Genome-wide analysis of mobile element insertions in human genomes

Running title: Mobile element insertion map of 5,675 genomes

Yiwei Niu[1,2,5], Xueyi Teng[1,3,5], Yirong Shi[1,3], Yanyan Li[1,2], Yiheng Tang[1,3], Peng Zhang[1], Huaxia Luo[1], Quan Kang[1], The Han100K Initiative[§], Tao Xu[2,4*], Shunmin He[1,3,6*]

1 Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.

2 College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

3 University of Chinese Academy of Sciences, Beijing 100049, China.

4 National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, 100101, China.

5 These authors contributed equally to this work.

6 Lead contact.

* Corresponding author. Email: xutao@ibp.ac.cn (T. X), heshunmin@ibp.ac.cn (S.M. H)

§ Full list of participants (collaborators) of the Han100K Initiative can be found online via http://www.pgghan.org/HCGD/about.

Keywords: mobile element insertion, MEI, transposable element, whole genome sequencing, variant

Abstract

Mobile element insertions (MEIs) are a major class of structural variants (SVs) and have been linked to many human genetic disorders, including hemophilia, neurofibromatosis, and various cancers. However, human MEI resources from large-scale genome sequencing are still lacking compared to those for SNPs and SVs. Here, we report a comprehensive map of 36,699 non-reference MEIs constructed from 5,675 genomes, comprising 2,998 Chinese samples (~26.2X, NyuWa) and 2,677 samples from the 1000 Genomes Project (~7.4X, 1KGP). We discovered that LINE-1 insertions were highly enriched at centromere regions, implying the role of chromosome context in retroelement insertion. After functional annotation, we estimated that MEIs are responsible for about 9.3% of all protein-truncating events per genome. Finally, we built a companion database named HMEID for public use. This resource represents the latest and largest genomewide study on MEIs and will have broad utility for exploration of human MEI findings.

Introduction

Transposable elements (TEs), also known as transposons or mobile elements, comprise a significant portion in mammalian genomes (Smit 1999; Deininger et al. 2003; Cordaux and Batzer 2009), approximately half of the human genome (Lander et al. 2001). Most TEs are transposition incompetent due to accumulated interior mutations and truncation or various host repression mechanisms (Goodier 2016). In humans, *Alu*, long interspersed nuclear element 1 (L1), SINE-VNTR-*Alu* (SVA), and HERV-K (also known as HML-2) are four families of TEs which are still active and capable of creating new insertions (Mills et al. 2007; Huang et al.

2012), termed mobile element insertions (MEIs). The transposition events have the potential to disrupt normal gene function and alter transcript expression or splicing at the sites of integration, contributing to disease (Payer and Burns 2019). For example, over 120 TE-mediated insertions have been associated with various human genetic diseases, including hemophilia, Dent disease, neurofibromatosis and various cancers (Hancks and Kazazian 2016). Apart from the impact through insertion events, intrinsic sequence properties of TEs endow some MEIs with functional effects on the host (Payer and Burns 2019), making MEIs differ qualitatively from typical forms of SVs like copy number variants (CNVs). Another important question related to MEIs is the integration site preference, which are usually non-random and influenced by various factors such as DNA sequences and chromatin context (Sultana et al. 2017).

However, despite these important functions, integrated resources for polymorphic TEs in human genomes is still lacking (Goerner-Potvin and Bourque 2018), which could offer a large pool of MEIs to explore TE diversity and serve as bedrock for phenotype-variant association studies. And MEIs are not routinely analyzed in most population-scale whole-genome sequencing (WGS) projects (The 1000 Genomes Project Consortium 2015; Wu et al. 2019, 2019; Cao et al. 2020). To date, the largest and most recent population study of MEIs using WGS remains the one conducted by the 1KGP, which included 2,504 genomes across 26 human populations (Sudmant et al. 2015; Gardner et al. 2017). However, the sequencing depth of the 1KGP is low, which may limit the MEI detection sensitivity and accuracy (Rishishwar et al. 2016). In addition, current MEI genetic resources are mainly from European ancestry cohorts, and the lack of Chinese cohort genomic study on MEIs is a critical part of the missing diversity.

In this study, we employed WGS of 5,675 members from newly sequenced Chinese samples and the 1KGP to construct a resource for non-reference MEIs. Although the 1KGP dataset has already been investigated for MEIs (Sudmant et al. 2015; Gardner et al. 2017), we included it here to increase population diversity and build a comprehensive MEI map. The NyuWa dataset has been used to study spectrum of small variant and build reference panel (Zhang et al. 2020), and the MEIs were not explored yet. Combining two cohorts enabled us to systematically analyze the genomic distribution, mutational patterns, and functional impacts of MEIs. From these analyses, we found that L1 MEIs were highly enriched in centromere regions, and we determined that MEIs represent about 9.3% of all protein-truncating events per individual, emphasizing the importance of detecting MEI routinely in WGS studies. We have built a companion database named HMEID (available at http://bigdata.ibp.ac.cn/HMEID/) for polymorphic MEIs, which could be explored for new insights into MEI biology.

Results

A Comprehensive Map of Non-reference Human MEIs

To generate a comprehensive map of MEIs from human genomes, we jointly analyzed two WGS datasets using MELT (Gardner et al. 2017), the low-coverage 1KGP dataset consisting of 2,677 individuals sequenced to ~7.4X coverage (Sudmant et al. 2015) and the high-coverage NyuWa dataset including 2,998 Chinese samples sequenced to ~26.2X coverage (Table S1) (Zhang et al. 2020). After site quality filtering, a total of 36,699 non-reference MEIs were kept, including 26,553 *Alu*s, 7,353 L1s, 2,667 SVAs and 126 HERV-Ks (Table 1). Most *Alu* and L1 MEIs were well-supported by split reads (Fig. S1A) and target site duplications (TSDs) (Fig. S1B). Using Hardy-Weinberg equilibrium (HWE) metrics as a rough proxy of genotyping accuracy, we found that about 87% autosomal MEI sites did not violate the HWE, and when restricted to the NyuWa dataset, almost all MEIs (97%) on autosomes had high genotyping accuracy (Fig. S2).

Table 1. MEI discovery in this study.

	Total sites	Mean sites per donor		Standard deviation	
		NyuWa	1KGP	NyuWa	1KGP
Alu	26,553	1,035	884	25.3	153
LINE-1	7,353	145	119	8.35	19.3
SVA	2,667	44.4	28.8	4.83	9.9
HERVK	126	11	8.23	1.86	2.12
Total	36,699	1,236	1,040	30	178

On average, we detected 1,236 MEIs with each genome in the NyuWa dataset and 1,040 MEIs in the 1KGP dataset (Table 1), which were expected as increased sequencing depth provides more power for MEI detection (Fig. S1C). The smaller correlation between MEI number and sequencing coverage in the NyuWa dataset than that of the 1KGP dataset reflected that MEI detection sensitivity was close to saturation in ~30X genomic coverage, consistent with the previous evaluation by the authors of MELT (Gardner et al. 2017). The distribution of MEI numbers per individual, MEI allele frequencies and length estimates largely fit the findings of previous studies (Fig. 1) (Gardner et al. 2017, 2019). About 70.7% MEIs are very rare (allele frequency < 0.1%), with over 30% singletons of all four MEI types (Fig. 1C; Fig. S1D). Since a large proportion of MEIs were individual-specific, we next sought to evaluate MEI discovery by increasing sample size. Through randomly down-sampling to different sizes with 100-

sample intervals, we estimated the total MEI variants and the increase of variants at different sample sizes (Fig. S1E-I). As expected, we found that the number of all four MEI types continued to rise with the increasing sample size, but the growth rate decreased. When looking at the subfamilies of MEIs, we found that the distributions of active *Alu* and L1 MEIs were in line with previous observations in humans (Gardner et al. 2017; Bennett et al. 2008; Stewart et al. 2011; Hormozdiari et al. 2013), e.g. *Alu*Ya5 and *Alu*Yb8 were found to be the most abundant two *Alu* subfamilies (Fig. S3), indicating their high retrotransposition activity in modern humans.

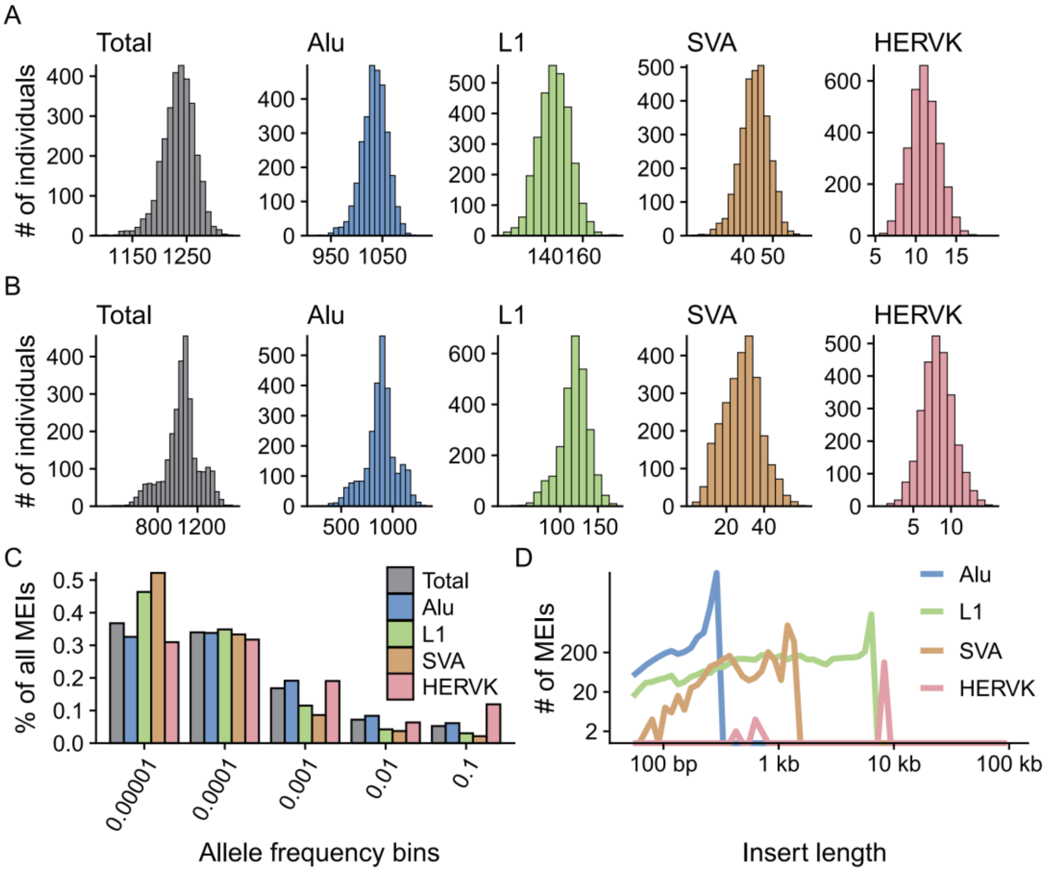

Fig. 1. The MEI call set. (**A**) Histograms of the number of MEIs identified per genome in the NyuWa dataset. (**B**) Histograms of the number of MEIs identified per genome in the 1KGP dataset. (**C**) Distribution of allele frequency of MEIs of four types: *Alu*, L1, SVA, and HERVK. "Total" combined the four types of MEIs. (**D**) Distribution of insert size estimated by MELT.

Compared to the previous MEI findings of 1KGP samples (Gardner et al. 2017), the total number of non-reference MEIs we detected has increased 55.4%, with 45.2% and 74.0% increase for *Alu* and L1 insertions respectively (Fig. S4A). In addition, large proportions of MEI calls detected by previous study were repeatedly identified in this study, and the allele frequency for overlapping sites also showed high consistency (Fig. S4B; Pearson's correlation coefficient = 0.95). Nonetheless, we noticed that many MEIs identified by Gardner *et al.* (Gardner et al. 2017) were missed in our call set. We conjectured that this may be due to differences of software version, reference genome build, and the way how the BAM files were generated etc. To test this, we performed three runs using three sample sets: 1) 100 samples from the 1KGP with reads mapping to the GRCh37 genome build; 2) 100 samples from the 1KGP with reads mapping to the GRCh38 genome build; 3) 100 samples from the 1KGP and 100 samples from the NyuWa, with reads mapping to the GRCh38 genome build. We found that more MEIs could be detected by using the GRCh38 genome build and/or by combining more samples (Table S2). This is also in line with the model used by MELT (Gardner et al. 2017), combining the 1KGP dataset with the high-coverage NyuWa dataset would improve MEI detection sensitivity as well as accuracy, with finer resolution of MEI break points. Collectively, our MEI call set represents a high-quality map of non-reference MEIs for humans.

Enrichment of Non-reference L1 insertions in Centromeres

It has been long noted that L1s occur preferentially in AT-rich regions but *Alu*s show the opposite trend (Lander et al. 2001). As expected, we also observed this tendency for MEIs (Fig.

S5A). In addition, the GC content of flanking DNA for *Alu*s and L1s were lower than background, while SVAs and HERV-Ks prefer DNA sequences with much higher GC content. We next compared the GC composition of rare MEIs (allele frequency < 1%) and common MEIs (allele frequency >= 1%) due to the reported bias shift in GC bias for older and younger short interspersed nuclear elements (SINEs) (Smit 1999; Hormozdiari et al. 2013; Medstrand et al. 2002; Waterson et al. 2005). Significant difference was only observed for HERV-K: rare HERV-K insertions occurred in much higher density at GC-rich regions (Fig. S5B). We did not observe marked bias for *Alu*s and SVAs, likely because most insertions we identified were already fixed in population.

We next sought to investigate the distribution of MEIs throughout the genome, like previously Collins *et al.* had done for common SVs (Collins et al. 2020). Interestingly, L1s were predominantly enriched at centromeric regions, whereas SVAs and HERV-Ks were enriched at telomeres (Fig. 2 A and B; Fig. S6). For comparison, similar analysis was applied to TEs in the reference genome, but no such patterns for L1s were found (Fig. S7B). Even in the latest telomere-to-telomere assembly of the human X chromosome, only a single L1 insertion was detected at the centromere region (Miga et al. 2020). When restricted to singleton L1 MEIs, we could still detect the enrichment in centromeres (Fig. 2C). Importantly, this finding was well-supported by non-reference L1s from euL1db (Fig. 2D) (Mir et al. 2015), which curated human polymorphic L1s from 32 different studies. Considering the reduced detection power of short-read WGS in repetitive regions, the enrichment of L1 insertions at centromeric regions could be still underestimated. The enrichment of non-reference L1 insertions at centromeric DNA could be partly attributed to lower GC content, as centromeres contain massive AT-rich alpha satellites (Manuelidis and Wu 1978). Also, active TEs have been found in neocentromere regions, and may contribute to centromere ontogenesis (Klein and O'Neill 2018; Contreras-Galindo et al. 2013; Zahn et al. 2015). The reasons for the dramatic enrichment of L1s in centromere regions are intriguing and further studies are needed in the future.

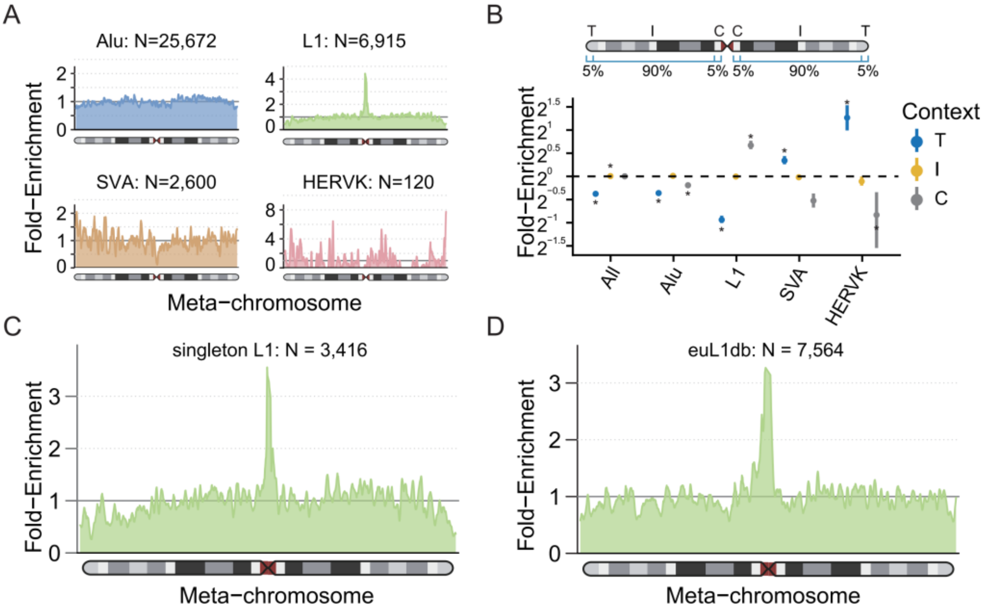

Fig. 2. Chromosome-level Distribution of MEI Density. (**A**) Smoothed enrichment of different types of MEIs ascertained in this study. The values were calculated per 100 kb window across the average of all autosomes and normalized by the length of chromosome arms (as "meta-chromosome"). (**B**) Enrichment of MEIs by class and chromosomal context. The dots are the mean values and point ranges represent 95% confidence intervals (CIs). P-values were computed using a two-sided t-test and adjusted using the Bonferroni method. *, p ≤ 0.05. C, centromeric; I, interstitial; T, telomeric. The way to compute the chromosomal enrichment and to represent data was from the gnomAD SV paper (Collins et al. 2020). (**C**) Smoothed enrichment of singleton L1s (L1 MEIs found in single genome) ascertained in this study. (**D**) Smoothed enrichment of non-reference L1s from euL1db database (Mir et al. 2015).

Strong Correlations between MEI Diversity and SNP Heterozygosity

Since mutations are ultimate sources of genetic innovation and significant causes of human birth defects and diseases, knowledge of mutation rate is a general population genetics question (Kumar and Subramanian 2002; Feusier et al. 2019). Here we employed the commonly-used Waterson's estimator (Watterson 1975) of Θ to estimate the mutation rate of each MEI type and found that mutation rates varied markedly by MEI class (Table S3). Since MEI detection and genotyping power is profoundly influenced by sample coverage (Gardner et al. 2017), we conducted the analysis separately for the NyuWa and the 1KGP datasets. The resulting calculation provided very close estimates of between 3.217×10^{-11} (NyuWa) and 2.928×10^{-11} (1KGP) de novo MEIs per bp per generation (μ), or roughly one new MEIs genome-wide every 11-16 live births, which is largely concordant with prior reports (Sudmant et al. 2015; Gardner et al. 2019).

The availability of SNP genotyping (both the NyuWa and the 1KGP dataset) for the same samples given us an opportunity to investigate the correlation between MEI diversity and SNP heterozygosity for each population. SNP heterozygosity was computed as the ratio of heterozygous SNPs across the individual's genome (Prado-Martinez et al. 2013) and was compared to the average MEI differences between samples in a given population (Hedges et al. 2004). The diversity for all types of MEIs showed strong correlation with SNP heterozygosity (R^2: 0.64~0.95), with African populations showing the highest MEI diversity and SNP heterozygosity (Fig. 3) — consistent with previous study (Stewart et al. 2011).

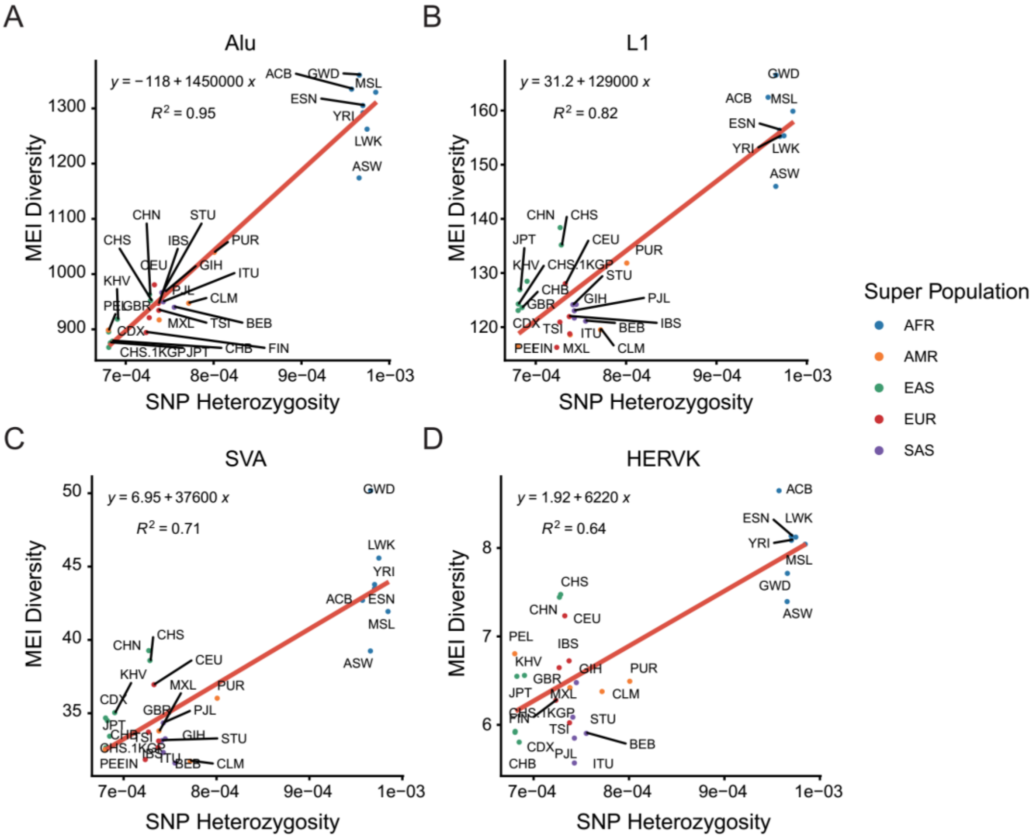

Fig. 3. **Correlation between SNP heterozygosity and MEI diversity**. SNP heterozygosities and diversity of (**A**) *Alu* MEIs, (**B**) L1 MEIs, (**C**) SVA MEIs and (**D**) HERV-K MEIs were compared in different populations. SNP heterozygosity was computed as the ratio of heterozygous SNPs across the individual's genome and MEI diversity was computed as the average allele difference in each population. Points were colored by super populations. AFR, African super population; AMR, American super population; EAS, East Asian super population; EUR, European super population; SAS, South Asian super population.

MEI Functional Properties

Via the local impacts by transposition events or more global post-insertion influence (Klein and O'Neill 2018), MEIs can disrupt normal gene functions and be disease-causing (Payer and Burns 2019; Hancks and Kazazian 2016). In principle, any MEIs can result in predicted loss-of-function (pLoF) by altering open-reading frames. To assess the functional impacts of MEIs, we annotated the MEI calls using Variant Effect Predictor (VEP) and BEDtools (see Methods). The vast majority (82.7%) of detected MEIs was in intergenic and intronic regions, while only ~2.7% MEIs impacted the coding sequences (CDS) (Fig. 4A). Varying enrichment levels on different genomic features were observed for different MEI types (Fig. 4B). For example, L1, SVA and HERV-K MEIs were significantly depleted in CDS and non-coding gene exons; L1 MEIs were enriched in coding introns and gene flanking regions; SVA and HERV-K sites were enriched in intergenic and non-coding introns. Focusing on protein-truncating variants (PTVs), each genome contained a mean of 24.8 MEIs (12.6 *Alu*, 7.4 L1, 1.3 SVA and 2.4 HERV-K) directly disrupting CDS, including 1.1 rare pLoF MEIs (allele frequency < 1%) (Fig. 4C; Table S4). By comparison, Karczewski *et al.* estimated 98.9 pLoF short variants (SNVs and InDels) per genome (Karczewski et al. 2020), and Collins *et al.* observed 144.3 pLOF SVs per genome (Collins et al. 2020). We thus estimated that MEIs account for about 9.3% (24.8/268) of all PTVs, among small variants and large SVs in each human genome.

Examining the degree to which evolutionary forces acting on coding MEI loci is important to understand the relationships between MEI variation and coding genes. Here we used three different metrics to investigate selective constraints: 1) the proportion of singleton variants (variants observed in only one individual), an established proxy for selection strengths (Lek et al. 2016); 2) the proportion of MEIs in genes with high probability of loss-of-function intolerance (pLI) (Lek et al. 2016); 3) the loss-of-function observed/expected upper bound fraction (LOEUF) of MEI-containing coding genes, where higher LOEUF scores suggest a relatively higher tolerance to inactivation for a given gene (Karczewski et al. 2020). HERV-K

MEI was not included in this analysis due to the relatively small number found in coding genes. Higher singleton proportions for *Alu* and L1 MEIs were found in CDS than that of introns (Fig. 4D; $\chi 2$ $p < 0.05$), while we did not find a statistically significant bias for SVA MEIs, though there were 166 and 949 SVA insertions found in CDS and coding introns, respectively. Likewise, lower proportions of *Alu*/L1 MEIs detected in genes with high pLI score (> 0.9) were found in CDS than that of intronic regions (Fig. 4E; $\chi 2$ $p < 0.05$). Observations from the perspective of enclosing genes fit these results: higher LOEUF score were found for genes with *Alu*/L1 MEIs (Fig. 4F, Wilcoxon $p < 0.05$). Our results sustained and expanded previous findings on human exome data (Gardner et al. 2019), in which Gardner *et al.* reported that exonic MEIs were under purifying selection.

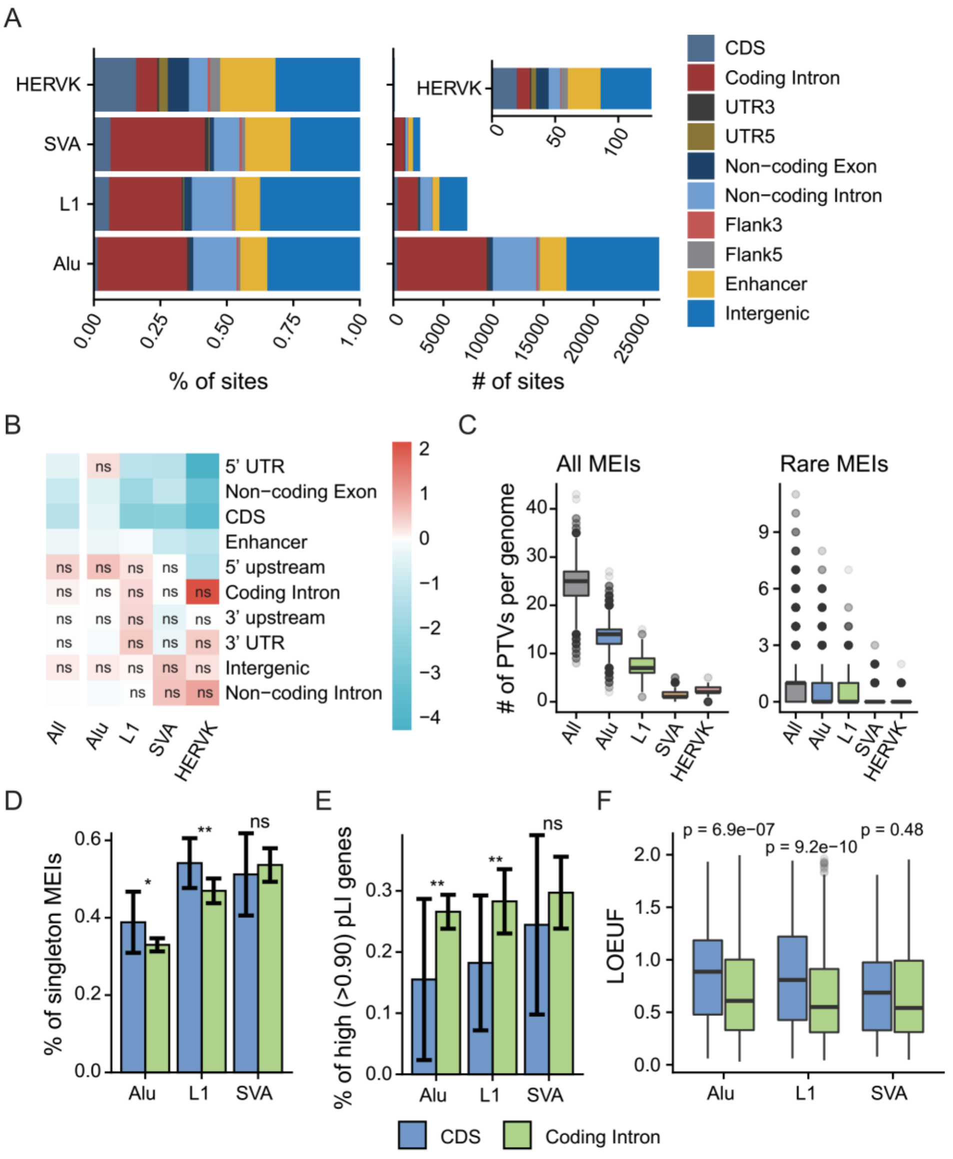

Fig. 4. MEI functional properties. (**A**) Predicted functional consequences for each type of MEI: (left) cumulative proportion, and (right) cumulative number. (**B**) Log2 fold enrichment of the MEI call set compared against the MEIs permutated. The permutation test was repeated 1000 times, and empirical p-values were commutated together with the enrichment values. The enrichment values were scaled row-wise. ns, not significant (p-value > 0.05). (**C**) Box plots of counts of predicted PTVs by MEI: (left) all the MEIs identified in this study, and (right) rare MEIs (allele frequency < 1%) in this study. (**D**) Proportions of singleton MEIs in CDS and coding introns for *Alu*, L1 and SVA. Error bars indicate 95% CIs based on population proportion. P-values were computed using chi-squared test. (**E**) Proportions of

high pLI genes (pLI > 0.9) for genes with MEIs in the CDS and genes with MEIs intron regions. Error bars represent 95% CIs based on population proportion. P-values were computed using chi-squared test. (**F**) Box plots of LOEUF scores of genes with MEIs in the CDS and genes with MEIs in their introns. Wilcoxon rank sum test was used to compute p-values. Figure D-F used the same legend beneath. ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$.

Although researchers have long noted that most of reference LTR elements and L1s in gene introns are in the antisense orientation with respect to the host genes (Smit 1999; Medstrand et al. 2002), possibly due to ill effects on transcript processing of sense-oriented elements (van de Lagemaat et al. 2006; Zhang et al. 2011), there are no established conclusions about the orientation tendency of non-reference MEIs (Gardner et al. 2019; Hormozdiari et al. 2013). Our large collection of MEIs found in genes allowed us to closely examine the strand bias of different MEIs. Although a bias for *Alu*, L1 MEIs and SVA MEIs to be in an antisense orientation when found within genes was observed (Hormozdiari et al. 2013), we did not find a statistically significant bias for L1 insertions (Fig. S8A). Conversely, *Alu*s were found to have strong strand bias when being inserted into protein-coding genes, non-coding genes, protein-coding introns, and non-coding introns (Fig. S8; $\chi2$ $p < 0.05$). For SVA MEIs, protein-coding genes, protein-coding exons, and protein-coding introns were regions where insertion orientation biases were detected (Fig. S8; $\chi2$ $p < 0.05$). Considering that *Alu* and SVA elements are non-autonomous TEs that are trans-mobilized by the L1 retrotransposition machinery (Dewannieux et al. 2003; Raiz et al. 2012), there may be some post-insertion selection forces on *Alu*/SVA elements which influence these patterns (Sultana et al. 2017). The genes themselves which had MEIs in sense or antisense strand in introns did not show clear differences in terms of selective constraints, by comparing the LOEUF scores of these two kinds genes (Fig. S8F). In addition, no significant orientation tendency against the neighboring genes were detected when MEIs were in gene upstream regions (Fig. S8I).

Alu MEIs have been found to be enriched in regions of genome associated with human disease risk, suggesting their potential effects on common diseases (Payer and Burns 2019; Payer et al. 2017). To identify MEIs potentially associated with human trait or disease, we mapped MEIs to regions in linkage disequilibrium (LD) with trait- or disease-associated loci identified by genome-wide association study (GWAS) ($P < 10^{-8}$) (Buniello et al. 2019). We found that 6,457 (about 17.6%) of the MEIs (17.5% for *Alu*, 15.3% for L1, 24.4% for SVA, and 16.6% for HERV-K) were in these regions that tagged by at least by one GWAS SNP (Table S5), with allele frequency of 738 MEIs over 1%, suggesting the remarkable potential for MEIs to contribute in disease and the utility of our MEI set in future phenotype-variant association studies.

L1 3’ Transduction and 5’ Inversion

Some L1 elements can bring a 3’ readthrough transcript to the offspring insert site, which is called 3’ transduction (Goodier et al. 2000). These L1 elements are usually near a strong Poly(A) sequence. Transcription of these L1 elements is not terminated by the original weak Poly(A) of the L1 element but by the stronger poly(A) sequence downstream. With the flanking sequences downstream L1 elements, we extracted the correspondence between L1s in different genomic positions. Totally, 446 offspring MEIs derived from 57 source MEIs were identified in our samples. These MEI relationships are both interchromosomal and intrachromosomal (Fig. 5A). Compared with L1 transduction source sites identified by 1KGP study (Gardner et al. 2017), we found most of the sites were overlapped (Fig. 5B). Among these sites, 2 of the 3 most active source sites (chr6:13190802, chr1:118858380) were also found in this study, while the site *L1RE3* (chr2: 155671336) is in a low complexity region and was filtered in the site filtering. Most of the sources transducts less than 20 offspring whereas site chrX:11713279 has 186 offspring (41% of all offspring detected). Source and offspring MEIs were distributed into families and population frequency was calculated (Fig. 5C and D). Most transduction classes were EAS specific. Comparing frequencies among subpopulations of EAS, we noticed 14 transduction classes only detected in Chinese people. Inside these classes, 5 classes only appear in samples of Northern Han Chinese (CHB, CHN) and 4 classes only appear in Southern Han Chinese (CHS, CHS.1KGP) (Table S6).

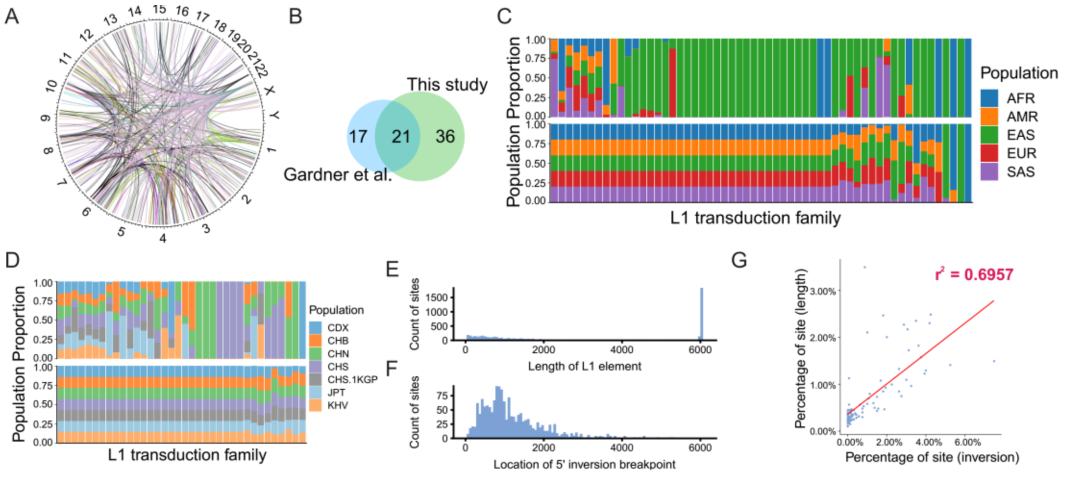

Fig. 5. L1 3' Transduction and 5' Inversion. (**A**) 3' transduction source-offspring relations across the whole genome. (**B**) Venn plot of 3' transduction sources found by our study and the 1KGP study (Gardner et al. 2017). (**C**) Source (bottom) and offspring (top) element frequencies in super populations. AFR, African super population; AMR, American super population; EAS, East Asian super population; EUR,

European super population; SAS, South Asian super population. (**D**) Source (bottom) and offspring (top) element frequencies in Asian subpopulations. CDX, Chinese Dai in Xishuangbanna; CHB, Han Chinese in Beijing; CHN, Northern Han Chinese, China; CHS, Southern Han Chinese; CHS.1KGP, Southern Han Chinese from the 1KGP; JPT, Japanese in Tokyo; KHV, Kinh in Ho Chi Minh City. (**E**) L1 length distribution within our call set. The length was estimated by MELT. (**F**) 5' inversion position distribution among all inverted sites. (**G**) Correlation plot between the distributions shown in (E) and (F). Full length L1 element was excluded in this comparison.

5' end of the L1 sequence can be inverted during insertion (Ostertag and Kazazian 2001). We extracted the 5' inversion information from the MELT result, and 1,606 L1 insertions were detected with a 5' inversion end. The nearest distance from the 5' inversion site to the 3' end of the L1 insertion is 602 bp, which is consistent with the 1KGP study (590 bp) (Gardner et al. 2017). It seems that inversion does not occur in the first ~600 bp from the 3' end, which may indicate that the inversion process requires at least ~600 bp DNA sequence. In the previous study, the distribution of the 5' inversion positions highly correlated with the distribution of L1 MEI lengths. MEIs in our study also showed this trend (R^2 = 0.696; Fig. 5E-G). We next calculated the percentage of 5' end inverted MEIs within each 3' transduction offspring class. The inversion rate across different classes varied and did not correlate with the class size (Table S7). For the biggest class which derived from chrX:11713279, only 25.3% of the offspring had 5' inversion while a class which only includes 15 offspring had a 40% inversion rate.

A Database for Polymorphic MEIs

Currently, resources for polymorphic TE findings in human genomes are in high demand (Goerner-Potvin and Bourque 2018). There were two dedicated databases for polymorphic human MEIs: dbRIP (Wang et al. 2006) and euL1db (Mir et al. 2015). However, the former had not been updated since 2012 and the latter was only for human-specific L1 insertions. To fill this gap, we have designed a companion database named HMEID to archive MEIs identified in this study, and to comprehensively catalog the variants on allele frequencies in the NyuWa dataset and the 1KGP dataset. Besides, variant quality metrics and functional annotations are also presented. Compared to dbRIP, HMEID contained more MEIs; the number of L1 insertions in HMEID was comparable with that of euL1db (Fig. 6). Importantly, HMEID contained MEIs detected from samples of Han Chinese, which is the largest ethnic group in the world. We anticipated that this resource would facilitate the exploration of TE polymorphisms and benefit future researches on TEs as well as human genetics.

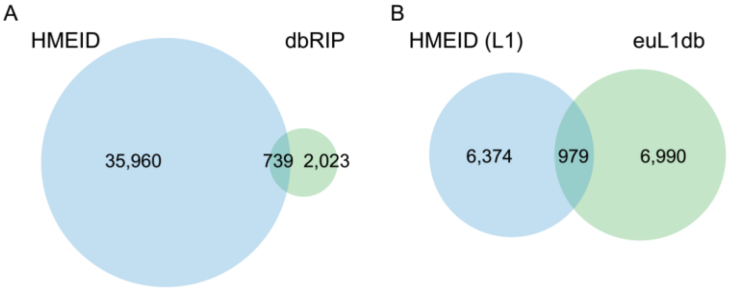

Fig. 6. Comparing HMEID with other MEI Databases. (**A**) Comparison the MEI set in the HMEID with that of from the dbRIP database (Wang et al. 2006). (**B**) Comparison the L1 MEIs in the HMEID with non-reference L1s from the euL1db database (Mir et al. 2015).

Discussion

MEIs, an endogenous and ongoing source of genetic variation, have not been investigated in many population-scale WGS projects. Here we leveraged 5,675 genomes from the NyuWa (Zhang et al. 2020) and the 1KGP (The 1000 Genomes Project Consortium 2015) dataset to identify non-reference MEIs. After describing the frequency spectrum of variants, we focused on the insertion site preference and functional impacts of MEIs. We provided an important resource of non-reference MEIs in humans.

We identified 36,699 non-reference MEIs for four types of TEs and determined that individuals harbour a mean of over 1,000 non-reference MEIs, mostly contributed by *Alu* insertions. In line with previous reports (Gardner et al. 2017, 2019; Stewart et al. 2011), most MEIs were rare and individual-specific, which was also observed for SNVs (The 1000 Genomes Project Consortium 2015) and SVs (Collins et al. 2020). With the newly sequenced 2,998 genomes from China, this study established a large-scale MEI resource for the genetics of Chinese as well as East Asians. Comparing to the previous study conducted by the 1KGP (Gardner et al. 2017), the number of MEIs detected by us has increased about 55%, representing what is to our knowledge the most comprehensive set of human non-reference MEIs.

We found that non-reference MEIs have non-random distributions along chromosomes, implicating the role of chromosome context in TE insertion. Of note, we found that non-reference L1 MEIs were drastically enriched in centromere regions, which was also supported by independent data from the euL1db (Mir et al. 2015). The genomic distribution of TEs is a result from insertion site preference and post-insertion selection on the host (Sultana et al. 2017). On the one hand, human centromeres are full of AT-rich alpha satellites (Manuelidis and Wu

1978), which could confer insertion preference for L1s, since the target specificity of L1 insertion machinery is TTTT/A (Feng et al. 1996). Certain centromeric histones and other centromeric proteins may also serve as preferred targets for TEs, as suggested by a study in maize (Schneider et al. 2016). Additionally, studies on HIV integration into the host genome implied that proximity to the nuclear periphery of centromere may facilitate TE targeting (Lelek et al. 2015; Marini et al. 2015). On the other hand, incorporation of L1s may facilitate the recurring evolutionary novelty of centromeres (Klein and O'Neill 2018). In support of this, Chueh *et al.* reported that RNA transcripts from a full-length L1 are the essential structural and functional components in the regulation of a human neocentromere (Chueh et al. 2009). Evidences were also found in the tammar wallaby (*Macropus eugenii*), where dramatic enrichment of L1s and endogenous retroviruses was found in a latent centromere site (Longo et al. 2009), and *Equus caballus,* where evolutionarily new centromeres locate in LINE- and AT-rich regions (Nergadze et al. 2018). In addition to centromere ontogenesis, a LINE-like element (G2/jockey3) contributes directly to the organization and function of centromeres of *D. melanogaster* (Chang et al. 2019). This is also likely true for the non-reference SVA, for which we found an enrichment in telomeres, as TEs were found to be essential in maintaining the telomere length homeostasis in insects (Pardue and DeBaryshe 2011). However, another plausible explanation for both the enrichment of non-reference L1 MEIs in centromere and non-reference SVA MEIs in telomere is that these regions contain few protein-coding genes, limiting insertional mutagenesis by TEs (Sultana et al. 2017). The reasons for this phenomenon are fascinating, and our study post an important question about the relationship between TEs and centromeres.

Knowing the functional impact of MEIs is fundamental to our understanding the impact of MEI with respect to human disease or trait and evolution (Goerner-Potvin and Bourque 2018). We have estimated that MEIs accounted for about 9.3% of all protein-truncating variants per genome, among small variants (Karczewski et al. 2020) and SVs (Collins et al. 2020). Our estimation was much higher than that determined by whole exome sequencing data (Gardner et al. 2019), possibly due to the limitation of exome baits. We found that a significant portion of polymorphic MEIs mapping to loci implicated in trait/disease association by GWAS, as increasingly recognized by recent studies (Payer et al. 2017; Wang et al. 2017). While previous GWAS have mainly focused on small variants (Visscher et al. 2017), future association studies should consider and evaluate the effects of MEIs in common disease. We anticipate that the HMEID will serve as a basis for such studies.

Our study is limited in that only one tool was used to identify MEIs. Though the overall performance of MELT outperformed existing MEI discovery tools (Gardner et al. 2017) and it has been successfully used in several large-scale studies (Gardner et al. 2017, 2019; Feusier et al. 2019; Werling et al. 2018; Torene et al. 2020), but the detection power could be compromised by modest sequencing depth and incompetence in complex genomic regions of short-read WGS *etc*. In addition, the overall genotyping accuracy by MELT v2 was 87.95% for non-reference *Alu*s (not excluding MEIs in low complexity regions), when compared with PCR generated genotypes (Goubert et al. 2020). As such, we have tried to ensure the site quality by strict filtering. In the future, we would consider combining different MEI identification and genotyping tools to improve the quality, which has been proved useful in previous reports (Ewing 2015; Goerner-Potvin and Bourque 2018; Rishishwar et al. 2016; Feusier et al. 2019).

Also, long-read WGS is promising in detecting MEIs, especially for genomic regions refractory to approaches using short-read sequencing technologies (Audano et al. 2019; Chaisson et al. 2019; Zhou et al. 2020). Another limitation of our MEI dataset is that reference MEIs (MEIs detected as deletions) were not included yet, for which the detection is underway and the results would be integrated into the HMEID for public use.

Methods

Experimental design

Data in this study were from two sources: low-coverage (~7.4X) WGS samples from the 1KGP (The 1000 Genomes Project Consortium 2015) and high-coverage (~26.2X) WGS samples from the NyuWa dataset (Zhang et al. 2020). For the 1KGP dataset, CRAM-format files of 2,691 individuals were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/, which were aligned to human genome build GRCh38 (Lowy-Gallego et al. 2018). The CRAM files were then converted to BAMs using SAMtools v1.9 (Li et al. 2009). The NyuWa dataset contained 2,999 individuals including diabetes and control samples collected from different provinces in China (Zhang et al. 2020), and this cohort was sequenced using the Illumina platform. The processing from raw FASTQs to BAMs was according to the GATK Best Practices Workflows germline short variant discovery pipeline (Poplin et al. 2018), as described in (Zhang et al. 2020). The median depth of the NyuWa samples after genome alignment (GRCh38 human genome build) and removal of PCR duplicates was about 26.2X.

Generation of MEI call set

MELT v2.15 (Gardner et al. 2017) was run with default parameters using “SPLIT” mode to identify non-reference MEIs, which detects a wide range of non-reference *Alu*, L1, SVA and HERV-K insertions. To get the BAM coverage for MELT analysis, we used goleft v0.1.8 (https://github.com/brentp/goleft) “covstats” function to estimate the genomic coverage for each sample. After initial generation of a unified VCF file by MELT “MakeVCF” function, variants that did not pass the following criteria were filtered to get a high-quality MEI call set: 1) not in low complexity regions; 2) be genotyped in greater than 25.0% of individuals; 3) split reads > 2; 4) MELT ASSESS score > 3; and 5) VCF FILTER column be PASS. 2,998 of 2,999 samples in NyuWa and 2,677 of 2,691 samples in 1KGP were successfully analyzed, with the final call set consisting of 36,699 MEIs from 5,675 genomes. Subfamily characterization for *Alu* MEIs and L1 MEIs was done using MELT’s CALU tool.

Detection of L1 3’ transduction and 5’ inversion

Following the generation of a high-quality MEI call set, MELT v2.15 was used to detect L1 3’ transduction. We followed the instruction of MELT 3’ transduction identification pipeline and extracted the METRANS and MESOURCE field in the resulting VCF manually. The population frequency was calculated with the AC/AN (for offspring MEI set, we used the sum of AC and AN) and normalized across different populations.

The MELT VCF provided the position of a 5' inversion site (from the 3’ end) through the “ISTP” field. We subtracted it from the full length of L1 (6,019 bp) to obtain the coordinate of the inversion site from the 5’ end. While comparing the inversion coordinate and the length of

L1, we removed the full-length L1 elements from the comparison set. Sites were distributed into 100 bins across the full length of L1. We compared the distribution of sequence length and inversion site position among these bins and calculated the Pearson correlation value.

Analysis of Hardy-Weinberg equilibrium

To evaluate the genotype distributions of each MEI under the null expectations set by the Hardy-Weinberg equilibrium (HWE), we tabulated genotype distributions of autosomal MEIs per dataset and performed exact tests by "HWExactStats" function in R package HardyWeinberg v1.6.3 (Graffelman 2015). While disequilibrium may indicate disease association or population stratification, it may be the result of confusion of heterozygotes and homozygotes. We thus used the HWE test for gross quality-check of genotyping accuracy (Fig. S2), as described in (Collins et al. 2020).

Comparison with the 1KGP MEI call set

To compare with the MEIs generated by the 1KGP (Gardner et al. 2017), we downloaded the GRCh38 version call set from the dbVar database (Lappalainen et al. 2013)**.** Then non-reference MEIs were extracted and compared with the MEIs identified in this study, using "window" function from BEDtools v2.26.0 (Quinlan and Hall 2010). When a site was located in ±500 bp of another site, it was considered as a hit.

Testing MELT for different genome build and joint calling

To test MELT's performance on different genome build, we randomly generated 100 samples from the 1KGP dataset, and we got the alignment files for both GRCh37 and GRCh38 version for these samples. After which we ran MEIL v2.15 on the two dataset and filtered sites as mentioned above. Finally, we compared the results using the function "intersect" from BEDtools v2.26.0 (Quinlan and Hall 2010).

To test MELT' performance with respect to sample size (joint calling), we randomly generated 100 samples from the NyuWa dataset and combined with the 100 random samples from the 1KGP above. Then we identified MEIs using the same pipeline as before on these 200 samples. After which we compared the call set with the MEIs detected from the 100 samples from the 1KGP with BEDtools "intersect".

Functional annotation

Variant Effect Predictor v99.2 (VEP) (McLaren et al. 2016) with Ensembl database version 99 (Zerbino et al. 2018) was used to annotate MEIs, with parameters "--pick --canonical --distance 1000,500". MEIs were also intersected with enhancers from GeneHancer database (Fishilevich et al. 2017) using BEDtools v2.26.0 "intersect" function (Quinlan and Hall 2010). Only one functional consequence was kept for each MEI, and enhancers were given higher priority when a MEI was also found in non-coding genes and intergenic regions.

Mapping MEIs to the GWAS signals was done as described in a previous study (Payer et al. 2017). GWAS SNPs and their related traits were obtained from GWAS Catalog v1.0.2 (Buniello et al. 2019). We first defined the LD block region for each GWAS SNP by its proxy SNPs ($r^2 > 0.8$). The LD between all the SNPs was calculated using the SNP call set generated by 1KGP phase III (The 1000 Genomes Project Consortium 2015), with plink v2.00a1LM (Chang et al. 2015). If there was no LD SNPs found in either side of the GWAS SNP, we used the median length of all predicted LD regions as the block length, centered by the target SNP. Then BEDtools v2.26.0 “intersect” function (Quinlan and Hall 2010) was employed to identify MEIs falling into these LD block regions. The complete set of these MEIs could be found in Table S5.

To qualify the enrichment of MEIs across different genomic features (Fig. 4B), we permuted 1,000 times for each MEI type with the same number as the real calls using GAT v1.3.4 (Heger et al. 2013). Each permutation set was annotated with VEP and BEDtools using the same rules as above. After counting the MEIs in each genomic feature, log2 fold changes and empirical p-values were computed. We repeated 3 times of the permutation procedure to verify the results.

Chromosome-level analyses of MEI density

To check the distribution of MEIs throughout the genome, we used the method described by Collins *et al.* (Collins et al. 2020) and we repeated it here for clarity. Focusing on 22 autosomes, each chromosome was segmented into consecutive 100kb bins and bins overlapped with centromeres were removed. For each MEI type (*Alu*, L1, SVA and HERV-K), the number of variants in each bin was recorded to get a matrix of MEI counts per 100kb bins per autosome. To smooth the MEI counts for each MEI type, an 11-bin (~1Mb) rolling mean per chromosome was computed. Each bin was then assigned to a percentile based on the position of that bin on its respective chromosome arm relative to the centromere. Specifically, a value of 0 corresponded to the centromere, and a value of -1 and 1 corresponded to the p-arm telomere and q-arm telomere, respectively. Finally, to compute “meta-chromosome” density shown in

Fig. 2, the normalized bin positions (i.e., -1 to 1) were cut into 500 uniform intervals, and values across all autosomes based on the normalized interval position were averaged. For the comparison of chromosome contexts (Fig. 2), normalized positions within the outermost 5% of each chromosome arm were considered as "telomeric", the innermost 5% as "centromeric" and the other 90% of each arm as "interstitial". Visualization of density of different MEIs on each chromosome shown in Fig. S6 was done using RIdeogram v0.2.2 (Hao et al. 2020).

Mutation rates

Before estimating mutation rate, we exclude the MEIs failed in the HWE test (adjusted p < 0.05). MEIs in low complexity regions (Li 2014) and in reference TE sequences were also filtered, due to the inability of MELT in these regions. Watterson's Theta (Watterson 1975) was then used to estimate the genome mutation rate of each MEI type:

$$\widehat{\theta_w} = \frac{K}{\sum_{i=1}^{n-1}\frac{1}{i}}$$

where K is the number of MEI site observed per MEI type in given population, and is the total number of chromosomes assessed. Then mutation rates were estimated as:

$$\widehat{\theta_{\mathrm{w}}} = 4N_{\mathrm{e}}$$

with an effective population size (i.e. N_e) of 10,000, consistent with previous studies (Sudmant et al. 2015; Gardner et al. 2019; Collins et al. 2020). To estimate mutation rates worldwide, the average mutation rate across all five continental populations was computed, with 95% confidence interval surrounding the mean based on t distribution (Collins et al. 2020).

SNP heterozygosity and MEI diversity

As described in a previous study (Hormozdiari et al. 2013), SNP heterozygosity was computed as the ratio of heterozygous SNPs over the length of the genome, and the mean value was used when multiple samples were considered. MEI diversity was defined as the average number of MEI differences between individuals in a population. For the NyuWa dataset (Zhang et al. 2020), high-quality SNP calls generated by the GATK v3.7 cohort pipeline (DePristo et al. 2011; Poplin et al. 2018) were used. For 1KGP3 samples, SNP calls on the human genome build GRCh38 of the were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/. Number of heterozygous SNPs was computed by VCFtools v0.1.15 (Danecek et al. 2011) and MEI diversity by “gtcheck” function in BCFtools v1.3.1 (Danecek and McCarthy 2017).

Database construction

We constructed the database with Bootstrap and Django. For each population, we calculated allele frequency of each MEI. All the data can be browsed in the database and downloaded from the “Download” page.

Statistical analysis

All statistical analyses in this study were briefly described in the main text and performed using R v3.6.2 (http://CRAN.R-project.org/).

Data Access

Complete MEI call set and other related information such as allele frequency and functional annotation are available in the companion database HMEID (available at http://bigdata.ibp.ac.cn/HMEID/).

Acknowledgments

We thank Eugene J. Gardner for helping us in using MELT. We thank Jing Wang for valuable comments in the data analysis and critical review of the manuscript. We thank Tingrui Song for assisting the use of high-performance computing platforms. We thank the people for generously contributing samples and sequencing data to the NyuWa dataset and the 1KGP dataset. Data analysis and computing resources were supported by the Center for Big Data Research in Health (http://bigdata.ibp.ac.cn), Institute of Biophysics, Chinese Academy of Sciences. This work was supported by the National Key R&D Program of China [2016YFC0901702, 2018YFA0106901]; National Natural Science Foundation of China [31871294, 31701117, 31970647]; the 13th Five-year Informatization Plan of Chinese Academy of Sciences Grant [XXH13505-05].

Author Contributions

T.X. and S.M.H. conceptualized and supervised the project. Y.W.N., X.Y.T., Y.R.S., Y.Y.L., Y.H.T. and Q.K. conducted data analysis. X.Y.T. built the database. Y.W.N., X.Y.T., H.X.L., P.Z. and S.M.H. drafted the manuscript, and all the primary authors reviewed, edited, and approved the manuscript.

Disclosure Declaration

The authors declare no competing interests.

References

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **0**. https://www.cell.com/cell/abstract/S0092-8674(18)31633-7 (Accessed January 21, 2019).

Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. 2008. Active Alu retrotransposons in the human genome. *Genome Res* **18**: 1875–1883.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**: D1005–D1012.

Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, et al. 2020. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res* 1–15.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* **10**: 1784.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**. https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533 (Accessed June 29, 2019).

Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen C-C, Erceg J, Beliveau BJ, Wu C-T, et al. 2019. Islands of retroelements are major components of Drosophila centromeres. *PLOS Biology* **17**: e3000241.

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA, Wong LH. 2009. LINE Retrotransposon RNA Is an Essential Structural and Functional Epigenetic Component of a Core Neocentromeric Chromatin. *PLOS Genetics* **5**: e1000354.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.

Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez MJ, Kappes

F, Dube D, Chan SM, Robinson D, Meng F, et al. 2013. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Res* **23**: 1505–1513.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* **10**: 691–703.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Danecek P, McCarthy SA. 2017. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**: 2037–2039.

Deininger PL, Moran JV, Batzer MA, Kazazian HH. 2003. Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development* **13**: 651–658.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* **35**: 41–48.

Ewing AD. 2015. Transposable element detection from whole genome sequence data. *Mob DNA* **6**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4696183/ (Accessed May 29, 2017).

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition. *Cell* **87**: 905–916.

Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**: 1567–1577.

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**. https://academic.oup.com/database/article/doi/10.1093/database/bax028/3737828 (Accessed November 27, 2018).

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Consortium 1000 Genomes Project, Devine SE. 2017. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res* gr.218032.116.

Gardner EJ, Prigmore E, Gallone G, Danecek P, Samocha KE, Handsaker J, Gerety SS, Ironfield H, Short PJ, Sifrim A, et al. 2019. Contribution of retrotransposition to developmental disorders. *Nat Commun* **10**: 1–10.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nature Reviews Genetics* **19**: 688–704.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mobile DNA* **7**: 16.

Goodier JL, Ostertag EM, Kazazian Jr HH. 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Human Molecular Genetics* **9**: 653–657.

Goubert C, Thomas J, Payer LM, Kidd JM, Feusier J, Watkins WS, Burns KH, Jorde LB, Feschotte C. 2020. TypeTE: a tool to genotype mobile element insertions from whole genome resequencing data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa074 (Accessed March 1, 2020).

Graffelman J. 2015. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *Journal of Statistical Software* **64**: 1–23.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4859970/ (Accessed June 20, 2019).

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci* **6**: e251.

Hedges DJ, Callinan PA, Cordaux R, Xing J, Barnes E, Batzer MA. 2004. Differential Alu Mobilization and Polymorphism Among the Human and Chimpanzee Lineages. *Genome Res* **14**: 1068–1075.

Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**: 2046–2048.

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, et al. 2013. Rates and patterns of great ape retrotransposition. *PNAS* **110**: 13457–13462.

Huang CRL, Burns KH, Boeke JD. 2012. Active Transposition in Genomes. *Annu Rev Genet* **46**: 651–675.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434–443.

Klein SJ, O'Neill RJ. 2018. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res* **26**: 5–23.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *PNAS* **99**: 803–808.

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

Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, et al. 2013. dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* **41**: D936–D941.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291.

Lelek M, Casartelli N, Pellin D, Rizzi E, Souque P, Severgnini M, Di Serio C, Fricke T, Diaz-Griffero F, Zimmer C, et al. 2015. Chromatin organization at the nuclear pore favours HIV replication. *Nature Communications* **6**: 6483.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843–2851.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Longo MS, Carone DM, Green ED, O'Neill MJ, O'Neill RJ, NISC Comparative Sequencing Program. 2009. Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. *BMC Genomics* **10**: 334.

Lowy-Gallego E, Fairley S, Zheng-Bradley H, Clarke L, Flicek P. 2018. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes. *F1000Research* **7**. https://f1000research.com/posters/7-1445 (Accessed May 18, 2020).

Manuelidis L, Wu JC. 1978. Homology between human and simian repeated DNA. *Nature* **276**: 92–94.

Marini B, Kertesz-Farkas A, Ali H, Lucic B, Lisek K, Manganaro L, Pongor S, Luzzati R, Recchia A, Mavilio F, et al. 2015. Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**: 227–231.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biology* **17**: 122.

Medstrand P, Lagemaat LN van de, Mager DL. 2002. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. *Genome Res* **12**: 1483–1495.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 1–9.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends in Genetics* **23**: 183–191.

Mir AA, Philippe C, Cristofari G. 2015. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res* **43**: D43–D47.

Nergadze SG, Piras FM, Gamba R, Corbo M, Cerutti F, McCarter JGW, Cappelletti E, Gozzo F, Harman RM, Antczak DF, et al. 2018. Birth, evolution, and transmission of satellite-free mammalian centromeric domains. *Genome Res* **28**: 789–799.

Ostertag EM, Kazazian HH. 2001. Twin Priming: A Proposed Mechanism for the Creation of Inversions in L1 Retrotransposition. *Genome Res* **11**: 2059–2065.

Pardue M-L, DeBaryshe PG. 2011. Retrotransposons that maintain chromosome ends. *PNAS* **108**: 20317–20324.

Payer LM, Burns KH. 2019. Transposable elements in human genetic disease. *Nat Rev Genet* **20**: 760–772.

Payer LM, Steranka JP, Yang WR, Kryatova M, Medabalimi S, Ardeljan D, Liu C, Boeke JD, Avramopoulos D, Burns KH. 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. *PNAS* **114**: E3984–E3992.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–475.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res* **40**: 1666–1683.

Rishishwar L, Mariño-Ramírez L, Jordan IK. 2016. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform*. https://academic.oup.com/bib/article/doi/10.1093/bib/bbw072/2562836 (Accessed October 31, 2017).

Schneider KL, Xie Z, Wolfgruber TK, Presting GG. 2016. Inbreeding drives maize centromere evolution. *PNAS* **113**: E987–E996.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development* **9**: 657–663.

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam

HYK, Lee W-P, et al. 2011. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genet* **7**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158055/ (Accessed March 10, 2020).

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics* **18**: 292–308.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Torene RI, Galens K, Liu S, Arvai K, Borroto C, Scuffins J, Zhang Z, Friedman B, Sroka H, Heeley J, et al. 2020. Mobile element insertion detection in 89,874 clinical exomes. *Genet Med* 1–5.

van de Lagemaat LN, Medstrand P, Mager DL. 2006. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol* **7**: R86.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**: 5–22.

Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. 2006. dbRIP: A Highly Integrated Database of Retrotransposon Insertion Polymorphisms in Humans. *Hum Mutat* **27**: 323–329.

Wang L, Norris ET, Jordan IK. 2017. Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes. *Front Microbiol* **8**. https://www.frontiersin.org/articles/10.3389/fmicb.2017.01418/full (Accessed August 20, 2020).

Waterson RH, Lander ES, Wilson RK, The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256–276.

Werling DM, Brand H, An J-Y, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**: 727–736.

Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, Soon WWJ, Bertin N, Lin CB, Khor CC, et al.

2019. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**: 736-749.e15.

Zahn J, Kaplan MH, Fischer S, Dai M, Meng F, Saha AK, Cervantes P, Chan SM, Dube D, Omenn GS, et al. 2015. Expansion of a novel endogenous retrovirus throughout the pericentromeres of modern humans. *Genome Biology* **16**: 74.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754–D761.

Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, Niu Y, Shi Y, Zhou H, Song T, et al. 2020. NyuWa Genome Resource: Deep Whole Genome Sequencing Based Chinese Population Variation Profile and Reference Panel. *bioRxiv* 2020.11.10.376574.

Zhang Y, Romanish MT, Mager DL. 2011. Distributions of Transposable Elements Reveal Hazardous Zones in Mammalian Introns. *PLOS Computational Biology* **7**: e1002046.

Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. 2020. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res* **48**: 1146–1163.

2019. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**: 106–111.