. doi:10.1093/nar/19.13.3619

Belancio, V. P., Hedges, D. J., & Deininger, P. (2008). Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Research, 18*(3), 343-358. doi:10.1101/gr.5558208

Bennettt, E. A., Coleman, L. E., Tsui, C., Pittard, W. S., & Devine, S. E. (2004). Natural genetic variation caused by transposable elements in humans. *Genetics, 168*(2), 933-951. doi:10.1534/genetics.104.031757

Bestor, T. H., & Bourc'his, D. (2004). Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harbor Symposia on Quantitative Biology, 69*, 381-387. doi:10.1101/sqb.2004.69.381

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., . . . Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology, 19*. doi:10.1186/s13059-018-1577-z

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica Et Biophysica Acta-Molecular Basis of Disease, 1842*(10), 1932-1941. doi:10.1016/j.bbadis.2014.06.015

Callinan, P. A., Wang, J. X., Herke, S. W., Garber, R. K., Liang, P., & Batzer, M. A. (2005). Alu retrotransposition-mediated deletion. *Journal of Molecular Biology, 348*(4), 791-800. doi:10.1016/j.jmb.2005.02.043

Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B., . . . Batzer, M. A. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *Journal of Molecular Biology, 311*(1), 17-40. doi:10.1006/jmbi.2001/4847

Chen, L. L., DeCerbo, J. N., & Carmichael, G. G. (2008). Alu element-mediated gene silencing. *Embo Journal, 27*(12), 1694-1705. doi:10.1038/emboj.2008.94

Chen, X. (2019). Genome analysis ERVcaller : identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. (March), 1-10. doi:10.1093/bioinformatics/btz205

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics, 18*(2), 71-86. doi:10.1038/nrg.2016.139

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics, 10*(10), 691-703. doi:10.1038/nrg2640

Cowley, M., & Oakey, R. J. (2013). Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *Plos Genetics, 9*(1). doi:10.1371/journal.pgen.1003234

de Souza, F. S. J., Franchini, L. F., & Rubinstein, M. (2013). Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? *Molecular Biology and Evolution, 30*(6), 1239-1251. doi:10.1093/molbev/mst045

Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biology, 12*(12). doi:10.1186/gb-2011-12-12-236

Deininger, P. L., & Batzer, M. A. (2002). Mammalian Retroelements. *Genome Research, 12*(10), 1455-1465. doi:10.1101/gr.282402

Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development, 13*(6), 651-658. doi:10.1016/j.gde.2003.10.013

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., & Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Research, 16*(12), 1548-1556. doi:10.1101/gr.5565706

Dewannieux, M., & Heidmann, T. (2005). LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. *Cytogenetic and Genome Research, 110*(1-4), 35-48. doi:10.1159/000084936

Elbarbary, R. A., Lucas, B. A., & Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science, 351*(6274). doi:10.1126/science.aac7247

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA, 6*. doi:10.1186/s13100-015-0055-3

Ewing, A. D. (2017). The mobilisation of processed transcripts in germline and somatic tissues. *Springer*, 95-106. doi:https://doi.org/10.1007/978-3-319-48344-3_4

Ewing, A. D., Ballinger, T. J., Earl, D., Harris, C. C., Ding, L., Wilson, R. K., & Haussler, D. (2013). Retrotransposition of gene transcripts leads to structural variation in mammalian genomes Article type Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biology, 14*(3), 1-14. doi:10.1186/gb-2013-14-3-r22

Ewing, A. D., & Kazazian, H. H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Research, 20*(9), 1262-1270. doi:10.1101/gr.106419.110

Fedele, A. O., Filocamo, M., Rocco, M. D., Sersale, G., Lübke, T., Natale, P., . . . Ballabio, A. (2007). Mutational Analysis of the HGSNAT Gene in Italian Patients With Mucopolysaccharidosis IIIC (Sanfilippo C Syndrome). *959*(November 2006). doi:10.1002/humu.9488

Finnegan, D. J. (2012). Retrotransposons. *Current Biology, 22*(11), R432-R437. doi:10.1016/j.cub.2012.04.025

Flasch, D. A., Macia, A., Sanchez, L., Ljungman, M., Heras, S. R., Garcia-Perez, J. L., . . . Moran, J. V. (2019). Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell, 177*(4), 837-+. doi:10.1016/j.cell.2019.02.050

Fujiwara, H. (2015). Site-specific non-LTR retrotransposons. *Microbiology Spectrum, 3*(2). doi:10.1128/microbiolspec.MDNA3-0001-2014

Garcia-Perez, J. L., Morell, M., Scheys, J. O., Kulpa, D. A., Morell, S., Carter, C. C., . . . Moran, J. V. (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature, 466*(7307), 769-773. doi:10.1038/nature09209

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., . . . Genomes Project, C. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Research, 27*(11), 1916-1929. doi:10.1101/gr.218032.116

Girilishena, Y. (2017). Complete computational sequence characterization of mobile element variations in the human genome using meta-personal genome data. MSc Thesis, Brock University

Goodier, J. L., Ostertag, E. M., & Kazazian, H. H. (2000). Transduction of 3 '-flanking sequences is common in L1 retrotransposition. *Human Molecular Genetics, 9*(4), 653-657. doi:10.1093/hmg/9.4.653

Grandi, F. C., Rosser, J. M., Newkirk, S. J., Yin, J., Jiang, X. L., Xing, Z., . . . An, W. F. (2015). Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites. *Genome Research, 25*(8), 1135-1146. doi:10.1101/gr.185132.114

Hamdi, H. K., Reznik, J., Castellon, R., Atilano, S. R., Ong, J. M., Udar, N., . . . Kenney, M. C. (2002). Alu DNA polymorphism in ACE gene is protective for age-related macular degeneration. *295,* 668-672.

Han, J. S. (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA, 1*. doi:10.1186/1759-8753-1-15

Hancks, D. C., & Kazazian, H. H. (2010). SVA retrotransposons: Evolution and genetic instability. *Seminars in Cancer Biology, 20*(4), 234-245. doi:10.1016/j.semcancer.2010.04.001

Hancks, D. C., & Kazazian, H. H. (2012). Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development, 22*(3), 191-203. doi:10.1016/j.gde.2012.02.006

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA, 7*. doi:10.1186/s13100-016-0065-9

Hasler, J., & Strub, K. (2006). Alu elements as regulators of gene expression. *Nucleic Acids Research, 34*(19), 5491-5497. doi:10.1093/nar/gkl706

Hoffmann, F. G., McGuire, L. P., Counterman, B. A., & Ray, D. A. (2015). Transposable elements and small RNAs: Genomic fuel for species diversity. *Mobile genetic elements*(5(5)), 63-66.

Howe, E., Holton, K., Nair, S., Schlauch, D., Sinha, R., & Quackenbush, J. (2010). *MeV: MultiExperiment Viewer*. Boston, MA: Springer.

Huda, A., Marino-Ramirez, L., & Jordan, I. K. (2010). Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mobile DNA, 1*. doi:10.1186/1759-8753-1-2

Jahic, A., Erichsen, A. K., Deufel, T., Tallaksen, C. M., & Beetz, C. (2016). A polymorphic Alu insertion that mediates distinct disease-associated deletions. *European Journal of Human Genetics, 24*(9), 1371-1374. doi:10.1038/ejhg.2016.20

Kang, H. X., Zhu, D., Lin, R. M., Opiyo, S. O., Jiang, N., Shiu, S. H., & Wang, G. L. (2016). A novel method for identifying polymorphic transposable elements via scanning of high-throughput short reads. *DNA Research, 23*(3), 241-251. doi:10.1093/dnares/dsw011

Kazazian, H. (2004). Mobile elements: Drivers of genome evolution. *Science, 303*(5664), 1626-1632. doi:10.1126/science.1089670

Kazazian, H., & Moran, J. (2017). Mobile DNA in Health and Disease. *New England Journal of Medicine, 377*(4), 361-370. doi:10.1056/NEJMra1510092

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). HEMOPHILIA-A RESULTING FROM DENOVO INSERTION OF L1 SEQUENCES REPRESENTS A NOVEL MECHANISM FOR MUTATION IN MAN. *Nature, 332*(6160), 164-166. doi:10.1038/332164a0

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics, 29*(3), 389-390. doi:10.1093/bioinformatics/bts697

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research, 12*(4), 656-664. doi:10.1101/gr.229202

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., . . . Eichler, E. E. (2010). A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. *Cell, 143*(5), 837-847. doi:10.1016/j.cell.2010.10.027

Kim, S., Cho, C.-S., Han, K., & Lee, J. (2016). Structural Variation of Alu Element and Human Disease. *Genomics & Informatics, 14*(3), 70-77.

Kobayashi, K., Nakahori, Y., Miyake, M., Matsumura, K., Kondo-Iida, E., Nomura, Y., . . . Toda, T. (1998). An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature, 394*(6691), 388-392.

Konkel, M. K., & Batzer, M. A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in Cancer Biology, 20*(4), 211-221. doi:10.1016/j.semcancer.2010.03.001

Konkel, M. K., Wang, J., Liang, P., & Batzer, M. A. (2007). Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. *Gene, 390*(1-2), 28-38. doi:10.1016/j.gene.2006.07.040

Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology, 11*(1), 39-46. doi:10.1016/s0959-440x(00)00167-6

Lander, E. S., Int Human Genome Sequencing, C., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., . . . Int Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., . . . Schatz, M. C. (2016). Third-generation sequencing and the future of genomics. (Table 1).

Lee, H.-E., Ayarpadikannan, S., & Kim, H.-S. (2015). Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates. *Genes & Genetic Systems, 90*(5), 245-257. doi:10.1266/ggs.15-00016

Leffler, E. M., Gao, Z. Y., Pfeifer, S., Segurel, L., Auton, A., Venn, O., . . . Przeworski, M. (2013). Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science, 339*(6127), 1578-1582. doi:10.1126/science.1234070

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Sam, T. (2009). The Sequence Alignment / Map format and SAMtools. *25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, K. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics, 19*(12), 1585-1586. doi:10.1093/bioinformatics/btg192

Li, W., Yang, W., & Wang, X.-J. (2013). Pseudogenes: Pseudo or Real Functional Elements? *Journal of Genetics and Genomics, 40*(4), 171-177. doi:10.1016/j.jgg.2013.03.003

Lonnig, W. E., & Saedler, H. (2002). Chromosome rearrangements and transposable elements. *Annual Review of Genetics, 36*, 389-410. doi:10.1146/annurev.genet.36.040202.092802

Lowe, C. B., Bejerano, G., & Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 8005-8010. doi:10.1073/pnas.0611223104

Luan, D. D., Korman, M. H., Jakubczak, J. L., & Eickbush, T. H. (1993). REVERSE TRANSCRIPTION OF R2BM RNA IS PRIMED BY A NICK AT THE CHROMOSOMAL TARGET SITE - A MECHANISM FOR NON-LTR RETROTRANSPOSITION. *Cell, 72*(4), 595-605. doi:10.1016/0092-8674(93)90078-5

Macia, A., Munoz-Lopez, M., Luis Cortes, J., Hastings, R. K., Morell, S., Lucena-Aguilar, G., . . . Luis Garcia-Perez, J. (2011). Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells. *Molecular and Cellular Biology, 31*(2), 300-316. doi:10.1128/mcb.00561-10

McClintock, B. (1950). THE ORIGIN AND BEHAVIOR OF MUTABLE LOCI IN MAIZE. *Proceedings of the National Academy of Sciences of the United States of America, 36*(6), 344-355. doi:10.1073/pnas.36.6.344

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. *Plos Genetics, 5*(10). doi:10.1371/journal.pgen.1000686

Metspalu, A. (2004). The Estonian Genome Project. *Drug Development Research, 62*(2), 97-101. doi:10.1002/ddr.10371

Mighell, A. J., Smith, N. R., Robinson, P. A., & Markham, A. F. (2000). Vertebrate pseudogenes. *Febs Letters, 468*(2-3), 109-114. doi:10.1016/s0014-5793(00)01199-6

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics, 23*(4), 183-191. doi:10.1016/j.tig.2007.02.006

Mills, R. E., Bennett, E. A., Iskow, R. C., Luttig, C. T., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). Recently mobilized Transposons in the human and chimpanzee Genomes. *American Journal of Human Genetics, 78*(4), 671-679. doi:10.1086/501028

Mine, M., Chen, J. M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P. S., . . . Marsac, C. (2007). A large genomic deletion in the PDHX gene caused by the retrotransposlitional insertion of a full-length LINE-1 element. *Human Mutation, 28*(2), 137-142. doi:10.1002/humu.20449

Mola, G., Vela, E., Fernandez-Figueras, M. T., Isamat, M., & Munoz-Marmol, A. M. (2007). Exonization of Alu-generated splice variants in the survivin gene of human and non-human primates. *Journal of Molecular Biology, 366*(4), 1055-1063. doi:10.1016/j.jmb.2006.11.089

Moran, J. V., DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science, 283*(5407), 1530-1534. doi:10.1126/science.283.5407.1530

Muro, E. M., Mah, N., & Andrade-Navarro, M. A. (2011). Functional evidence of post-transcriptional regulation by pseudogenes. *Biochimie, 93*(11), 1916-1921. doi:10.1016/j.biochi.2011.07.024

Nekrutenko, A., & Li, W. H. S. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics, 17*(11), 619-621. doi:10.1016/s0168-9525(01)02445-3

Obbard, D. J., Gordon, K. H. J., Buck, A. H., & Jiggins, F. M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philosophical Transactions of the Royal Society B-Biological Sciences, 364*(1513), 99-115. doi:10.1098/rstb.2008.0168

Ostertag, E. M., & Kazazian, H. (2001). Biology of mammalian L1 retrotransposons. *Annual Review of Genetics, 35*, 501-538. doi:10.1146/annurev.genet.35.102401.091032

Otieno, A. C., Carter, A. B., Hedges, D. J., Walker, J. A., Ray, D. A., Garber, R. K., . . . Batzer, M. A. (2004). Analysis of the human Alu Ya-lineage. *Journal of Molecular Biology, 342*(1), 109-118. doi:10.1016/j.jmb.2004.07.016

Pavlicek, A., Gentles, A. J., Paces, J., Paces, V., & Jurka, J. (2006). Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends in Genetics, 22*(2), 69-73. doi:10.1016/j.tig.2005.11.005

Payer, L. M., Steranka, J. P., Yang, W. R., Kryatova, M., Medabalimi, S., Ardeljan, D., . . . Burns, K. H. (2017). Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proceedings of the National Academy of Sciences of the United States of America, 114*(20), E3984-E3992. doi:10.1073/pnas.1704117114

Pedersen, I. M., & Zisoulis, D. G. (2016). Transposable elements and miRNA: Regulation of genomic stability and plasticity. *Mobile genetic elements, 6*(3), e1175537-e1175537. doi:10.1080/2159256x.2016.1175537

Peixoto, A., Pinheiro, M., Massena, L., Santos, C., Pinto, P., Rocha, P., . . . Teixeira, M. R. (2013). Genomic characterization of two large Alu-mediated rearrangements of the BRCA1 gene. *Journal of Human Genetics, 58*(2), 78-83. doi:10.1038/jhg.2012.137

Pickeral, O. K., Makalowski, W., Boguski, M. S., & Boeke, J. D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Research, 10*(4), 411-415. doi:10.1101/gr.10.4.411

Platt, R. N., Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research, 26*(1-2), 25-43. doi:10.1007/s10577-017-9570-z

Puurand, T., Kukuškina, V., Pajuste, F.-d., & Remm, M. (2019). AluMine : alignment-free method for the discovery of polymorphic Alu element insertions.

Qin, S., Jin, P., Zhou, X., Chen, L., & Ma, F. (2015). The Role of Transposable Elements in the Origin and Evolution of MicroRNAs in Human. *Plos One, 10*(6). doi:10.1371/journal.pone.0131365

Quinlan, A. R., & Hall, I. M. (2010). BEDTools : a flexible suite of utilities for comparing genomic features. *26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Raghavan, M. S. M. H. K. S. S. R. S. D. M., & Eriksson, A. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *349*(6250), 1-36. doi:10.1126/science.aab3884.Genomic

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., . . . Schumann, G. G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Research, 40*(4), 1666-1683. doi:10.1093/nar/gkr863

Rand, E., Ben-Porath, I., Keshet, I., & Cedar, H. (2004). CTCF elements direct allele-specific undermethylation at the imprinted H19 locus. *Current Biology, 14*(11), 1007-1012. doi:10.1016/j.cub.2004.05.041

Rishishwar, L., Villa, C. E. T., & Jordan, I. K. (2015). Transposable element polymorphisms recapitulate human evolution. *Mobile DNA, 6*. doi:10.1186/s13100-015-0052-6

Roberts, T. C. (2014). The MicroRNA Biology of the Mammalian Nucleus. (July), 1-8. doi:10.1038/mtna.2014.40

Roy, A. M., Carroll, M. L., Kass, D. H., Nguyen, S. V., Salem, A. H., Batzer, M. A., & Deininger, P. L. (1999). Recently integrated human Alu repeats: finding needles in the haystack. *Genetica, 107*(1-3), 149-161. doi:10.1023/a:1003941704138

Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H., . . . Deininger, P. L. (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics, 159*(1), 279-290.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, A., Kutter, C., . . . Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages (vol 148, pg 335, 2012). *Cell, 148*(4), 832-832. doi:10.1016/j.cell.2012.02.001

Schorn, A. J., Gutbrod, M. J., LeBlanc, C., & Martienssen, R. (2017). LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell, 170*(1), 61-+. doi:10.1016/j.cell.2017.06.013

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., . . . Hayes, V. M. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature, 463*(7283), 943-947. doi:10.1038/nature08795

Scott, E. C., & Devine, S. E. (2017). The Role of Somatic L1 Retrotransposition in Human Cancers. *Viruses-Basel, 9*(6). doi:10.3390/v9060131

Shin, W., Lee, J., Son, S. Y., Ahn, K., Kim, H. S., & Han, K. (2013). Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome. *Plos One, 8*(4). doi:10.1371/journal.pone.0060605

Shintani, M., Tada, M., Kobayashi, T., Kajiho, H., Kontani, K., & Katada, T. (2007). Characterization of Rab45/RASEF containing EF-hand domain and a coiled-coil motif as a self-associating GTPase. *Biochemical and Biophysical Research Communications, 357*(3), 661-667. doi:10.1016/j.bbrc.2007.03.206

Solyom, S., & Jr, H. H. K. (2012). Mobile elements in the human genome : implications for disease. *1*, 1-8.

Spengler, R. M., Oakley, C. K., & Davidson, B. L. (2014). Functional microRNAs and target sites are created by lineage-specific transposition. *Human Molecular Genetics, 23*(7), 1783-1793. doi:10.1093/hmg/ddt569

Stewart, C., Kural, D., Stromberg, M. P., Walker, J. A., Konkel, M. K., Stutz, A. M., . . . Genomes, P. (2011). A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *Plos Genetics, 7*(8). doi:10.1371/journal.pgen.1002236

Strachan, T., & Read, A. (2011). Human Molecular Genetics 4th Edition. *Human Molecular Genetics 4th Edition*, 1-781.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Genomes Project, C. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-+. doi:10.1038/nature15394

Sverdlov, E. D., & Wiley, J. (2000). Retroviruses and primate evolution. 161-171.

Szak, S. T., Pickeral, O. K., Makalowski, W., Boguski, M. S., Landsman, D., & Boeke, J. D. (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biology, 3*(10).

Szitenberg, A., Cha, S., Opperman, C. H., Bird, D. M., Blaxter, M. L., & Lunt, D. H. (2016). Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biology and Evolution, 8*(9), 2964-2978. doi:10.1093/gbe/evw208

Tang, W., Mun, S., Joshi, A., Han, K., & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14Mbp sequence increase. *DNA research : an international journal for rapid publication of reports on genes and genomes, 25*(5), 521-533. doi:10.1093/dnares/dsy022

Taniguchi-Ikeda, M., Kobayashi, K., Kanagawa, M., Yu, C., Mori, K., Oda, T., . . . Toda, T. (2011). Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. *Nature, 478*(7367), 127-U143. doi:10.1038/nature10456

Tucker, B. A., Scheetz, T. E., Mullins, R. F., Deluca, A. P., Hoffmann, J. M., Johnston, R. M., . . . Stone, E. M. (2011). Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell- associated kinase (MAK) as a cause of retinitis pigmentosa. *108*(34), 569-576. doi:10.1073/pnas.1108918108

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., . . . Genomes, P. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *Bmj-British Medical Journal, 361*, 7. doi:10.1136/bmj.k1687

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell, 154*(1), 26-46. doi:10.1016/j.cell.2013.06.020

Valencia, J. D., & Girgis, H. Z. (2019). LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. *Bmc Genomics, 20*. doi:10.1186/s12864-019-5796-9

van de Lagemaat, L. N., Gagnier, L., Medstrand, P., & Mager, D. L. (2005). Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Research, 15*(9), 1243-1249. doi:10.1101/gr.3910705

Venter, J. C. (2001). The sequence of the human genome (vol 292, pg 1304, 2001). *Science, 292*(5523), 1838-1838.

Visser, M., Palstra, R. J., & Kayser, M. (2014). Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. *Human Molecular Genetics, 23*(21), 5750-5762. doi:10.1093/hmg/ddu289

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M. A., & Liang, P. (2006). dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation, 27*(4), 323-329. doi:10.1002/humu.20307

Wang, J. X., Song, L., Gonder, M. K., Azrak, S., Ray, D. A., Batzer, M. A., . . . Liang, P. (2006). Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. *Gene, 365*, 11-20. doi:10.1016/j.gene.2005.09.031

Wang, L., & Jordan, I. K. (2018). Transposable element activity, genome regulation and human health. *Current Opinion in Genetics & Development, 49*, 25-33. doi:10.1016/j.gde.2018.02.006

Wang, L., Norris, E. T., & Jordan, I. K. (2017). Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. *Frontiers in Microbiology, 8*. doi:10.3389/fmicb.2017.01418

Wang, L., Rishishwar, L., Marino-Ramirez, L., & Jordan, I. K. (2017). Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Research, 45*(5), 2318-2328. doi:10.1093/nar/gkw1286

Watanabe, M., Kobayashi, K., Jin, F., Park, K. S., Yamada, T., Tokunaga, K., & Toda, T. (2005). Founder SVA Retrotransposal Insertion in Fukuyama-Type Congenital Muscular Dystrophy and Its Origin in Japanese and Northeast Asian Populations. *348*(June), 344-348. doi:10.1002/ajmg.a.30978

Wessler, S. R. (2006). Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America, 103*(47), 17600-17601. doi:10.1073/pnas.0607612103

Wimmer, K., Callens, T., Wernstedt, A., & Messiaen, L. (2011). The NF1 Gene Contains Hotspots for L1 Endonuclease-Dependent De Novo Insertion. *Plos Genetics, 7*(11). doi:10.1371/journal.pgen.1002371

Witherspoon, D. J., Xing, J. C., Zhang, Y. H., Watkins, W. S., Batzer, M. A., & Jorde, L. B. (2010). Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *Bmc Genomics, 11*. doi:10.1186/1471-2164-11-410

Wu, J. T., Lee, W. P., Ward, A., Walker, J. A., Konkel, M. K., Batzer, M. A., & Marth, G. T. (2014). Tangram: a comprehensive toolbox for mobile element insertion detection. *Bmc Genomics, 15*. doi:10.1186/1471-2164-15-795

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The Ensembl Regulatory Build. 1-8. doi:10.1186/s13059-015-0621-5

Zhang, Z., Harrison, P., & Gerstein, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Research, 12*(10), 1466-1482. doi:10.1101/gr.331902

Appendix

Table S1. Identified cases of transductions mediated by polymorphic ME insertions

dbRIP ID	Insertion	ME type	5' Transduction Length	3' Transduction Length
2000241	ref	L1	0	2
2000274	ref	L1	0	39
2000320	ref	L1	0	9
2000335	non-ref	L1	2	21
2000363	ref	L1	0	30
2000436	non-ref	L1	2	0
2000437	non-ref	L1	3	0
2000497	non-ref	L1	10	41
2000507	non-ref	L1	2	3
2000508	non-ref	L1	3	8
2000516	non-ref	L1	0	39
2000523	non-ref	L1	1	11
2000524	non-ref	L1	0	14
2000525	non-ref	L1	0	433
2000526	non-ref	L1	9	54
2000544	non-ref	L1	0	48
2000551	non-ref	L1	1	24
2000565	non-ref	L1	0	9
2000567	non-ref	L1	0	52
2000568	non-ref	L1	0	53
2000581	non-ref	L1	0	5
2000584	non-ref	L1	1	27
2000598	non-ref	L1	1	11
2000641	ref	L1	2	2
2000653	ref	L1	1	18
4000028	ref	LTR	23	36

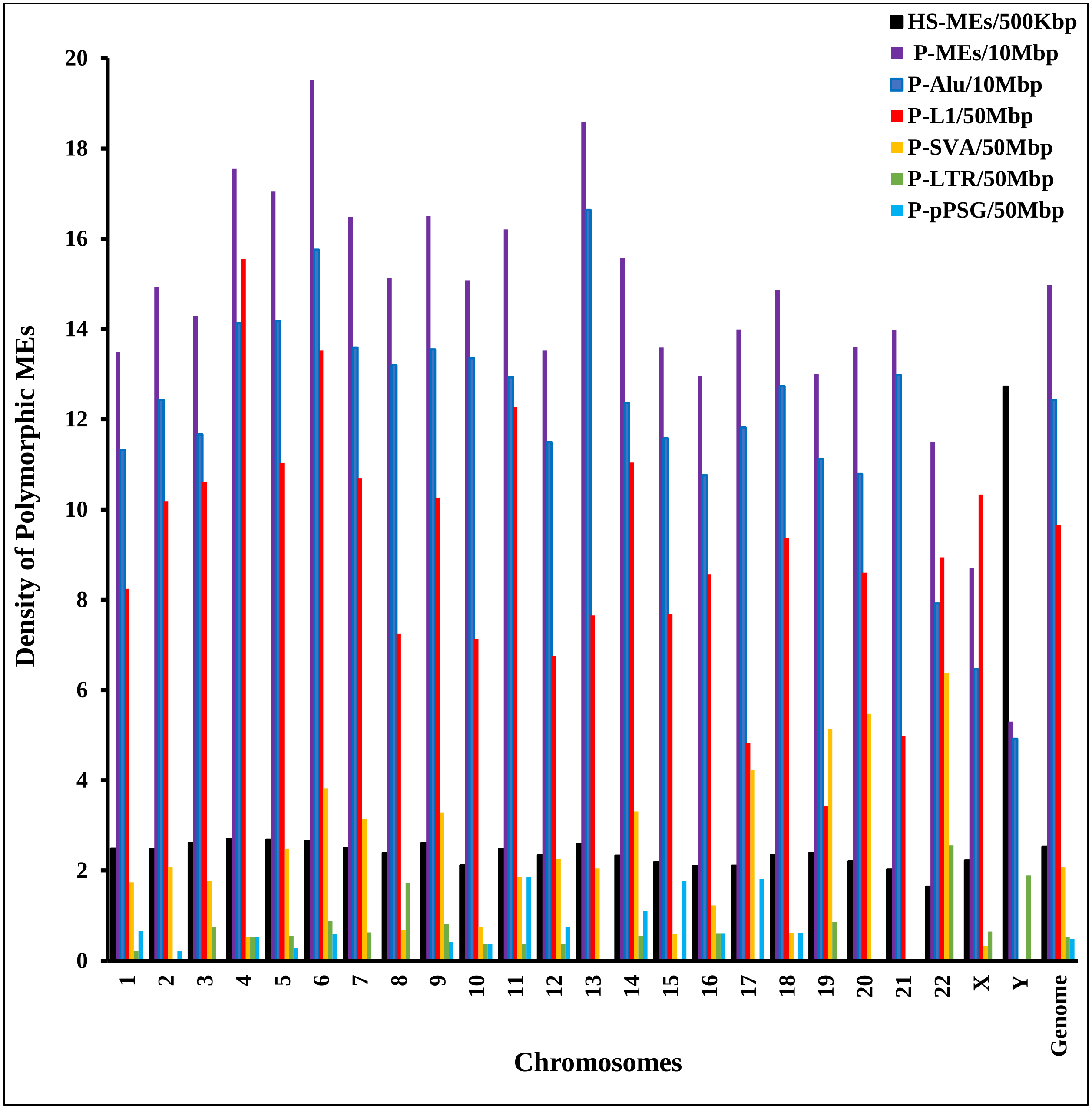

Figure S1. Overview of genome distribution patterns of polymorphic MEs (P-MEs) and each polymorphic ME type. Bar plots showing the density of P-MEs for each type of MEs among 24 chromosomes and the entire genome.

Table S2. Activity level of ME subfamilies based on polymorphic MEs

Family	Subfamily	Total MEs	Polymorphic MEs	Polymorphic ME Ratio
Alu	AluYa5	3861	1315	34.1%
	AluYb8	2828	822	29.1%
	AluYb9	327	111	33.9%
	AluYd8	237	19	8.0%
	AluYg6	835	106	12.7%
	AluYi6	455	23	5.1%
	AluYk12	201	1	0.5%
	AluYa8	343	36	10.5%
	AluYe5	1318	60	4.6%
	AluYi6 4d	149	0	0.0%
	AluYh7	153	0	0.0%
	AluYc3	543	2	0.4%
	AluYk11	1256	2	0.2%
	AluYk4	1010	1	0.1%
	AluYh3	2627	0	0.0%
	AluYe6	194	0	0.0%
	AluYc	4521	19	0.4%
	AluYh9	142	15	10.6%
	AluYh3a3	313	0	0.0%
	Alu	4280	0	0.0%
	AluYk3	1152	0	0.0%
	AluY	102,844	428	0.4%
	AluYj4	3487	0	0.0%
L1	L1HS	1346	447	33.2%
	L1PA2	4096	32	0.8%
	L1P1	2851	2	0.1%
	L1PA3	8780	16	0.2%
	L1PA8	6541	1	0.0%
	L1P2	1231	1	0.1%
	L1P	140	0	0.0%
SVA	SVA_D	1325	13	1.0%
	SVA_F	821	56	6.8%
	SVA_E	595	42	7.1%
	SVA_C	418	2	0.5%
	SVA_A	1001	3	0.3%
	SVA_C	768	2	0.3%
LTR	ERVK	7369	10	0.1%
	ERV1	103982	0	0.0%
PPSG	Processed	12079	28	0.2%

Table S3. Chromosome distributions of polymorphic MEs organized by ME type

Chr	Chr Length (Mb)	All MEs	All MEs/1Kbp	All HS-MEs	HS-MEs/500Kbp	All P-MEs	pME/10Mb	P-Alu	P-Alu/10Mbp	P-L1	P-L1/50Mbp	P-SVA	P-SVA/50Mbp	P-LTR	P-LTR/50Mbp	P-PPSG	P-PPSG/50Mbp	Gene	Gene per Mb
chr1	230.48	227225	9.86	1147	2.5	311	13.49	261	11.32	38	8.24	8	1.74	1	0.22	3	0.65	5620	24.38
chr2	240.55	224251	9.32	1190	2.5	359	14.92	299	12.43	49	10.19	10	2.08	0	0.00	1	0.21	4264	17.73
chr3	198.10	181925	9.18	1036	2.6	283	14.29	231	11.66	42	10.60	7	1.77	3	0.76	0	0.00	3230	16.30
chr4	189.75	169168	8.92	1025	2.7	333	17.55	268	14.12	59	15.55	2	0.53	2	0.53	2	0.53	2699	14.22
chr5	181.27	163567	9.02	970	2.7	309	17.05	257	14.18	40	11.03	9	2.48	2	0.55	1	0.28	3086	17.02
chr6	170.08	155616	9.15	902	2.7	332	19.52	268	15.76	46	13.52	13	3.82	3	0.88	2	0.59	3160	18.58
chr7	158.97	161811	10.18	794	2.5	262	16.48	216	13.59	34	10.69	10	3.15	2	0.63	0	0.00	3084	19.40
chr8	144.77	134045	9.26	692	2.4	219	15.13	191	13.19	21	7.25	2	0.69	5	1.73	0	0.00	2507	17.32
chr9	121.79	120268	9.87	635	2.6	201	16.50	165	13.55	25	10.26	8	3.28	2	0.82	1	0.41	2443	20.06
chr10	133.26	130160	9.77	564	2.1	201	15.08	178	13.36	19	7.13	2	0.75	1	0.38	1	0.38	2433	18.26
chr11	134.53	121067	9.00	667	2.5	218	16.20	174	12.93	33	12.26	5	1.86	1	0.37	5	1.86	3434	25.53
chr12	133.14	134351	10.09	625	2.3	180	13.52	153	11.49	18	6.76	6	2.25	1	0.38	2	0.75	3104	23.31
chr13	97.98	87599	8.94	507	2.6	182	18.57	163	16.64	15	7.65	4	2.04	0	0.00	0	0.00	1464	14.94
chr14	90.57	88829	9.81	423	2.3	141	15.57	112	12.37	20	11.04	6	3.31	1	0.55	2	1.10	2341	25.85
chr15	84.64	86430	10.21	370	2.2	115	13.59	98	11.58	13	7.68	1	0.59	0	0.00	3	1.77	2310	27.29
chr16	81.81	96569	11.80	344	2.1	106	12.96	88	10.76	14	8.56	2	1.22	1	0.61	1	0.61	2639	32.26
chr17	82.92	99020	11.94	350	2.1	116	13.99	98	11.82	8	4.82	7	4.22	0	0.00	3	1.81	3153	38.02
chr18	80.09	68671	8.57	376	2.3	119	14.86	102	12.74	15	9.36	1	0.62	0	0.00	1	0.62	1227	15.32
chr19	58.44	90253	15.44	280	2.4	76	13.00	65	11.12	4	3.42	6	5.13	1	0.86	0	0.00	3098	53.01
chr20	63.94	66872	10.46	282	2.2	87	13.61	69	10.79	11	8.60	7	5.47	0	0.00	0	0.00	1496	23.40
chr21	40.09	38474	9.60	162	2.0	56	13.97	52	12.97	4	4.99	0	0.00	0	0.00	0	0.00	903	22.53
chr22	39.16	44598	11.39	128	1.6	45	11.49	31	7.92	7	8.94	5	6.38	2	2.55	0	0.00	1417	36.19
chrX	154.89	156577	10.11	689	2.2	135	8.72	100	6.46	32	10.33	1	0.32	2	0.65	0	0.00	2550	16.46
chrY	26.42	27470	10.40	672	12.7	14	5.30	13	4.92	0	0.00	0	0.00	1	1.89	0	0.00	601	22.75
Genome	2937.64	2874816	9.79	14830	2.5	4400	14.98	3652	12.43	567	9.65	122	2.08	31	0.53	28	0.48	62263	21.19

Table S4. Monte-Carlo simulation output for the chromosomal distribution of polymorphic MEs

Chromosome	Observed	Mean	SD	Z-score	P-value
1	311	346.20	15.78	-2.23	0.008
2	359	367.84	16.36	-0.54	0.290
3	283	302.20	16.86	-1.14	0.118
4	333	289.09	16.04	2.74	0.003
5	309	270.57	13.10	2.93	0.002
6	332	258.57	15.73	4.67	1.73E-06
7	262	236.30	14.82	1.73	0.042
8	219	219.93	13.48	-0.07	0.472
9	201	184.81	12.93	1.25	0.106
10	201	200.23	15.09	0.05	0.479
11	218	202.24	15.36	1.03	0.153
12	180	197.29	12.06	-1.43	0.066
13	182	145.28	13.65	2.69	0.004
14	141	136.70	11.68	0.37	0.356
15	115	125.75	12.19	-0.88	0.181
16	106	121.65	10.98	-1.43	0.067
17	116	120.51	10.35	-0.44	0.328
18	119	116.48	10.54	0.24	0.405
19	76	85.44	9.45	-1.00	0.150
20	87	91.65	8.83	-0.53	0.295
21	56	54.87	7.05	0.16	0.436
22	45	52.78	6.66	-1.17	0.112
X	135	233.98	15.16	-6.53	2.78E-16
Y	14	38.64	5.94	-4.15	6.25E-07

Table S5. Polymorphic ME entries failed to convert from hg19 to hg38 using liftOver tools

Chromosome	Start	End	dbRIP ID	Reason for conversion fail
chrX	149551149	149551472	1001383	Partially deleted in new
chr2	91689543	91689850	1001222	Split in new
chrX	45587575	45587866	1002651	Partially deleted in new
chr10	49022912	49023239	1001087	Split in new
chr9	41011834	41012140	1001372	Split in new
chrX	114670506	114670787	1000763	Partially deleted in new
chr17	35143438	35143742	1002589	Partially deleted in new
chr22	18047376	18047718	1000788	Split in new
chr22	17884244	17884561	1000118	Partially deleted in new
chr14	23105130	23107948	3000092	Split in new
chr1	1584472	1584473	1001461	Deleted in new

Table S6. Specific function among genes containing polymorphic MEs within CDS regions

Gene Names	**Full Name**	**Function**
RASEF	RAS and EF-Hand Domaining Containing	Protein transport and GTPase activity
FSTL4	Follistatin-Like 4	Calcium ion binding
AL135745.1	NA	Antibody
AL133335.1	NA	Antibody
HGSNAT	Heparan-alpha-glucosaminide N-acetyltransferase	Lysomal degradation of heparin sulfate
ZNF83	Zinc Finger Protein 83	DNA binding TF activity
HMBS	Hydroxymethylbilane Synthase	Heme biosynthetic pathway
BRCA2	Breast Cancer Susceptibility Gene 2	Genome stability, dsDNA repair
F9	Coagulation Factor IX	Acts as zymogen in the blood
BCHE	Butyrylcholinesterase	Protein binding and hydrolase activity
CLCN5	Chloride Voltage-Gated Channel 5	Ion channel and antiporter activity
DMD	Dystrophin	Calcium ion binding and structural constituent of cytoskeleton
SPTA1	Spectrin Alpha, Erythrocytic 1	Calcium ion binding and actin filament binding
CSNK2A3	Casein Kinase 2 Alpha 3	Transferase and tyrosine kinase activity