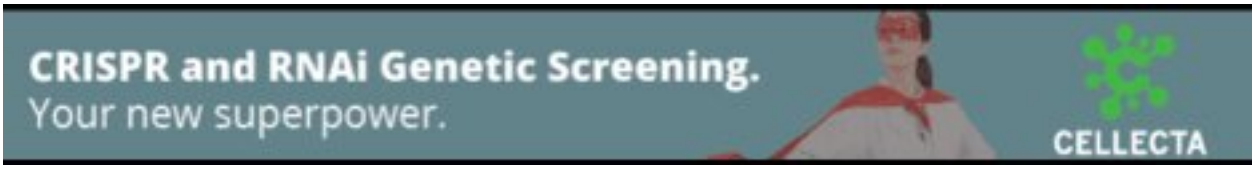
The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology

Eugene J. Gardner, Vincent K. Lam, Daniel N. Harris, et al.

Genome Res. published online August 30, 2017
Access the most recent version at doi:10.1101/gr.218032.116



The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology

Eugene J. Gardner[1,2], Vincent K. Lam[2,3,], Daniel N. Harris[1,2], Nelson T. Chuang[1,2,4,5], Emma C. Scott[1,2], W. Stephen Pittard[6], Ryan E. Mills[7,8], The 1000 Genomes Project Consortium, and Scott E. Devine[1,2,3,4,9]

1) Program in Molecular Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, U.S.A.;
2) Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, U.S.A.;
3) Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201, U.S.A.;
4) Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, U.S.A.;
5) Division of Gastroenterology, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, U.S.A.;
6) Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.;
7) Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, U.S.A.;
8) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, U.S.A.;
9) Corresponding author: sdevine@som.umaryland.edu.

Corresponding Author:
Scott E. Devine, Ph.D.
Institute for Genome Sciences
University of Maryland School of Medicine
801 W. Baltimore Street
Rm 615 BioPark II
Baltimore, MD 21201
Phone: (410) 706-2343
Email: sdevine@som.umaryland.edu

Running title: Mobile Element Locator Tool (MELT)
Keywords: MELT, mobile element insertion, *Alu*, LINE-1, SVA

August 5, 2017

Abstract

Mobile element insertions (MEIs) represent ~25% of all structural variants in human genomes. Moreover, when they disrupt genes, MEIs can influence human traits and diseases. Therefore, MEIs should be fully discovered along with other forms of genetic variation in whole genome sequencing (WGS) projects involving population genetics, human diseases, and clinical genomics. Here, we describe the Mobile Element Locator Tool (MELT), which was developed as part of the 1000 Genomes Project to perform MEI discovery on a population scale. Using both Illumina WGS data and simulations, we demonstrate that MELT outperforms existing MEI discovery tools in terms of speed, scalability, specificity and sensitivity, while also detecting a broader spectrum of MEI-associated features. Several run modes were developed to perform MEI discovery on local and cloud systems. In addition to using MELT to discover MEIs in modern humans as part of the 1000 Genomes Project, we also used it to discover MEIs in chimpanzees and ancient (Neanderthal and Denisovan) hominids. We detected diverse patterns of MEI stratification across these populations that likely were caused by: i) diverse rates of MEI production from source elements, ii) diverse patterns of MEI inheritance, and iii) the introgression of ancient MEIs into modern human genomes. Overall, our study provides the most comprehensive map of MEIs to date spanning chimpanzees, ancient hominids, and modern humans, and reveals new aspects of MEI biology in these lineages. We also demonstrate that MELT is a robust platform for MEI discovery and analysis in a variety of experimental settings.

Introduction

Population-scale, whole genome sequencing (WGS) projects have rapidly expanded over the past several years (The 1000 Genomes Project Consortium, 2010; The 1000 Genomes Project Consortium, 2012; Genome of the Netherlands Consortium, 2014; The 1000 Genomes Project Consortium, 2015; The UK10K Consortium, 2015; Gudbjartsson et al. 2015; Telenti et al. 2016). As we look to the future, projects are planned or underway to sequence many thousands of additional human genomes for studies involving population genetics, human diseases, and clinical genomics. Although new technologies such

as the Illumina HiSeq X platform can produce the massive amounts of WGS data that are required for such studies, many of the analysis tools that were developed over the past decade cannot be scaled up to meet the informatics demands of these data-intensive projects. Tools that detect mobile element insertions (MEIs) are no exception, and as a consequence, MEIs are not being routinely detected in most population-scale WGS projects (e.g., The UK10K Consortium, 2015; Gudbjartsson et al. 2015; Telenti et al. 2016). Thus, there is a clear need for innovative MEI discovery approaches that can address this important gap in variant detection.

MEIs should be fully discovered along with other forms of genetic variation because they can alter human traits or cause diseases when they disrupt genes. For example, at least five reported cases of hemophilia A have been linked to germline MEIs that disrupted the Factor XIII (*F8*) gene (Kazazian et al. 1988; Van de Water et al. 1998; Sukarova et al. 2001; Ganguly et al. 2003), and another six cases of hemophilia B have been linked to germline MEIs that disrupted the Factor IX (*F9*) gene (Vidaud et al. 1993; Wulff et al. 2000; Li et al. 2001; Mukherjee et al. 2004; Nakamura et al. 2015). 10/11 (90.1%) of these insertions disrupted coding exons, while the remaining insertion caused exon "skipping" (Ganguly et al. 2003). Similar germline MEIs have been implicated in a range of other human diseases, including neurofibromatosis (Wallace et al. 1991), Duchenne muscular dystrophy (Narita et al. 1993), cystic fibrosis (Chen et al. 2008), retinitis pigmentosis (Schwahn et al. 1998), beta-thalassemia (Divoky et al.1996; Kimberland et al. 1999; Lanikova et al. 2013), various cancers (Miki et al. 1996; Teugels et al. 2005), and other diseases (e.g., Janicic et al. 1995; Claverie-Martin et al. 2003; Watanabe et al. 2005). As above, most of these diseases were caused by MEIs that disrupted the coding exons of genes or caused exon skipping, although disease-causing MEIs also have been identified in the promoters (Lanikova et al. 2013) and untranslated regions (UTRs) of protein-coding genes (Watanabe et al. 2005). Thus, MEIs can influence human traits and diseases by disrupting a range of gene features.

MEIs also are mobilized in somatic human tissues, including epithelial cancers, suggesting that somatic MEIs might help to drive tumorigenesis (Miki et al. 1992; Iskow et al. 2010; Lee et al. 2012; Solyom et al. 2012; Shukla et al. 2013; Helman et al. 2014; Tubio et al. 2014; Rodic et al. 2015; Ewing et al. 2015; Doucet-O'Hare et al. 2015; Scott et al. 2016). Likewise, somatic MEIs are produced in at least

some normal somatic tissues such as the colon and brain (Muotri et al. 2005; Coufal et al. 2009; Baillie et al. 2011; Evrony et al. 2012; Upton et al. 2015; Scott et al., 2016), where they have been linked to colorectal cancer (Scott et al. 2016), schizophrenia (Bundo et al. 2014), and Aicardi-Goutieres syndrome (Upton et al. 2015). Thus, MEIs ideally would be discovered routinely in somatic WGS projects involving normal/tumor pairs and individual cells in order to gain a better understanding of cancers and other diseases.

Three major classes of mobile elements, i.e., *Alu*, L1, and SVA elements, remain actively mobile in human genomes and continue to generate new offspring MEIs (Mills et al. 2007; Iskow et al. 2010; Beck et al. 2010; Huang et al. 2010; Ewing and Kazazian 2010; Stewart et al. 2011; Sudmant et al. 2015; The 1000 Genomes Project Consortium 2015). All three of these element classes are mobilized by the L1 retrotransposition machinery, and as a consequence, all of these elements have at least some characteristic features of L1 elements. For example, the target site duplications (TSDs) that flank new *Alu*, L1, and SVA insertions are all very similar because they are created by the same L1-encoded proteins and target-primed-reverse-transcriptase (TPRT) mechanism (Luan et al. 1993; Moran et al. 1996; Dewannieux et al. 2003; Hancks et al. 2011; Raiz et al. 2012). Likewise, interior mutations frequently are introduced into *Alu*, L1, and SVA copies by the error-prone L1 reverse transcriptase that replicates all three of these element classes (Gilbert et al. 2005). Interior mutations have been useful for identifying and tracking active subfamilies of *Alu*, L1, and SVA elements (Boisonott et al. 2000; Batzer and Deininger 2002; Wang et al. 2005; Konkel et al. 2015), and for tracking relationships between source elements and their offspring (Scott et al. 2016). Several additional hallmark features of human MEIs include: i) 3'-transductions that are caused by alternative, downstream poly(A) signals (Moran et al. 1999), ii) 5' inversions that are caused by twin priming (Ostertag et al. 2001), and iii) 5' truncations that are caused by incomplete replication (Beck et al. 2011). Ideally, MEI discovery tools would fully detect all of these associated genetic features because such features are useful for studying the biological impact of MEIs in humans.

Results

Overview of MELT

As outlined above, there is an unmet need for a robust MEI discovery package that can comprehensively detect MEIs and their associated genetic features on a population scale in humans. As members of the 1000 Genomes Project, we developed the Mobile Element Locator Tool (MELT) to address this need. The 1000 Genomes Project was ideal for this purpose because we were faced with the challenge of performing MEI discovery in 2,534 human genomes. This included 2,504 low coverage Illumina whole genome sequences (averaging 7.4X coverage), and 30 high coverage Illumina whole genome sequences (averaging 60X coverage). We also sought to leverage the data that were collected across populations to construct comprehensive MEI models at each MEI site, which allowed us to more accurately discover MEI-associated features and genotypes. Finally, we wished to develop an expanded toolkit to track and study the MEIs that were discovered in these genomes. The MELT package includes a robust MEI discovery algorithm and a suite of MEI analysis tools that collectively achieve these goals.

MELT detects *Alu*, L1, and SVA MEIs by searching for signatures of discordant read pairs (DRPs) and split reads (SRs) in Illumina WGS data that are enriched at sites containing new, non-reference (non-REF) MEIs (Supplemental Fig. S1). MELT was designed to work with BAM files that are generated with the Burroughs-Wheeler Alignment tool (BWA-ALN or -MEM) (Li and Durbin 2009; Li and Durbin 2010), since most Illumina WGS data sets are directly available in this format. MELT first scans BAM files to identify a specific type of DRP (i.e., DRPs where one mate maps to the reference genome and the other maps to an *Alu*, L1, or SVA reference mobile element sequence), thus indicating the presence of a candidate non-REF MEI at the site. MELT uses SRs to further refine the precise breakpoints and TSDs at each candidate MEI site. When applied on a population scale, MELT constructs MEI models using all of the available DRP and SR data from multiple samples to accurately discover each MEI site and its features. MELT identifies a comprehensive set of MEI-associated features, including: the chromosomal insertion site, MEI orientation, TSD, internal mutation profile, subfamily, and other features, if present (Supplemental Fig. S2; Supplemental Table S1). MELT performs genotyping across all samples for both

novel (non-REF) and reference (REF) mobile element copies to provide a comprehensive map of polymorphic MEIs in a given genome. MELT also evaluates the potential impact of each MEI on nearby genes and lists the gene features that are impacted (e.g., promoter, coding exon, intron, UTR, or terminator). Finally, we developed a quality tranche system that leverages the evidence at each MEI site to estimate the quality of the MEI breakpoint (Supplemental Table S2; Supplemental Methods). These tools and features are all included in the comprehensive MELT ver. 2.0 package, which has several improvements over the original MELT ver. 1.0 that was used for the 1000 Genomes Project (Supplemental Table S1; Supplemental Table S3; see below).

MELT ver 2.0 was developed to work with diverse computational architectures and experimental designs. In this regard, several run modes were developed to provide flexibility in implementation, including 1) the single-sample (MELT-Single) mode, 2) the multiple-sample (MELT-Split) mode, where the four major steps of MELT are sequentially launched by the user, and 3) the multiple-sample (MELT-SGE) automated mode, where Sun Grid Engine (SGE) is used to automate the submission and processing of multiple samples (Supplemental Fig. S1). We also developed an Amazon Machine Image (AMI) version of MELT to facilitate MEI discovery in the cloud. The MELT-Single mode is useful for discovering and annotating MEIs in a relatively small number of genomes, whereas the MELT-Split and MELT-SGE multiple sample modes are engineered for population-scale studies involving hundreds or thousands of genomes.

As outlined above, an additional advantage of the multiple sample modes is that evidence for each MEI is drawn from multiple genomes (instead of just one) to identify the MEI site, associated features, and genotypes. To illustrate the increased sensitivity that is gained with the multiple sample MELT-SGE mode vs. the MELT-Single mode, we analyzed 10 low coverage (6.0 to 17.0X) CEU whole genome sequences and five high coverage (60X) whole genome sequences with both modes and compared the outcomes (please see Supplemental Table S4 for the genomes that were analyzed; The 1000 Genomes Project Consortium, 2015). The multiple sample MELT-SGE mode clearly increases the sensitivity of MEI detection compared to the MELT-Single mode for all three MEI classes (*Alu*, L1, and SVA; Supplemental Fig. S3).

Scalability of MELT with Illumina WGS data

We next compared the clock speed and scalability of MELT ver. 2.0 with four existing MEI detection tools, i.e., TEMP (Zhuang et al. 2014), RetroSeq (Keane et al. 2013), Mobster (Thung et al. 2014), and Tangram (Wu et al. 2014). MELT ver 2.0 had the fastest runtimes among the five MEI detection tools at both 6X and 30X coverages using Illumina WGS data from sample NA12878 (Figure 1A; Supplemental Table S5; The 1000 Genomes Project Consortium 2015). We also compared the scalability of these tools using ten low coverage CEU genomes (Figures 1B,C). For this test, we examined each tool for its ability to perform MEI discovery in one to ten low coverage (6.0 to 17.0X) CEU whole genome sequences (Supplemental Table S4; The 1000 Genomes Project Consortium 2015). Again, MELT had the best performance in these scalability tests (Figures 1B,C; Supplemental Table S5).

By extrapolating these scalability curves to all 2,504 low coverage genomes that were sequenced by the 1000 Genomes Project, MELT was predicted to require 21.9 days to perform MEI discovery in all 2,504 genomes (Figure 1B—top). This estimate is in good agreement with the MELT runtimes that we actually observed when we processed all 2,504 genomes for the 1000 Genomes Project on our computational cluster (approximately 2.5 weeks). In contrast, Tangram was estimated to require 84 days (12 weeks), Mobster 154.8 days (22.1 weeks), and TEMP 299.7 days (42.8 weeks), to complete MEI discovery in these genomes (Figure 1B). Therefore, on the basis of these tests, MELT is estimated to perform population-scale MEI discovery ~3.8-fold faster than Tangram, ~7.1-fold faster than Mobster, and ~13.7-fold faster than TEMP in head-to-head comparisons (Figure 1B; Supplemental Table S5).

Simulation studies to evaluate sensitivity and specificity

To evaluate the sensitivity and specificity of MELT ver. 2.0, we conducted a series of simulation studies (Figures 1D-K; Supplemental Figs. S4,S5; Supplemental Tables S5,S6). Briefly, a test set of *Alu*, L1, and SVA MEIs was inserted randomly into the reference human genome multiple times to generate a series of simulated genomes containing new MEIs. We used similar sets of non-REF MEIs that were discovered in the NA12878 genome (The 1000 Genomes Project Consortium, 2015, Sudmant et al. 2015; n = 1,114, including 922 *Alu*s, 146 L1's, 46 SVAs) but redistributed them randomly in the reference

human genome, and replicated this process 50 independent times (Methods; Supplemental Methods). We then generated FASTQ files from these genomes using an Illumina paired end read simulator at 7.5X, 15X, 30X and 60X coverage (Li and Durbin 2010). Finally, we mapped these reads to the reference human genome, generated BAM files, and tested MELT's ability to detect these artificially-inserted MEIs. In head-to-head comparisons with TEMP, RetroSeq, Mobster, and Tangram, these simulations indicated that MELT had the best sensitivity and specificity curves for all three classes of MEIs over the range of simulated WGS coverages that were tested (Figures 1D-K; Supplemental Table S5). Extensive PCR-based validations with sites selected from the 1000 Genomes Project data set (64 *Alu* sites, 54 L1 sites, and 59 SVA sites—a total of 177 sites that were discovered with MELT ver. 1.0) were in agreement with these MELT simulations (Sudmant et al. 2015). Likewise, an additional 90 PCR validations with MEI sites that were discovered with MELT ver. 2.0 also were in agreement with these simulations (Supplemental Methods; Supplemental Figs. S6,7; Supplemental Table S7—See below). Overall, these benchmarking, simulation, and PCR validation tests indicate that MELT ver 2.0 outperforms existing MEI discovery tools in terms of speed, scalability, sensitivity, and specificity, while also detecting a broader spectrum of MEI-associated features (Figure 1; Supplemental Figs. S4-7; Supplemental Tables S1,S5-7).

New 1000 Genomes Project and chimpanzee call sets

We next used the improved MELT ver. 2.0 package to rediscover *Alu*, L1, and SVA MEIs in the 2,504 low coverage and 30 high coverage human genomes that were sequenced for Phase III of the 1000 Genomes Project. The resulting MEI call sets are more extensive than the original 1000 Genomes Project Phase III MEI call sets and include additional MEI features (Sudmant et al. 2015; Supplemental Fig. S6). Our new call set includes 6,089 or 36.6% additional MEI calls, the majority of which are rare MEIs that were not detected in our previous 1000 Genomes Project call sets (Supplemental Fig. S6; Sudmant et al. 2015). MELT ver 2.0 employs slightly different rules for using DRPs and SRs compared to MELT ver 1.0 (Methods), and these changes led to the improved detection of rare MEIs while retaining similar overall results in benchmarking and validation tests (Figure 1; Supplemental Figs. S6,7;

Supplemental Tables S5-7). The *Alu*, L1, and SVA MEIs that we discovered with MELT ver. 2.0 had frequency distribution curves that were remarkably similar to the SNP frequency curve that was generated for the same samples by the 1000 Genomes Project (Supplemental Fig. S6). In contrast, the MEIs that were called with MELT ver. 1.0 did not resemble this SNP curve as closely, and lacked sensitivity in the lowest allele frequency bin (labeled 0.00; Supplemental Fig. S6).

We also adapted MELT ver 2.0 to perform MEI discovery in chimpanzees. MELT was designed to be flexible in this regard, and the requirements for adapting MELT to a new species are fairly minimal: 1) a reference genome sequence must be available for the species and 2) a set of reference endogenous MEI sequences must be available (or generated) for the species. Both of these requirements were met for chimpanzees, and thus, we used MELT to perform MEI discovery on 25 chimpanzees whose genomes had been sequenced previously (Prado-Martinez et al. 2013). Our chimpanzee MEI data set includes 7,278 *Alu* and 4,381 L1 MEIs (Supplemental Fig. S8). These chimpanzee data, and similar data sets generated with other organisms including canines (Gardner and Devine, unpublished), demonstrate that MELT can be readily adapted to other organisms.

Interior mutations and subfamily analysis

Mutations are encountered frequently within the interior sequences of MEIs due to the error-prone L1 reverse transcriptase that replicates *Alu*, L1, and SVA elements (Gilbert et al. 2005). These mutations (and patterns of mutations) have been useful for determining whether a given MEI belongs to a lineage that is known to be active in humans (Boisonott et al. 2000; Batzer and Deininger 2002; Wang et al. 2005; Konkel et al. 2015). Thus, we developed new MELT tools to identify interior mutations in MEIs and assign MEIs to lineages (or subfamilies). Since *Alu* and L1 MEIs together represent 95.3% of the MEIs that were discovered in the 1000 Genomes Project, we initially focused on developing tools for these two element classes. Specifically, we developed a tool named C*Alu* that identifies interior mutations within *Alu* elements and then uses these mutations to assign *Alu*s to known subfamilies, and a similar tool named LINEu that carries out comparable functions for L1 elements. These tools were validated in the simulation studies outlined above (Supplemental Fig. S4) and then used to identify interior mutations in

the updated 1000 Genomes Project call set. The resulting subfamily analysis revealed that the distributions of active *Alu* and L1 MEIs in the 1000 Genomes Project data sets were similar to those observed in previous studies in humans, thus providing additional validation for our methods. For example, *Alu*Ya5, and *Alu*Yb8 elements, which are known to be the most abundant *Alu* subfamilies in humans (Batzer and Deininger 2002; Bennett et al. 2008), also were the most abundant *Alu* MEIs that we discovered in the updated 1000 Genomes Project data set (Supplemental Fig. S9). Likewise, extensive testing of LINEu on fully-sequenced FL-L1 elements indicated that LINEu accurately identifies known human-specific L1 subfamilies (L1-Ta0, L1-Ta1, L1-Ta1d, and L1-Ta1nd; Scott et al. 2016).

Stratification of MEIs in 1000 Genomes Project populations

We next examined the population stratification of MEIs in the updated call sets that we generated from the 1000 Genomes Project. Varying degrees of *Alu*, L1, and SVA sharing were observed across the four major continental groups of the 1000 Genomes Project. *Alu* subfamilies such as *Alu*Ya5, *Alu*Yb8, *Alu*Yc1, *Alu*Y, *Alu*Yg6, and *Alu*Yk13 included copies that were shared by all non-admixed continental groups, as well as those that were found in a subset of groups, or in a single group (Figure 2A). None of these copies was found in the chimpanzee data set. These data indicate that several major human-specific *Alu* subfamilies have been active for most of modern human history, since these subfamilies have generated some MEI loci that are sufficiently old to be found in all modern humans, while generating other loci that are sufficiently young to be restricted to a single continental group or subpopulation.

We also examined the distributions of 79 novel *Alu* subfamilies that we identified in the updated 1000 Genomes Project MEIs through analysis of shared interior mutation patterns (Figures 2B,C; Supplemental Table S8; Supplemental Fig. S10). Although some of these novel *Alu* subfamilies were found in all four of the major continental groups of modern humans, others had very unequal distributions. For example, Family F is mostly restricted to the AFR continental population, whereas Family G is enriched in the EAS continental population (Figures 2B,C). Many of these very young *Alu* subfamilies have unique blends of sharing and a sampling of the spectrum of sharing is depicted in

Figures 2D to I. For example, Family I includes copies that are shared only by AFR and SAS individuals, suggesting that these copies might have been influenced by population bottlenecks during migration out of Africa (OOA) or by admixture. Overall, these data reveal complex patterns of *Alu* subfamily stratification in the 1000 Genomes Project populations, likely reflecting diverse patterns of demographic histories and other population forces affecting MEI dynamics (see below).

Active full-length L1 (FL-L1) source elements in human populations

We also developed two new tools to identify FL-L1 source elements that recently have generated MEI offspring in humans. One of these tools leverages 3' transduction events (Figure 3), and the other leverages interior mutation profiles, to track source/offspring relationships (Scott et al. 2016). A 3' transduction event occurs when a short segment of adjacent genomic sequence is incorporated into an offspring MEI during retrotransposition (Moran et al. 1999). A 3'-transduction is initiated at the level of transcription: transcripts originating from a FL-L1 source element bypass a weak poly(A) signal at the 3' end of the element and instead use an alternative poly(A) signal that is encountered in the adjacent downstream genomic region. When the resulting chimeric FL-L1 transcript is used as a replication template during retrotransposition, 3' adjacent genomic sequences are replicated and mobilized along with the L1 source element. These 3' transductions can be used to track source/offspring relationships because they serve as unique address tags that are associated with a single L1 source element and its offspring (Moran et al. 1999; Tubio et al. 2014). Thus, we developed a new tool to study source/offspring relationships using 3' transductions and validated it with simulations (Supplemental Methods; Supplemental Fig. S4F; Supplemental Table S6).

By applying this 3' transduction tool to the 2,504 low coverage human genomes that were sequenced by the 1000 Genomes Project, we identified 121 L1 offspring insertions that carried 3' transductions (Figure 3). We then used these unique 3' transduction tags to identify the 38 FL-L1 source elements that produced these insertions. Three of the 38 FL-L1 source elements, the Chr2:156527848, Chr6:13191033, and Chr1:119394974 elements, were exceptionally active and collectively generated more than half (68/121 or 56.2%) of the offspring MEIs (Figures 3A-E). The most active element among

these FL-L1 elements, the Chr2:156527848 element, is a previously-identified “hot L1” source element known as *LRE3* (Brouha et al. 2002), and it alone generated 41/121 (33.9%) of the offspring insertions (Figures 3A-C). The Chr6:13191033 and Chr1:119394974 FL-L1 elements generated an additional 14 and 13 offspring, respectively (Figures 3A,B,D,E), while the remaining 35 FL-L1 source elements each generated between one and four offspring (Figures 3A,B; Supplemental Table S9).

We found that the three highly active FL-L1 source elements (i.e., *LRE3*, Chr6:13191033, and Chr1:119394974) had diverse patterns of stratification among the 1000 Genomes Project populations. *LRE3* is clearly enriched in OOA populations (Figures 3G,H), and has very low allelic frequencies in African (AFR) populations (with MAFs ranging from 0 to 0.025 in the six AFR subpopulations). Likewise, offspring insertions produced by *LRE3* followed a similar pattern of enrichment in OOA populations, with the majority of offspring localized to one or more OOA populations (Figure 3G,H). *LRE3* itself has generated at least 20 FL-L1 offspring insertions that could, in principle, serve as new FL-L1 source elements (Figure 3F,G). We fully sequenced eight of these 20 FL-L1’s (Figure 3F), and all eight had two intact open reading frames (ORFs), providing further support that they might serve as active source elements. Interestingly, five of these eight FL-L1 insertions were found in only one OOA population (Figure 3F). The Chr6:13191033 FL-L1 source element and its offspring had population distributions that were very similar to *LRE3*, whereas the Chr1:119394974 FL-L1 source element and its offspring were more evenly distributed among the four continental groups (Figure 3H). These correlated patterns of FL-L1 source elements and their offspring suggest that L1 mutagenesis is influenced by the stratification of these highly active source elements. Moreover, the birth of new source elements within these lineages might further enhance L1 mutagenesis in these lineages over time.

The remaining 35 FL-L1 elements that produced 3’ transductions fell into three categories: i) source elements that were shared almost equally across the four continental populations (Figure 3H left, middle panel), ii) those that were shared unequally by the four continental populations (Figure 3H center, middle panel), and iii) those that were found in less than four (between one and three) continental populations (Figure 3H right, middle panel). Unlike the three highly active elements described above (*LRE3*, Chr6:13191033, and Chr1:119394974), which had fairly correlated patterns of source and

offspring distributions, the offspring patterns for these remaining 35 FL-L1 source elements were very diverse and often were not correlated with the distributions of their respective source elements (Figure 3H, lower panel). For example, even though some FL-L1 elements were present in all four continental groups (Figure 3H center, middle panel), the offspring from these elements often were found in only one or two continental populations (Figure 3H center, lower panel). In some cases, this might be due to an ascertainment bias caused by the small number of offspring that were produced by these elements. However, it is also possible that some of these FL-L1 source elements are active in a subset of the populations in which they reside. Overall, these data reveal diverse relationships of FL-L1 source elements and their offspring in modern human populations.

The complete interior sequences of the FL-L1 source elements that produced 3' transductions in our study were obtained for 34/38 (89.5%) of the elements (either from the reference genome or from PacBio sequencing; Supplemental Methods; Supplemental Table S9). Most of the sequenced elements (25/34 or 73.5%) have two intact ORFs and belong to one of four active L1Ta subfamilies (Ta0, Ta1, Ta1d, Ta1nd). Many of these elements also have been found to be active in a cell culture-based assay for L1 retrotransposition (Brouha et al. 2002; Brouha et al. 2003; Beck et al. 2010). Thus, at least some of these elements are likely to have retained the ability to produce MEIs in humans. The remaining 9/34 (26.5%) of the sequenced source elements had only one or zero intact ORFs, and thus, appear to have accumulated deleterious mutations that have rendered them inactive. These include one PA2 element, four Ta elements (Ta0, Ta1, Ta1d, Ta1nd), and one non-canonical L1 element. Our analysis also revealed a range of poly(A) signal configurations for these FL-L1 elements, suggesting that the underlying reasons for the production of 3' transductions are complex (Supplemental Discussion, Supplemental Table S9).

Comparisons with L1 source elements that are somatically active in cancer genomes

We next compared the active FL-L1 source elements that we identified through 3' transductions in the 1000 Genomes Project samples (Figure 3) with those that had been reported previously in the literature (Supplemental Table S9). A total of 113 non-redundant FL-L1 source elements were identified

in these collective studies (including our study; Figure 4A). 46/113 (40.7%) of these FL-L1 source elements were active exclusively in the germline, whereas 48/113 (42.5%) were active exclusively in somatic cancer tissues (Figure 4A; Supplemental Fig. S11). Another 19/113 (16.8%) were active in both the germline and somatic tissues (Figure 4A). The three most active FL-L1 source elements that we identified in our study also were among the most active elements in a large-scale somatic cancer study (Tubio et al. 2014; Figure 4B). Thus, some FL-L1 source elements are highly active in both the germline and somatic tissues, whereas others are active in only one of these tissue types.

We also compared these germline and somatic FL-L1 source element activities (Figure 4) with those that had been measured previously in a cell culture-based assay for L1 retrotransposition (Moran et al. 1996; Brouha et al. 2002; Brouha et al. 2003; Beck et al. 2010; Figure 4C). The *LRE3* FL-L1 source element, which is highly active in both the germline and somatic cancer tissues, is also highly active in the cell culture-based retrotransposition assay (Figure 4C; Brouha et al. 2002; Beck et al. 2010). Although several other FL-L1 source elements had similarly correlated levels of retrotransposition in the germline, somatic tissues, and cultured cells, other elements behaved quite differently in these three cellular environments. For example, the Chr22:29059271 FL-L1 source element is highly active in somatic cancer genomes, but is relatively inactive in both the germline and the cell culture-based assay (Figure 4C; Brouha et al. 2003; Tubio et al. 2014). Other discordant patterns of activity also were observed among these tissues and assays (Figure 4C). Overall, these data indicate that a given FL-L1 source element can have remarkably different levels of activity in these three cellular environments.

5’ inversions

L1 elements often produce MEI offspring that have 5’ inversions (i.e, the 5’ portion of the MEI is inverted relative to the 3’ portion). These 5’ inversions have been proposed to be caused by a mechanism termed “twin priming”, whereby the TPRT process is initiated simultaneously from both strands of DNA at the genomic integration site (Ostertag et al. 2001). Another feature of these 5’ inverted L1 MEIs is that they often have small insertions, deletions, or duplications at the inversion junctions.

However, these MEIs otherwise appear to have all of the typical features of L1 insertions, including poly(A) tails, TSDs, and the other features outlined in Supplemental Table S1.

Because 5' inversions occur frequently, we developed a new tool to identify 5' inversions in L1 offspring elements and validated it in simulation studies (Supplemental Fig. S4G,H; Supplemental Table S6). We then applied this tool to the 1000 Genomes Project samples and found that 298/1634 (18.2%) of the L1 MEIs discovered in the 1000 Genomes Project samples had 5' inversions (Note: in some cases we could not measure 5' inversions due to a lack of DRP/SR evidence at one end of the L1 MEI). The inversion junctions for these 298 non-REF MEIs generally were located throughout the reference L1 sequence as reported previously for older REF L1 elements (Szak et al. 2002; Figures 5A,B). However, we noted a depletion of 5' inversions near the 5' end of L1. In fact, we did not identify a single 5' inversion junction in the first 590 bp of L1, despite the fact that 33.8% of the non-REF L1 MEIs discovered in the 1000 Genomes samples were either full-length or otherwise contained sequences that spanned this region (Figures 5A,B). We verified that MELT had sufficient sensitivity in the first 590 bp to detect 5' inversions in this region (Supplemental Fig. S4H; Supplemental Methods). These data suggest the possibility that the 5' inversion mechanism requires ~500 bp of free RNA or DNA at the 5' end of L1, perhaps to loop back and serve as a priming site for DNA replication.

We also examined the rates at which 5' inversions are produced from diverse FL-L1 source elements, and whether these rates vary from one FL-L1 source element to the next. To explore this question, we examined the 5' inversion rates for the FL-L1 source elements that were associated with 3' transductions in the 1000 Genomes Project data set (Figure 3). Interestingly, the three most active FL-L1 source elements from this data set (i.e., *LRE3*, Chr6:13191033, and Chr1:119394974; Figure 3) generated 5' inversions at very different rates (i.e., 55.6% of offspring had 5' inversions for the Chr1:119394974 FL-L1 source element, 14.3% for the *LRE3* FL-L1 element, and 0.0% for the Chr6:13191033 FL-L1 element; Figure 5D; Supplemental Table S10). We also examined the 5' inversion rates in several additional studies involving germline and somatic MEIs and likewise observed a range of 5' inversion rates in these studies (Figure 5D-F; Supplemental Table S10). These data suggest that the

rates at which FL-L1 source elements produce 5' inversions may vary among FL-L1 source elements and across diverse cellular environments.

Ancient MEIs in Neanderthal and Denisovan genomes

Ancient genomes from Neanderthals and Denisovans have been technically challenging to sequence and analyze, and thus far, no MEIs have been successfully discovered in these archaic hominids. We next determined whether MELT could detect *Alu*, L1, and SVA MEIs in these genomes. Indeed, MELT successfully detected 41 ancient *Alu* MEIs in Neanderthals and 127 ancient *Alu* MEIs in Denisovans that were not found in chimpanzees or modern humans (Figure 6A). We also discovered another ten ancient *Alu* MEIs that were shared by Neanderthals and Denisovans but were absent from chimpanzees and modern humans (Figure 6A). Similarly, 26 ancient L1 insertions, and three ancient SVA insertions, were identified in Neanderthals and Denisovans that were absent from chimpanzees and modern humans (Figure 6B and Supplemental Table S11). Thus *Alu*, L1, and SVA elements appear to have been active in ancient hominids during the period when they were temporally and geographically separated from modern humans (~86,000 to 800,000 years ago; Sankararaman et al. 2014; Vattathil and Akey 2015).

We also identified 272 *Alu*, 39 L1, and 13 SVA elements that were shared by ancient hominids and modern humans but were absent from chimpanzees (Figure 6A-D; Supplemental Table S11). In some cases, these shared MEIs likely were generated in a common ancestor prior to the migration of ancient hominids and modern humans out of Africa. However, 49 MEIs (42 *Alu* and 7 L1 MEIs) were shared exclusively between ancient hominids and modern OOA populations (i.e., they were absent from chimpanzee and AFR populations; Figure 6E; Supplemental Table S11). This class of sharing suggested the possibility that at least some of these MEIs initially were generated in ancient hominid genomes and then moved into modern human genomes through introgression during periods when ancient and modern humans cohabitated Europe and Asia. Indeed, archaic SNP haplotype maps (Sankararaman et al. 2014) indicated that these MEIs were almost exclusively embedded within Neanderthal and Denisovan haplotypes, supporting the idea that these ancient MEIs were integrated in the context of ancient

genomes and then migrated into modern humans through introgression (Figure 6E). Among these introgressed MEIs, we identified a FL-L1 element that might have been active in both ancient and modern humans (Figure 6G; See Discussion). These MEI introgression data are in agreement with SNP-based models of introgression, whereby Neanderthal introgression is observed in all OOA modern populations (Figure 6C) and Denisovan introgression is found mainly in East Asian populations (Figure 6D). Thus, introgression of ancient *Alu*, L1, and SVA MEIs into modern human genomes is yet another mechanism of MEI dynamics in humans.

Discussion

We have developed the MELT package of computational tools to efficiently discover and study MEIs in WGS projects. In head-to-head tests, we found that MELT outperformed existing MEI discovery tools in terms of speed, scalability, sensitivity, and specificity, while also detecting a broader range of MEI-associated features (Figure 1; Supplemental Table S1). In addition to the basic MELT discovery algorithm, we also developed a set of companion tools to study the MEIs that are discovered by MELT (Supplemental Table S1). We likewise developed several run modes to improve the portability of MELT and to provide flexibility in experimental design. In addition to using MELT with the 1000 Genomes Project samples (Sudmant et al. 2015; Supplemental Fig. S6), we also have used it to discover MEIs in chimpanzees, ancient Neanderthal and Denisovan genomes (Figure 6), cancer genomes (Scott et al. 2016), and canines (Gardner and Devine, unpublished). Thus, in principle, MELT could be used with any species or experimental design, provided that a reference genome sequence and a set of reference mobile element sequences is available for the organism of interest.

Population dynamics of MEIs

Our study revealed extensive MEI stratification across diverse human populations. For example, the active FL-L1 source elements that we identified through 3' transductions often were unequally distributed among the major continental groups of modern humans, and this appears to have led to

differences in the production of L1 offspring in some cases (Figure 3). Since FL-L1 source elements also are responsible for generating *Alu* and SVA MEIs, a highly active population-specific source element could have a major impact on the stratification of these other two classes of MEIs as well. In some cases, MEIs were shared by all continental groups and ancient hominids, but were absent from chimpanzees, indicating that the MEI was generated very early in human history and is shared throughout the human and hominid lineages as a consequence of common ancestry. In other cases, MEIs were found only in ancient Neanderthal or Denisovan genomes, or were restricted to AFR or OOA populations, suggesting that they were generated more recently. However, the complexity of sharing that we observed suggests that other forces such as admixture and bottlenecks likely have influenced the stratification of these elements as well. Our data provide an extensive map of MEIs extending from chimpanzees, ancient hominids, and modern humans, and document these diverse MEI sharing patterns on the largest scale that has been examined to date.

We also noted a novel process that affected the population dynamics of MEIs: the introgression of ancient MEIs from Neanderthals and Denisovans into modern humans. Our data indicate that the mobilome of modern humans has been shaped at least partly by ancient *Alu*, L1 (and likely SVA) elements that initially were generated in archaic genomes and subsequently were introduced into modern humans through introgression. Ancient Neanderthals and Denisovans were exposed to harsh selective pressures in the face of new environmental challenges (Vattathil and Akey, 2015) and some ancient MEIs could, in theory, have produced adaptive changes in these hominids. Such MEIs could, in turn, have been rapidly passed into modern humans through introgression, which could have conferred a selective advantage to recipients as they faced the same challenges. Thus, ancient MEIs that predated modern humans could have helped to influence modern human phenotypes through novel genetic mechanisms. As additional Neanderthal and Denisovan genomes are sequenced, it should be possible to further explore this possibility.

We also identified a FL-L1 element among the introgressed MEIs that might have served as an active source element in both ancient hominids and modern humans (Figure 6G). This element was fully sequenced from a modern human genome using a PacBio-based approach that we developed recently

(Scott et al. 2016). We found that the element is 6,015 bp in length, has two intact ORFs, and belongs to the L1-Ta1d subfamily, which is one of the most active human-specific L1 subfamilies (Supplemental Table S9). The same L1-Ta1d subfamily gave rise to the highly active *LRE3* FL-L1 element (Figures 3,4) and other "hot" L1 source elements that have caused human diseases (e.g., see Scott et al. 2016). These data indicate that the L1-Ta1d subfamily originated sufficiently early to be present in ancient Neanderthal and modern human genomes, suggesting that this highly active subfamily was generated in a common ancestor at least ~800,000 years ago. Since the L1-Ta1d subfamily has generated a number of diseases in modern humans, it likely caused diseases in ancient hominids as well.

Recently-active FL-L1 source elements identified with 3' transductions

Our 3' transduction tracking data provide evidence for many novel FL-L1 source elements that have been active recently in modern humans. Although we identified a total of 38 active FL-L1 source elements in our studies, just three of these elements produced the majority of germline insertions that are associated with 3' transductions in the 1000 Genomes Project populations (Figures 3,4). Many of the FL-L1 elements that produced 3' transductions in our data sets (Figures 3,4) also have been reported to be active in somatic tissues or in cell culture (Moran et al. 1996; Brouha et al. 2003; Beck et al. 2010; Macfarlane et al. 2013; Tubio et al. 2014). In some cases, FL-L1 elements were active in germline tissues, somatic cancers, and cultured cells (e.g., the Chr2:156527848 element (*LRE3*); Figure 4C). However, other FL-L1 source elements behaved very differently in these three cellular contexts. For example, the Chr22:29059271 element was very active in somatic cancers, but had very low levels of activity in germline tissues and cultured cells (Figure 4C; Supplemental Table S9). The rates at which 5' inversions were generated also varied in these diverse settings (Figures 5D-F, Supplemental Table S10).

There are several factors that might help to explain these differences in FL-L1 behavior. First, FL-L1 elements occasionally can have two or more alleles that occupy the same locus, and these alleles can support very different levels of retrotransposition due to internal mutations that affect functionally important sites within L1 (Seleme et al. 2006). Under this scenario, the same FL-L1 locus might appear to have different levels of activity because different alleles of the element are being examined. A source

element also could be heterozygous or homozygous at a particular locus, which might also influence the apparent rate of offspring production. In other cases, the methylation status or chromatin state of the FL-L1 element may influence the rate of retrotransposition in various tissues (Borc'his and Bestor, 2004; Iskow et al. 2010; Scott et al. 2016). A range of additional host factors that are exclusively expressed in germline or somatic tissues also might be envisioned to differentially influence the activity of a given FL-L1 element. Irrespective of the mechanism(s) underlying these differences, our data suggest that FL-L1 elements likely help to shape human traits and diseases in complex ways, depending on the populations and tissues in which they are active.

Methods

Description of the MELT pipeline

MELT is coded in Java (release 1.8) and uses several external libraries (Supplemental Table S12). For each genome analyzed, MELT parses WGS data that is aligned with BWA-MEM or -ALN (Li and Durbin 2010) for DRPs (defined as mates that are either aligned to different chromosomes, or separated by at least 1 Mbp). DRPs are then aligned to mobile element (ME) reference sequences (Supplemental Table S13; Dombroski et al. 1993; Stewart et al. 2011) using Bowtie 2 (Langmead and Salzberg 2012). DRPs where one mate maps to the human reference sequence and the other maps to a ME reference sequence (Supplemental Table S13) are then used for MEI discovery by 'walking' across the reference genome using the reference-aligned mate, seeking clusters of at least four DRPs. Sites are filtered based on proximity to reference MEs (Smit AFA 1996-2010; Karolchik et al. 2004), surrounding sequencing depth, location in relation to reference sequence gaps, and the mapping quality of the reads. After MEI sites are identified, DRP and SR evidence is used to discover MEI-associated features and precise breakpoints (Supplemental Table S1). MEI sites are then genotyped in all samples using a modified version of the algorithm described in Li (2011). Following genotyping, sites are filtered based on 5' and 3' supporting evidence, total percentage of no-call (i.e. './.') genotypes, and total number of SRs. Sites are then merged into a VCF 4.2 format file (Danecek et al. 2011).

Simulated data sets and validation of MELT features

To facilitate *in silico* analyses, we generated 50 simulated human genomes with computationally inserted MEIs containing diverse features. For simulated insertions, the NA12878 genome was selected to represent a typical distribution of elements (*Alu* 922, L1 124, SVA 46). For each MEI type, a full-length consensus sequence (Supplemental Table S13; Dombroski et al. 1993; Stewart et al. 2011) was randomly modified with known ME features (Supplemental Table S1, Supplemental Fig. S2, Supplemental Methods). Using bedtools (Quinlan 2014), each MEI then was randomly inserted into an accessible locus of the hg19 genome, as demarcated by the 1000 Genomes Pilot Accessibility Mask (The 1000 Genomes Project Consortium 2015). Simulated reads were then generated with wgsim (Li and Durbin 2010) at 60X coverage (read length 100 bp, fragment length 500 bp, zero base error rate), and aligned with BWA (Li and Durbin 2009). Each BAM file was additionally down-sampled to 30X, 15X, and 7.5X with Picard Tools (http://broadinstitute.github.io/picard/) to evaluate MEI discovery performance at various coverage levels. For deletion analysis, we simulated 25 genomes with 400 *Alu* and 50 L1 polymorphic sites randomly selected from a collection of 1719 *Alu* and 139 L1 reference sites that are known to be polymorphic in humans (Smit AFA 1996-2010; Karolchik et al. 2004; Sudmant et al. 2015). We performed MEI discovery using MELT-Single (n = 50) and MELT-DEL (n = 25) (Supplemental Fig. S1). The wgsim read simulator that we used enabled us to simulate diploid human genomes, which allowed us to model both heterozygous and homozygous MEI sites. The ratio of heterozygous to homozygous non-REF MEIs was weighted according to the relative frequencies of these events in actual data (95% heterozygous and 5% homozygous non-REF). Although wgsim does not model the error profile of Illumina sequencing, the FDRs of our simulations were in good agreement with those obtained in our PCR validations, suggesting that sequencing errors likely do not have a major impact on MEI discovery (Figure 1; Supplemental Fig. S7).

Comparison of runtime, scalability, sensitivity, and specificity

To test the relative runtime of MELT and four additional MEI detection pipelines (Keane et al. 2013; Thung et al. 2014; Wu et al. 2014; Zhuang et al. 2014) we analyzed the NA12878 genome at ~6X

(100 bp paired-end Illumina WGS) and ~30X (250 bp paired-end Illumina WGS) coverages without multithreading or distributed computing. Genomes were downloaded as raw FASTQ files (The 1000 Genomes Project Consortium 2015), and alignments were performed using BWA-MEM (Li and Durbin 2010). For Tangram and Mobster, Mosaik (Lee et al. 2014) alignments were generated with identical input FASTQ files, and used for MEI discovery. Scalability was evaluated using the default parameters for each algorithm on one to ten genomes with five replicates (Supplemental Table S4). Samples were added in the same order for each algorithm and each replicate. Multithreading or parallelization was enabled for algorithms that support these approaches (i.e., MELT, TEMP, and Tangram). Tangram was run with several different multithreading parameters (with either 1, 2, or 4 cores per chromosome during the tangram_detect stage; Figure 1; Supplemental Table S5). All reported times are actual runtimes (i.e. start to finish, not CPU time; Figure 1B,C; Supplemental Table S5). System specifications for the machines that were used for testing are reported in the Supplemental Methods. Sensitivity and specificity was tested in the five MEI detection algorithms using the 50 simulated data sets described above. Each tool was run according to the algorithm documentation using default parameters at four coverage levels (7.5X, 15X, 30X, and 60X) for the three human MEs (*Alu*, L1, and SVA). A site was considered correct if it fell within +/- 500 bp of the actual site (Quinlan 2014).

MEI data sets and analysis

MEI discovery was performed in the 1000 Genomes Project Phase III data sets, chimpanzees, and ancient humans with MELT ver. 2.0 as outlined in the Supplemental Methods. PCR validations with 90 new MEI sites from the MELT ver. 2.0 1000 Genomes Project dataset were conducted as outlined previously (Sudmant et al. 2015; Supplemental Methods; Supplemental Fig. S7; Supplemental Table S7). *Alu* subfamily annotation and stratification analysis was conducted as outlined previously (Batzer et al. 1996; Supplemental Methods; Supplemental Table S8). Analysis of L1 3' transductions, L1 5' inversions, L1 germline vs. somatic activity, and archaic MEIs was performed as outlined in the Supplemental Methods, Supplemental Fig. S11, and Supplemental Tables S9-11.

Data access

All MELT ver. 2.0 call sets from this study have been deposited to the dbVar database at NCBI (https://www.ncbi.nlm.nih.gov/dbvar) under accession number nstd144. The MELT ver. 2.0 package is available at the MELT download site (http://melt.igs.umaryland.edu).

Acknowledgements

We thank The 1000 Genomes Project Consortium structural variation group for valuable input on MELT development. We thank Naomi Sengamalay, Sandra Ott, Kelly Klega, Xuechu Zhao, Alvaro Godinez, Lisa Sadzewicz, and Luke Tallon for assistance with PacBio sequencing. We thank Clara Daly and Victor Felix for assistance with the MELT download site. This work was funded by the following grants: F31 HG009223 (EJG), T32 DK067872 (NTC), R01 CA166661 (SED), and R01 HG002898 (SED).

Author contributions

E.J.G. developed MELT and associated tools, performed PCR validation studies, and generated MEI call sets in chimpanzees and modern humans; V.L. and E.J.G. performed simulation studies; D.H. and E.J.G. performed MEI discovery in ancient genomes; N.T.C. and E.J.G. performed long-range PCR and PacBio sequence analysis of FL-L1 elements; members of The 1000 Genomes Project Consortium provided valuable input on MELT development, testing, and validation; W.S.P. converted MELT into an AMI for cloud computing; E.C.S. and E.J.G. examined 5' inversions; R.E.M. and E.J.G. developed C*Alu* and LINEu; E.J.G. and S.E.D. designed MELT and associated tools, designed experiments, analyzed data, developed display items, and wrote the manuscript.

Disclosure declaration

The authors have no conflicts of interest to disclose.

Figure legends

Figure 1. Comparisons of MEI discovery algorithms.

(**A**-**C**) Runtime comparisons between MELT and four other MEI discovery algorithms: RetroSeq Mobster, Tangram, and TEMP. (**A**) Runtime in minutes on either a 6X or 30X coverage genome using a single processor (numbers are mean ± SD), with the best time for each coverage indicated in red. (**B**) Time required for each algorithm to analyze between one and 10 genomes using a distributed computing cluster. Shown to the right of experimental data are extrapolated estimates of the total run-times for 2,504 genomes with each algorithm. (**C**) Identical to (**B**), but depicting the median runtime for only MELT and Tangram. Tangram was run with 23, 46, or 92 threads (numbers to the right of lines). (**D**-**G**) Comparison of sensitivities for MELT and the MEI detection algorithms outlined above. False negative rates (FNRs) are plotted for (**D**) Aggregate, (**E**) *Alu*, (**F**) L1, and (**G**) SVA. (**H**-**K**) Comparison of specificities for MELT and the MEI detection algorithms outlined above. False discovery rates (FDRs) are plotted for (**H**) Aggregate, (**I**) *Alu*, (**J**) L1, and (**K**) SVA (Supplemental Table S5).

Figure 2. Complex patterns of *Alu* subfamily expansion in diverse human populations.

Six known *Alu* subfamilies (**A**) and 79 novel *Alu* subfamilies (**B**) that were identified using interior sequence changes were analyzed for sharing among the 1000 Genomes Project non-admixed continental populations. Plotted are: (top) Log_{10} total sites in each subfamily; (bottom) proportion of sites shared among all continental populations (blue); proportion of sites shared by two or three continental populations (green); proportion of sites that are specific to one continental population (brown). The average proportion for each category is indicated by a horizontal dotted line. (**C**) Tree of 79 novel *Alu*Y subfamilies. We required at least five independent copies with a novel set of interior mutations (excluding CpG sites) to establish new subfamilies. This threshold is fairly conservative, and eliminates errors introduced by Illumina sequencing. After C*Alu* classification (Supplemental Fig. S9), novel *Alu* subfamilies were placed on a tree of known *Alu*Y families (a, b, and c shown) and subfamilies (small black circles). Each pie chart represents the sum of allele counts for all constituent sites of a particular

novel subfamily with the total number of identical loci represented by the diameter of the pie (see Supplemental Fig. S10 for figure key). (**D**-**I**) Families with unique population sharing are shown with each pie representing the proportion of total alleles from each of the four major continental populations of the 1000 Genomes Project. Each site is placed into one of three categories based on population sharing: present in all four continental populations (left); in two or three continental populations (middle); or in one continental population (right). Pies are sized based on the Log_{10} allele frequency of each site. (*) actual AF is 0.52446. *Alu* subfamilies were named as outlined in Batzer et al. 1996 (Supplemental Table S8).

Figure 3. Analysis of L1 source-offspring relationships using 3' transductions.

(**A**) Pie chart depicting the proportion of offspring attributable to each of the 38 FL-L1 source elements identified in this study. 121/4,118 (2.9%) of the L1s identified had 3' transductions that could be used to identify the FL-L1 source elements that produced these offspring insertions (Supplemental Table S9). Note that this method can only be used to track source/offspring relationships for L1's that produce 3' transductions. The *LRE3*, Chr6:13191033, and Chr1:119394974 FL-L1 source elements are indicated in red, blue, and green, respectively. (**B**) Circos plot depicting the genomic landscape of source-offspring relationships summarized in (**A**). Red, blue, and green arrows indicate the three FL-L1 source elements highlighted in (**A**). (**C**-**E**) Individual Circos plots tracking offspring for the three most active FL-L1 source elements from (**A**). Each source-offspring relationship is colored based on the population in which the offspring element is found (grey if found in multiple populations). (**F**) *LRE3* was sequenced from an individual of European descent (top), along with eight FL-L1 *LRE3* offspring. Sequence changes compared to the L1.3 FL-L1 element (Dombroski et al. 1993) are shown as blue, green, yellow, red, or black vertical lines representing C, A, G, T, or deletion mutations, respectively. All eight sequenced FL-L1 offspring of *LRE3* have two intact ORFs (dark grey bars). The first poly(A) tail is shown in bright green, with transduced sequence shown in light grey. Offspring elements that have a 3' transduction also have a second poly(A) tail (bright green). The five population-specific FL-L1 elements are indicated by the 1000 Genomes Project population colors next to the elements. (**G**) *LRE3* transduction family, displayed in a similar manner to Figure 2 (AMR-specific offspring not shown; n=2). Each pie chart

represents either *LRE3* (labeled) or one *LRE3* offspring locus from (**C**). Borders of each pie are colored red if the element is a FL-L1 (n = 20), or purple if it has a 5' inversion (n = 2). (**H**) Source and offspring element population distributions. Shown for each source element is the total number of offspring (top), the population distribution of the source element (middle), and the aggregate population distribution of all offspring (bottom). Highlighted with colored arrows are source elements from (**A**). Red bars indicate where offspring were only found in the American continental population. Vertical black lines separate source elements into one of three classes: found in all populations (left), found predominately in OOA populations (middle), and found in a subset of other populations (right).

Figure 4. Recently-active FL-L1 source elements in human populations and cancers.

(**A**) Circos plot of the human genome with coordinates of known active FL-L1 source elements producing 3' transductions (circles; Supplemental Table S9). FL-L1s are further separated into one of three categories based on the tissue type(s) in which activity was recorded. The three most active FL-L1 source elements identified in this study are represented as circles corresponding to their colors in Figure 3A. (**B**) Log-Log plot depicting 14 L1 source elements that were active in the germline (this study) and somatic tumors (Tubio et al. 2014). The three most active germline elements (this study) are highlighted according to their colors in Figure 3A, and are active in somatic cancers as well. Note that one of the dots represents two separate L1 source elements, as both have the same number of germline and somatic offspring. (**C**) Comparison of FL-L1 element activity in the cell culture-based retrotransposition assay (percent of L1.3 or L1RP activity, light blue; Brouha et al. 2002; Brouha et al. 2003; Beck et al. 2010) with the total number of offspring identified in this study (light green) and the Tubio et al. study (light orange; Tubio et al. 2014). Only elements that were active in the germline (this study), cancers (Tubio et al.), and the cell culture-based assay were displayed.

Figure 5. L1 5' inversions in germline and somatic tissues.

(**A**) L1 length distribution among sites discovered in the 1000 Genomes Project Phase III samples. (**B**) 5' inversion positions discovered in the 1000 Genomes Project Phase III samples

(Supplemental Table S10). (**C**) Correlation between the distributions shown in (**A**) and (**B**), with the linear trend line and r^2 correlation shown in red. FL-L1 (red arrow in **A**) sites were excluded from this comparison because no correlation was observed in the first ~590 bp of FL-L1 elements. (**D**) 5' inversion rates among all L1 sites in the 1000 Genomes Project, chimpanzee, and among particularly active 3'-transducers highlighted in Figure 3. (**E**) Total number of 5' inverted sites from 1000 Genomes Project MEIs (this study) compared with other germline and somatic studies. The proportion of 5' inverted sites is significantly different (*, p = 0.0207) between germline and somatic insertions (Supplemental Table S10). (**F**) Comparison of germline 5' inversion rates (1000 Genomes Project MEIs) and several different tumor types analyzed by various studies (Supplemental Table S10).

Figure 6. Mobile element activity in ancient human genomes and introgression of ancient MEIs into modern humans.

(**A**,**B**) Sharing of (**A**) *Alu* and (**B**) L1 MEIs between Neanderthal, Denisovan, modern humans, and chimpanzees (Supplemental Table S11). (**C**,**D**) Sharing of Neanderthal and Denisovan *Alu* MEIs in each of the 26 1000 Genomes Project Phase III populations. For each population, we determined the average percentage per individual of *Alu* MEIs shared with (**C**) Neanderthal or (**D**) Denisovan. Heat maps represent multiple comparison ANOVA p-values between each population (key at right). (**E**) Analysis of Neanderthal MEI introgression in non-African individuals. Each bar represents one MEI site that was shared between Neanderthal and non-African individuals (i.e. the site was found only in SAS, EUR, and/or EAS). Bars are colored by MEI overlap with Neanderthal haplotypes (Supplemental Methods; Sankararaman et al. 2014), with sites to the left of the chart likely contributed to modern humans by introgression from Neanderthals, and sites to the right likely due to common ancestry. "HAP" indicates whether the Neanderthal haplotype is present at the site (HAP+ or HAP-). "MEI" indicates whether the MEI is present at the site (MEI+ or MEI-). Blue bars indicate a high degree of linkage disequilibrium (LD) between the Neanderthal haplotype and the MEI (HAP+/MEI+), and brown bars indicate little or no correlation between the Neanderthal haplotype and the MEI (HAP-/MEI+). Blue sites have r^2 values for HAP+/MEI+ of >0.5, whereas brown sites segregate independently (Supplemental Table S11;

Supplemental Methods). The black arrow indicates the FL-L1 element described in (**G**). (**F**) Analysis of Neanderthal MEI introgression in all individuals. Identical analysis to E, but for sites with an AFR allele frequency greater than zero. (**G**) Cartoon of a FL-L1 element sequenced from a GBR individual, with differences shown as in Figure 3F. The quality of ancient MEI calls was comparable to those called in modern humans (Supplemental Fig. S12).

References

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534-537.

Batzer MA, Deininger PL, Hellman-Blumberg U., Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E., and Zuckerkandl E. 1996. Standarized nomenclature for Alu repeats. *J. Mol. Evol.* **42**: 3-6.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370-379.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159-1170.

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

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.

Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17**: 915-928.

Borc'his D, Bestor TH. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96-99.

Brouha B, Meischl C, Ostertag E, de Boer M, Zhang Y, Neijens H, Roos D, Kazazian HH, Jr. 2002. Evidence consistent with human L1 retrotransposition in maternal meiosis I. *Am J Hum Genet* **71**: 327-336.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian, HH Jr. 2003. Hot L1s account for the bulk of retrotransposition activity in the human population. *Proc Natl Acad Sci USA* **100**: 5280-5285.

Bundo M, Toyoshima M, Okada Y, Akamatsu W, Ueda J, Nemoto-Miyauchi T, Sunaga F, Toritsuka M, Ikawa D, Kakita A, et al. 2014. Increased L1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* **81**: 306-313.

Chen JM, Masson E, Macek M, Raguenes O, Piskackova T, Fercot B, Fila L, Cooper DN, Audrezet MP, Ferec C. 2008. Detection of two Alu insertions in the CFTR gene. *J Cyst Fibros* **7**: 37-43.

Claverie-Martin F, Gonzalez-Acosta H, Flores C, Anton-Gamero M, Garcia-Nieto V. 2003. De novo insertion of an Alu sequence in the coding region of the CLCN5 gene results in Dent's disease. *Hum Genet* **113**: 480-485.

Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**: 1127-1131.

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.

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

Dombroski BA, Scott AF, Kazazian HH, Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc Natl Acad Sci USA* **90**: 6513-6517.

Doucet-O'Hare TT, Rodic N, Sharma R, Darbari I, Abril G, Choi JA, Young Ahn J, Cheng Y, Anders RA, Burns KH, et al. 2015. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc Natl Acad Sci USA* **112**: E4894-E4900.

Divoky V, Indra K, Mrug M, Brabec V, Huisman THJ, Prchal JT. 1996. A novel mechanism of beta-thalassemia. The insertion of L1 retrotransposable element into beta globin IVSII. *Blood* **88**: 148a

Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A et al. 2012. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**: 483-496.

Ewing AD, Kazazian HH, Jr. 2010. High throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**: 1262-1270.

Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, Abril G, Pereira G, Makohon-Moore A. et al. 2015. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res* **25**: 1536-1545.

Ganguly A., Dunbar T, Chen P, Godmilow L, Ganguly T. 2003. Exon skipping caused by an intronic insertion of a young AluYb9 element leads to severe hemophilia A. *Hum Genet* **113**: 348-352.

Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genet* **46**: 818-825.

Gilbert N, Lutz S, Morrish TA, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**: 7780-7795.

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet* **47**: 435-444.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet* **20**: 3386-3400.

Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. 2014. Somatic retrotransposition in human cancer revealed by whole-genome and whole exome sequencing. *Genome Res* **24**: 1053-1063.

Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T, et al. 2010. Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**: 1171-1182.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253-1261.

Janicic N, Pausova Z, Cole DE, Hendy GN. 1995. Insertion of an Alu sequence in the Ca(2+)-sensing receptor gene in familial hypocaliuric hypercalcemia and neonatal severe hyperparathyroidism. *Am J Hum Genet* **56**: 880-886.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-496.

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

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389-390.

Kimberland ML, Kivoky V, Prchal J, Schwahn U, Berger W, Kazazian HH Jr. 1999. Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum Mol Genet* **8**: 1557-1560.

Konkel MK, Walker JA, Hotard AB, Ranck MC, Fontenot CC, Storer J, Stewart C, Marth GT, Batzer MA. 2015. Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. *Genome Biol Evol* **7**: 2608-2622.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lanikova L, Kucerova J, Indrak K, Divoka M, Issa JP, Papayannopoulou T, Prchal JT, Divoky V. 2013. B-Thalassemia due to intronic LINE-1 insertion in the B-globin gene (HBB): molecular mechanisms underlying reduced transcript levels of the B-globin (L1) allele. *Hum Mutat* **34**: 1361-1365.

Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ III, Lohr JG, Harris CC, Ding L, Wilson RK, et al. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**: 967-971.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS ONE* **9**: e90581.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li X, Scaringe W, Hill KA, Roberts S, Mengos A, Careri D, Pinto MT, Kasper CK, Sommers SS. 2001. Frequency of recent retrotransposition events in the human factor IX gene. *Hum Mutat* **17**: 511-519.

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

Macfarlane CM, Collier P, Rahbari R, Beck CR, Wagstaff JF, Igoe S, Moran JV, Badge RM. 2013. Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. *Hum Mutat* **34**: 974-985.

Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. 1992. Disruption of the *APC* gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**: 643-645.

Miki Y, Katagiri T, Kasumi F, Yoshimot T, Nakamura Y. 1996. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat Genet* **13**: 245-247.

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

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

Moran JV, DeBerardinis RJ, Kazazian HH, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.

Mukherjee S, Mukhopadhyay A, Banerjee D, Chandak GR, Ray K. 2004. Molecular pathology of haemophilia B, identification of five novel mutations including a LINE 1 insertion in Indian patients. *Haemophilia* **10**: 259-263.

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903-910.

Nakamura Y, Murata M, Takagi Y, Kozuka Y, Hasebe R, Takagi A, Kitazawa J, Shima M, Kojima T. 2015. SVA retrotransposition in exon 6 of the coagulation factor IX gene causing severe hemophilia B. *Int J Hematol* **102**: 134-139.

Narita N, Nisho H, Kitoh Y, Ishikawa Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M. 1993. Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest* **91**: 1862-1867.

Ostertag EA, Kazazian HH Jr. 2001. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**: 2059-2065.

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. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics* **47**: 11.12.11-11.12.34.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Lower J, Stratling WH, Lower 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.

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**: 636-639.

Rodic N, Steranka JP, Makohon-Moore A, Moyer A, Shen P, Sharma R, Kohutek ZA, Huang CR, Ahn D, Mita P, et al. 2015. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med* **21**: 1060-1064.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**: 354-357.

Schwahn U, Lenzner S, Dong J, Feil S, Hinzmann B, van Duijnhoven G, Kirschner R, Hemberger M, Bergen AA, Rosenberg T, et al. 1998. Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nat Genet* **19**: 327-332.

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* **26**, 745-755.

Seleme MC, Vetter MR, Cordaux R, Bastone L, Batzer MA, Kazazian HH, Jr. 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 6611-6616.

Shukla R, Upton KR, Munoz-Lopez M, Gearhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**: 101-111.

Smit AFA, Green P. 1996-2010. RepeatMasker Open-3.0.

Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B et al. 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**: 2328-2338.

Stewart C, Kural D, Stromberg MP, Walker JA, Konkel MK, Stutz AM, Urban AE, Grubert F, Lam HY, Lee WP, et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**: e1002236.

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

Sukarova E, Dimovski AJ, Tchacarova P, Petkov GH, Efremov GD. 2001. An Alu insert as the cause of a severe form of hemophilia A. *Acta Haematol.* **106**: 126-129.

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3**: research0052.1-research0052.18.

Telenti A, Pierce LC, Biggs WH, di lulio J, Wong EHM, Fabiani MM, Kirkness EF, Moustafa A, Shah N, Xie C, et al. 2016. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci USA* **113**: 11,901-11,906.

Teugels E, De Brakeleer S, Goelen G, Lissens W, Sermijn E, De Greve J. 2005. *De novo* Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. *Hum Mutat* **26**: 284.

The 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-73.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.

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

The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82-90.

Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Dwa JY 2014. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol* **15**: 488.

Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, et al. 2014. Mobile DNA in cancer: Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**: 1251343.

Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sanchez-Luque FJ, Bodea GO, Ewing A, Salvador-Palomeque C, van der Knapp MS, Brennan PM, et al. 2015. Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**: 228-239.

Van de Water N, Williams R, Ockelford P, Browett P. 1998. A 20.7 kb deletion within the factor VIII gene associated with LINE-1 element insertion. *Thromb Haemost* **79**: 938-942.

Vattathil S, Akey, JM. 2015. Small amounts of archaic admixture provide big insights into human history. *Cell* **163**: 281-284.

Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. 1993. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *Eur J Hum Genet* **1**: 30-36.

Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. 1991. A *de novo* Alu insertion results in neurofibromatosis type 1. *Nature* **353**: 864-866.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354**: 994-1007.

Watanabe M, Kobayashi K, Jin F, Park KS, 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. *Am J Med Genet* **138**: 344-348.

Wu J, Lee WP, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. 2014. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* doi: 10.1186/1471.

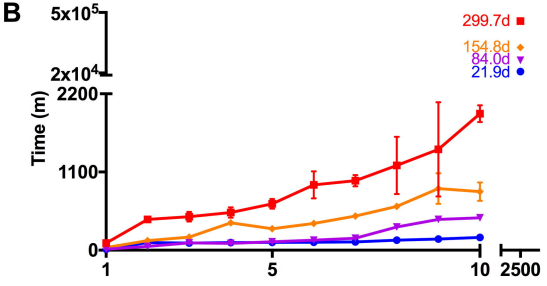
Wulff K, Gazda H, Schroder W, Robicka-Milewska R, Herrmann FH. 2000. Identification of a novel large F9 gene mutation—an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. *Hum Mutat* **15**: 299.

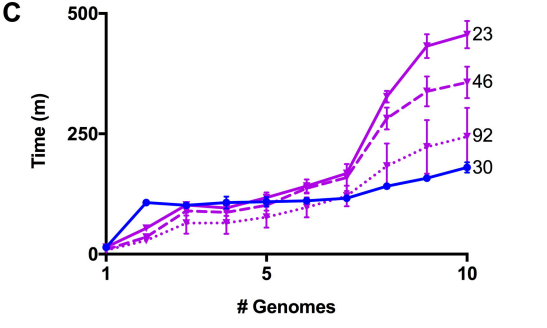
Zhuang J, Wang J, Theurkauf W, Weng Z. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* **42**: 6826-6838.

Figure 1.

A

Individual Genome Runtime (m)		
	6x	30x
MELT	10.7±0.15	93.3±1.13
TEMP	15.0±0.21	149.0±12.68
RetroSeq	41.9±0.25	445.1±1.17
Tangram	63.6±0.05	3164.3±0.94
Mobster	27.9±0.14	DNF





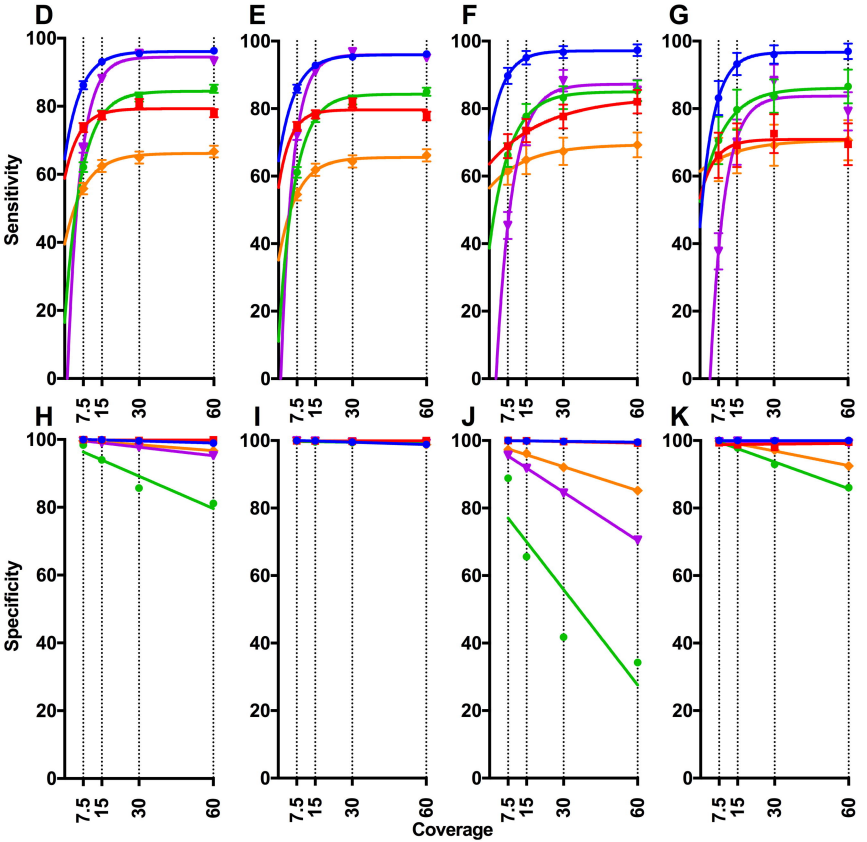


Figure 2.

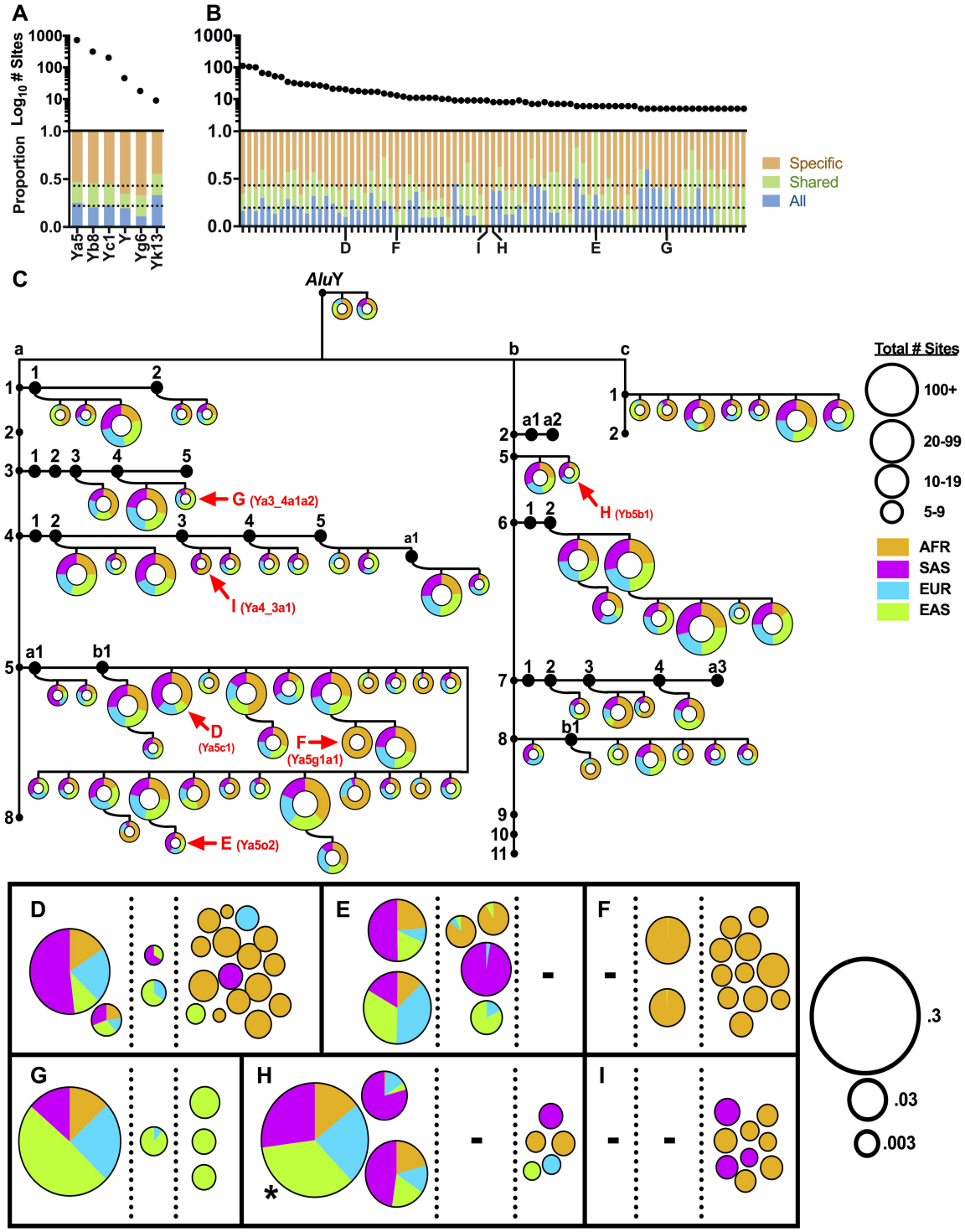

A
B
Log10 # Sites
Proportion
1000
100
10
1.0
0.5
0.0
Ya5
Yb8
Yc1
Y
Yg6
Yk13
D
F
I
H
E
G
Specific
Shared
All
C
AluY
a
b
c
G (Ya3_4a1a2)
I (Ya4_3a1)
D (Ya5c1)
F (Ya5g1a1)
E (Ya5o2)
H (Yb5b1)
Total # Sites
100+
20-99
10-19
5-9
AFR
SAS
EUR
EAS
.3
.03
.003

Figure 3.

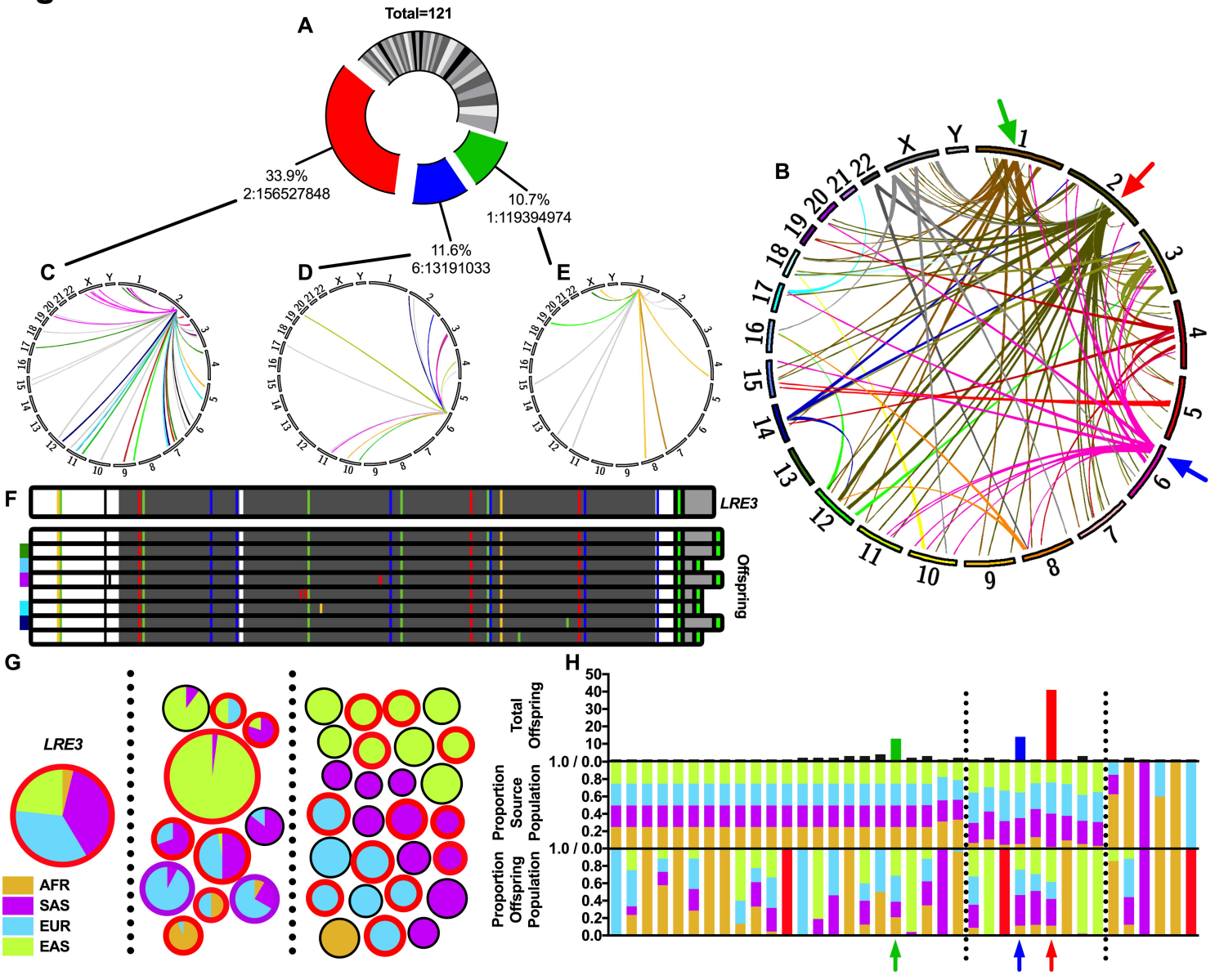

A
Total=121
33.9%
2:156527848
11.6%
6:13191033
10.7%
1:119394974
B
C
D
E
F
LRE3
Offspring
G
LRE3
AFR
SAS
EUR
EAS
H
Total Offspring
Proportion Source Population
Proportion Offspring Population
1.0 / 0.0

Figure 4.

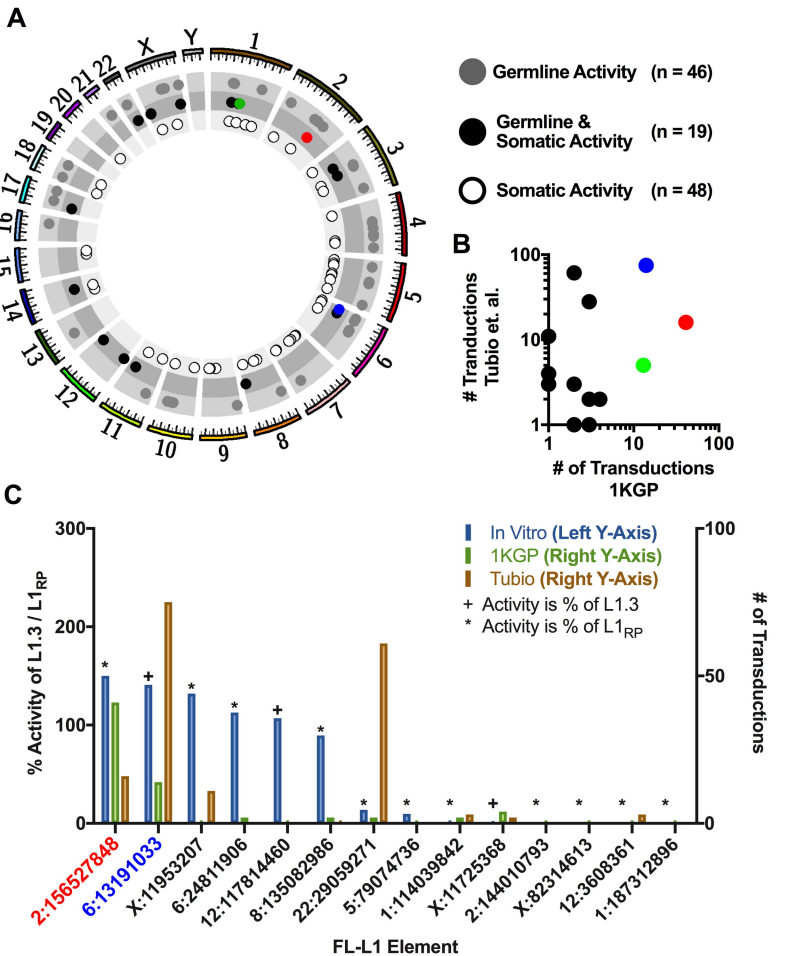

A
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Germline Activity (n = 46)
Germline & Somatic Activity (n = 19)
Somatic Activity (n = 48)
B
Tranductions Tubio et. al.
100
10
1
1
10
100
# of Transductions 1KGP
C
In Vitro (Left Y-Axis)
1KGP (Right Y-Axis)
Tubio (Right Y-Axis)
+ Activity is % of L1.3
* Activity is % of L1RP
% Activity of L1.3 / L1RP
# of Transductions
300
200
100
0
100
50
0
2:156527848
6:13191033
X:11953207
6:24811906
12:117814460
8:135082986
22:29059271
5:79074736
1:114039842
X:11725368
2:144010793
X:82314613
12:3608361
1:187312896
FL-L1 Element

Figure 5.

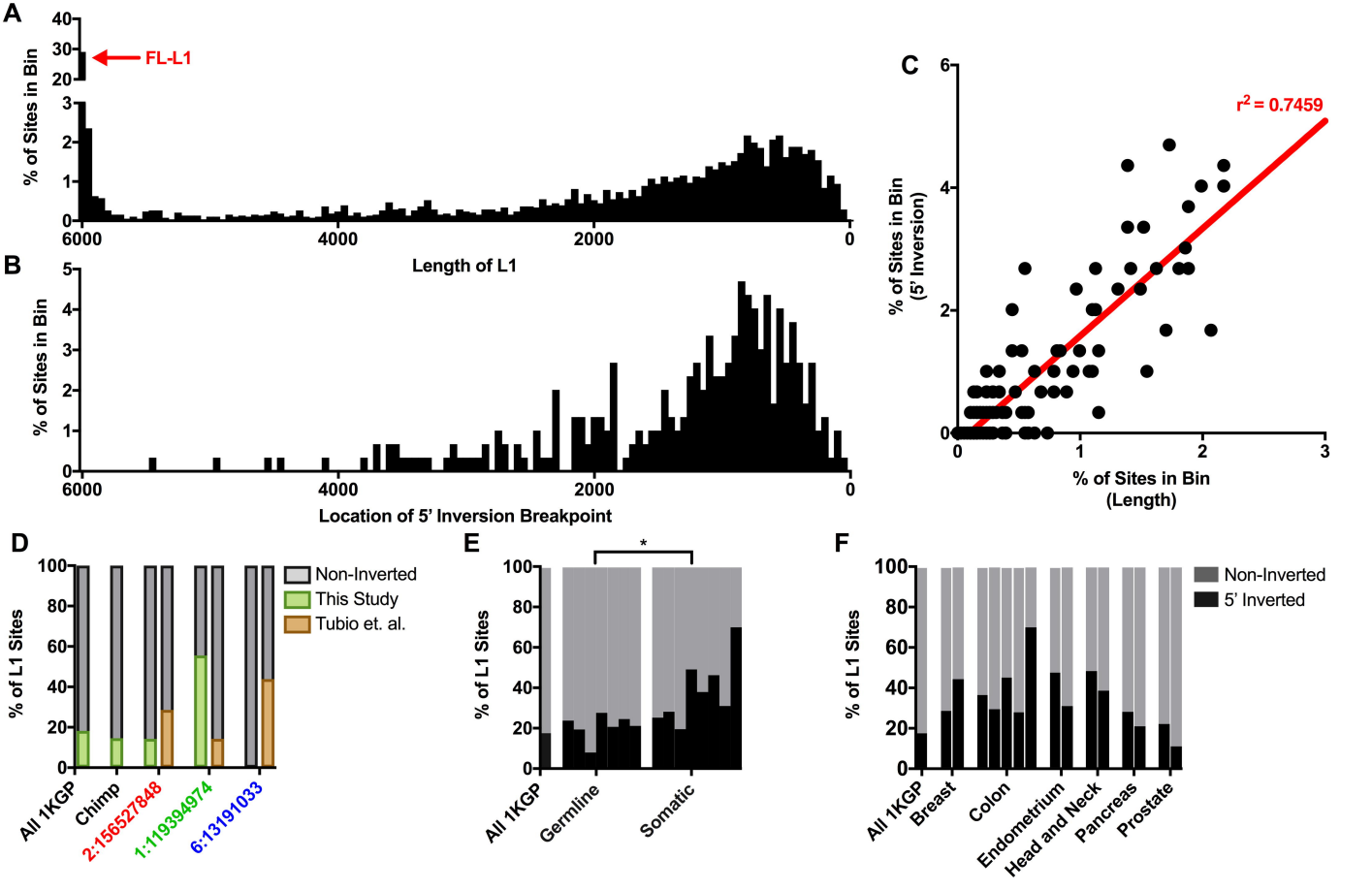
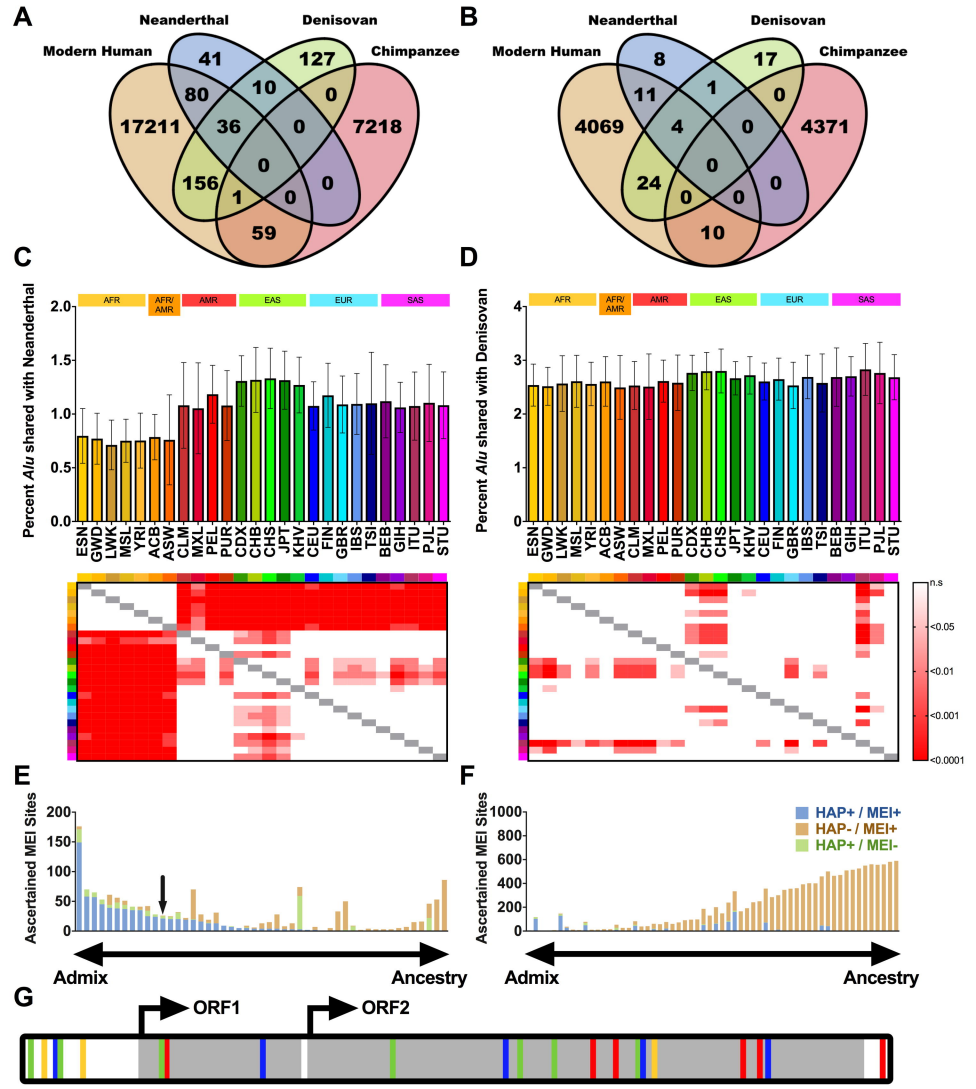

A
% of Sites in Bin
FL-L1
Length of L1
B
% of Sites in Bin
Location of 5' Inversion Breakpoint
C
r2 = 0.7459
% of Sites in Bin (5' Inversion)
% of Sites in Bin (Length)
D
% of L1 Sites
Non-Inverted
This Study
Tubio et. al.
All 1KGP
Chimp
2:156527848
1:119394974
6:13191033
E
% of L1 Sites
*
All 1KGP
Germline
Somatic
F
% of L1 Sites
Non-Inverted
5' Inverted
All 1KGP
Breast
Colon
Endometrium
Head and Neck
Pancreas
Prostate

Figure 6.


A
Neanderthal
Denisovan
Modern Human
Chimpanzee
41
127
80
10
0
17211
36
0
7218
0
156
1
0
59
B
Neanderthal
Denisovan
Modern Human
Chimpanzee
8
17
11
1
0
4069
4
0
4371
24
0
0
0
10
C
Percent Alu shared with Neanderthal
AFR
AFR/ AMR
AMR
EAS
EUR
SAS
ESN GWD LWK MSL YRI ACB ASW CLM MXL PEL PUR CDX CHB CHS JPT KHV CEU FIN GBR IBS TSI BEB GIH ITU PJL STU
D
Percent Alu shared with Denisovan
n.s
<0.05
<0.01
<0.001
<0.0001
E
Ascertained MEI Sites
Admix
Ancestry
F
Ascertained MEI Sites
HAP+ / MEI+
HAP- / MEI+
HAP+ / MEI-
Admix
Ancestry
G
ORF1
ORF2