Supplemental Materials

Mutagenesis of human genomes by endogenous mobile elements on a population scale.

Nelson T. Chuang, Eugene J. Gardner, Diane M. Terry, Jonathan Crabtree, Anup A. Mahurkar, Guillermo L. Rivell, Charles C. Hong, James A. Perry, and Scott E. Devine.

Table of Contents **Page**

Supplemental Methods

Additional details on improvements to MELT code.

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

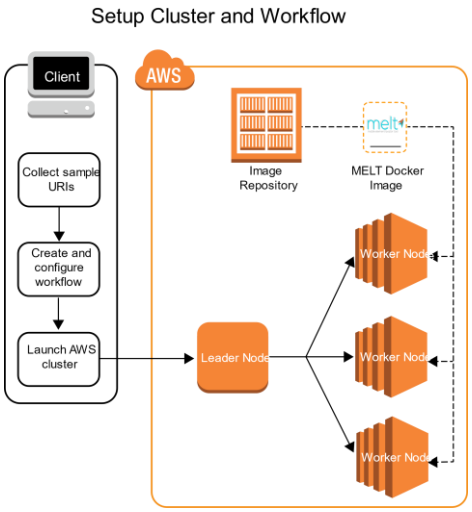
Accuracy of long PCR and PacBio sequencing of FL-L1Hs elements.

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

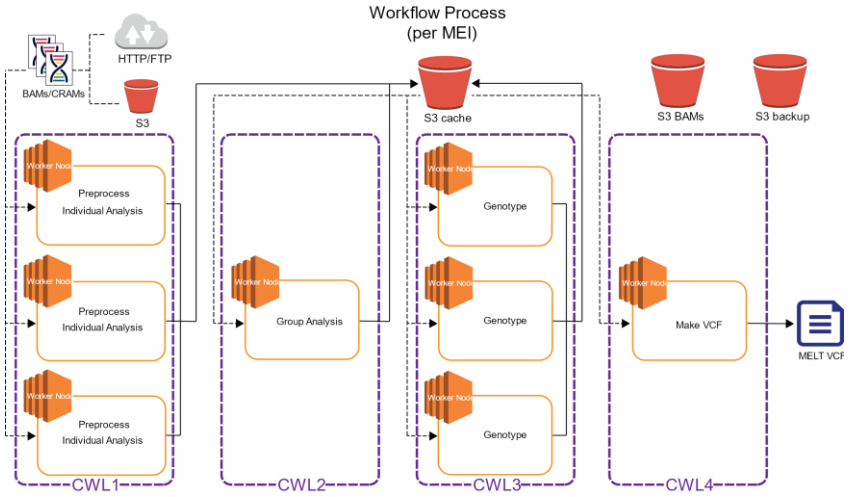
Supplemental Figures

Supplemental_Fig_S1.

A

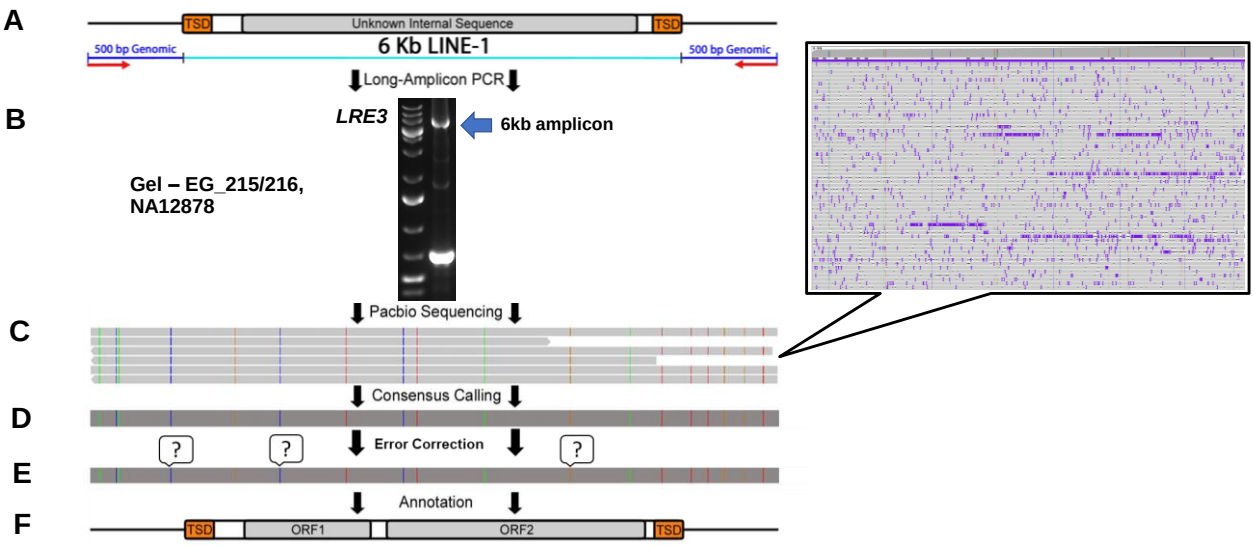


B



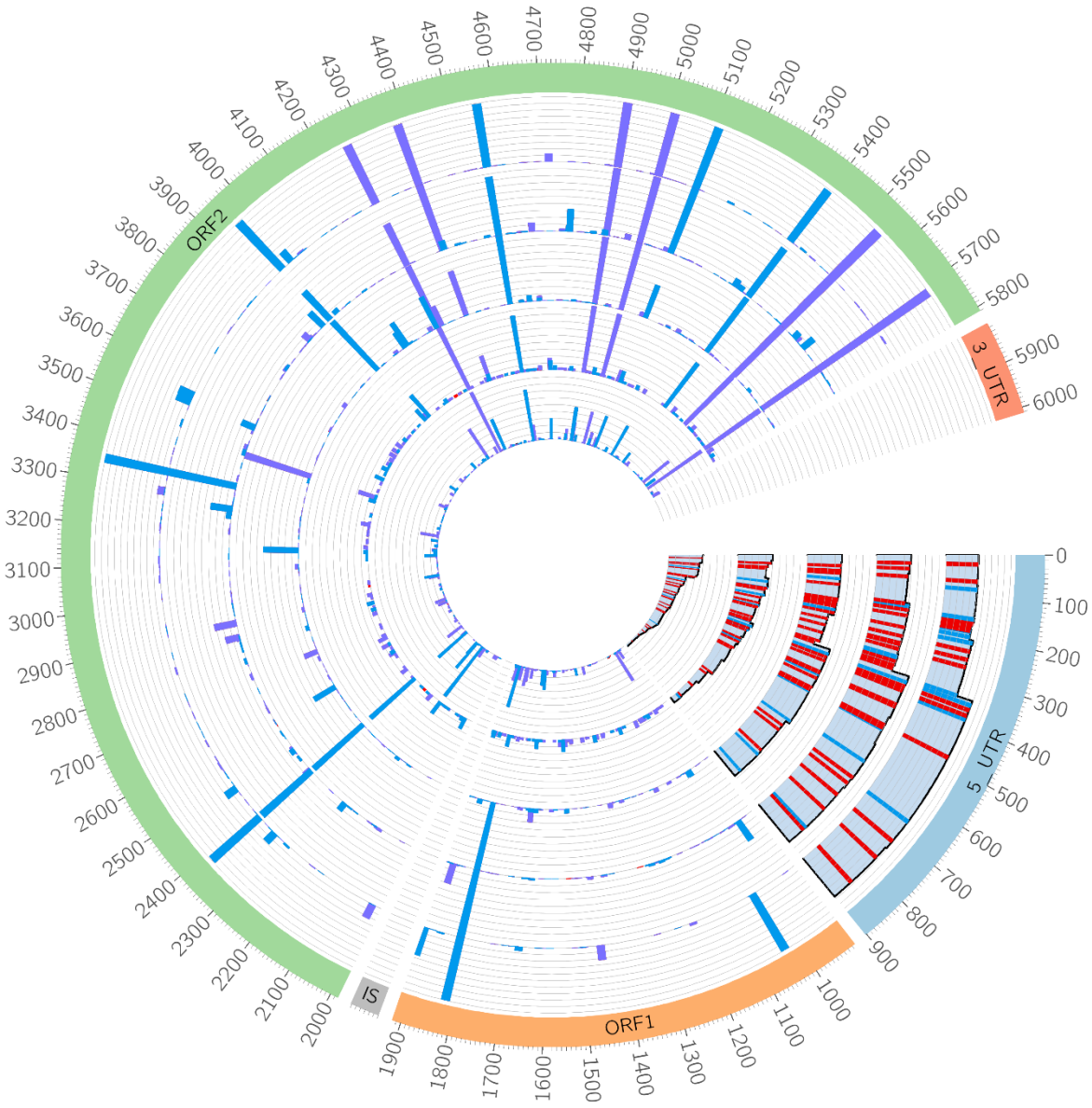
Supplemental_Fig_S1. CloudMELT workflow. A. In order to launch and run CloudMELT on Amazon Web Services (AWS), initial parameters need to be configured. These parameters are used by Toil to communicate with AWS for the desired number of compute nodes. **B.** Each vCPU in a compute node will run in parallel the standard MELT Preprocess and Individual Analysis step. The results are then aggregated to perform the Group Analysis step. Once that is completed, the results are then used in the Genotype step for each genome. Finally, the results from the Genotype step are gathered to create the final VCF file. The intermediate files are saved separately on S3 for debugging or to re-run steps that fail.

Supplemental_Fig_S2.



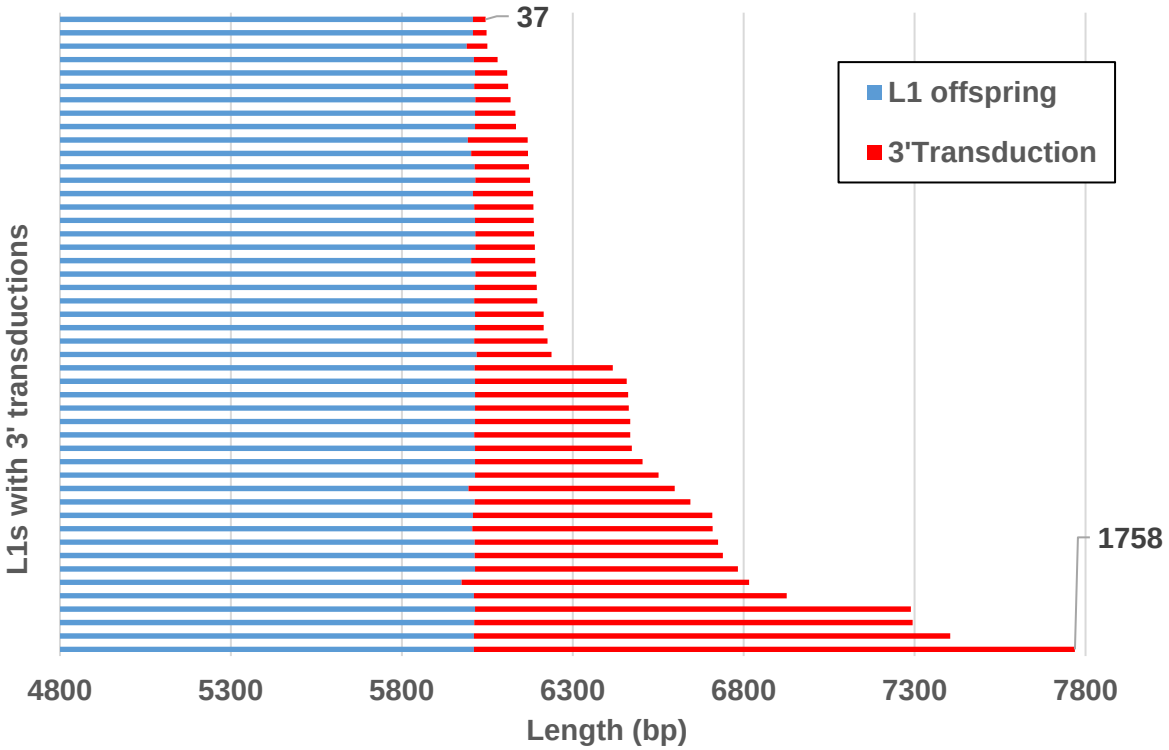
Supplemental_Fig_S2. PacBio amplification and sequencing of FL-L1Hs elements. A. Long-range PCR primers are designed to amplify each target FL-L1Hs element (red arrows). **B.** The initial long PCR for the *LRE3* FL-L1Hs element is examined on an agarose gel to determine whether a 6 kb FL-L1Hs fragment was successfully amplified (blue arrow). **C.** Several dozen successful amplicons from unrelated sites are then pooled and the 6 kb FL-L1Hs fragments are eluted from an agarose gel to eliminate the unoccupied allele. The eluted 6 kb pools are then subjected to PacBio sequencing. A sampling of the actual PacBio reads for *LRE3* is depicted in the upper right. **D.** The long reads for each FL-L1Hs site are identified using the unique flanking genomic sequences, aligned, and the consensus sequences are determined with the SMRT Analysis v2.3.0 software. Vertical colored lines indicate internal sequence changes within a given FL-L1Hs element compared to the reference FL-L1Hs element sequence. **E.** Systematic error correction is performed on the consensus sequences (see Methods), and **F.** the final sequence is annotated with a custom analysis pipeline (Supplemental_Code_S2.tar).

Supplemental_Fig_S3.



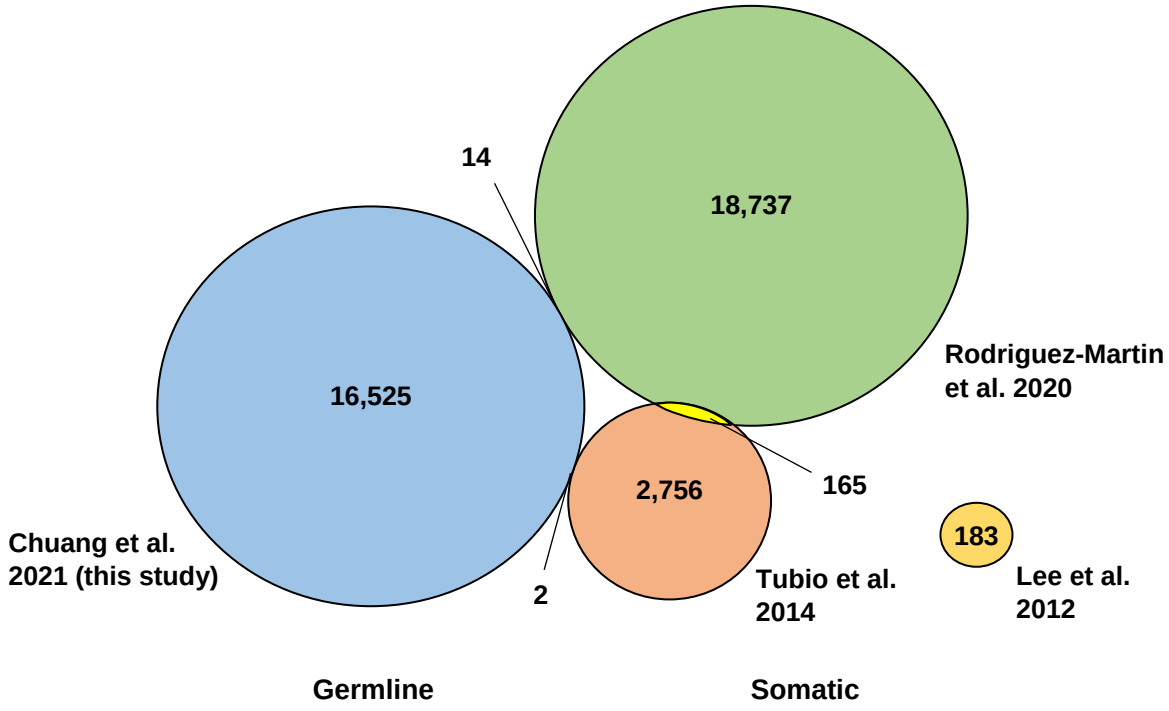
Supplemental_Fig_S3. Circos plot demonstrating the frequency of variation within our collection of sequenced FL-L1Hs elements arranged by subfamily. All subfamilies displayed were aligned to the PreTa subfamily consensus as the reference. The inner track is the oldest subfamily after PreTa, Ta0, and the youngest is Ta1d-TCA on the outermost track. The 5' UTR shows the frequency of CpG changes from the reference number of CpGs found in the PreTa consensus sequence. The blue bars indicate a position where there is a CpG gain, whereas the red bars indicate a position where there is a CpG loss. ORF1 and ORF2 changes are shown as blue bars for synonymous mutations and purple bars for nonsynonymous mutations. Red bars are nonsense mutations.

Supplemental_Fig_S4.



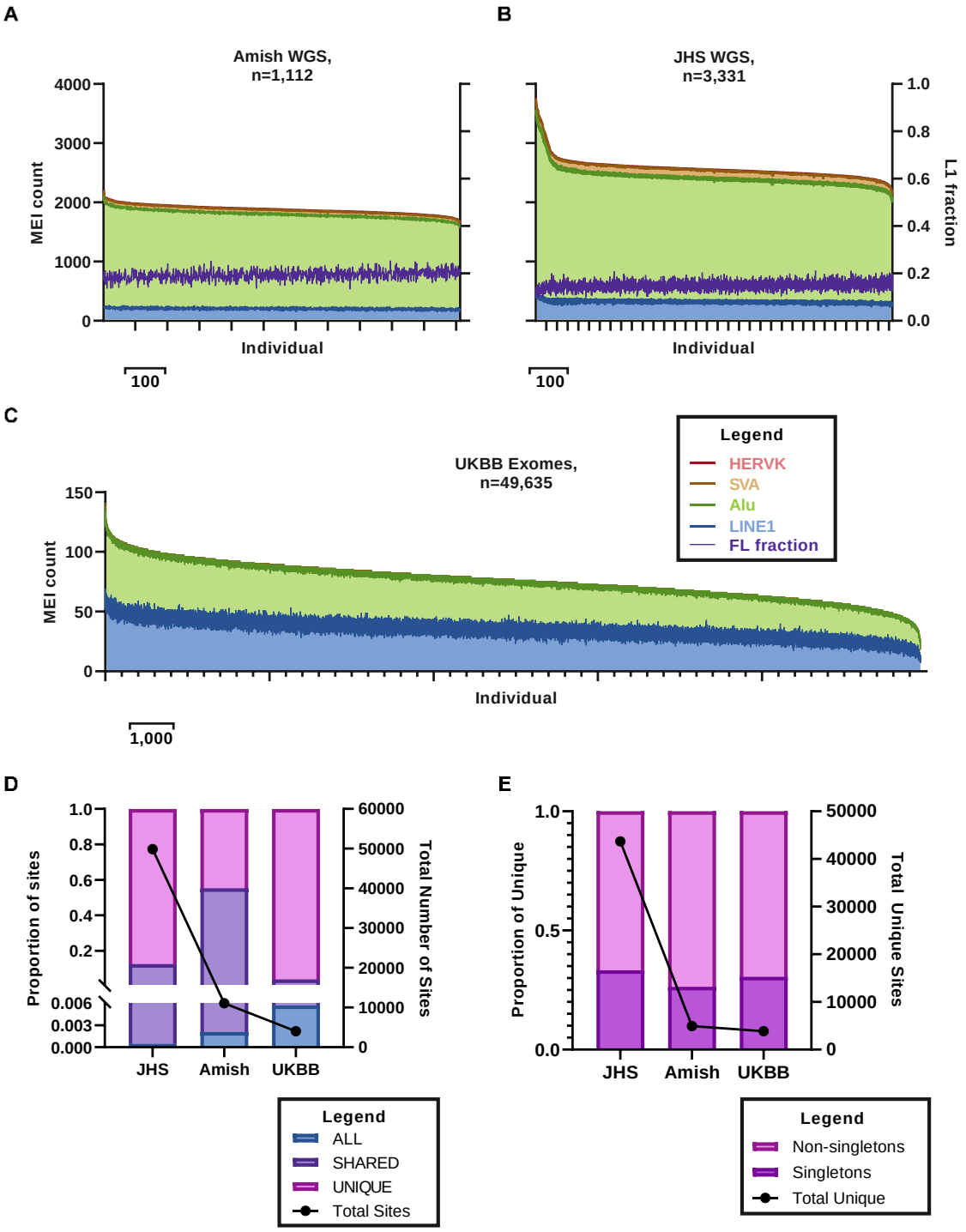
Supplemental_Fig_S4. 3' transductions associated with FL-L1Hs elements. The 3' ends of the 49 FL-L1Hs elements that are associated with 3' transductions from our PacBio sequenced elements are depicted (blue) along with the 3' transduced sequences (red). The full 3' transduced sequences were determined along with the adjacent FL-L1Hs elements. Transductions ranged from 37 bp to 1758 bp in length. A total of 76 of our sequenced FL-L1Hs elements have produced MEIs with 3' transductions (Supplemental_Table_S3A.xlsx).

Supplemental_Fig_S5.



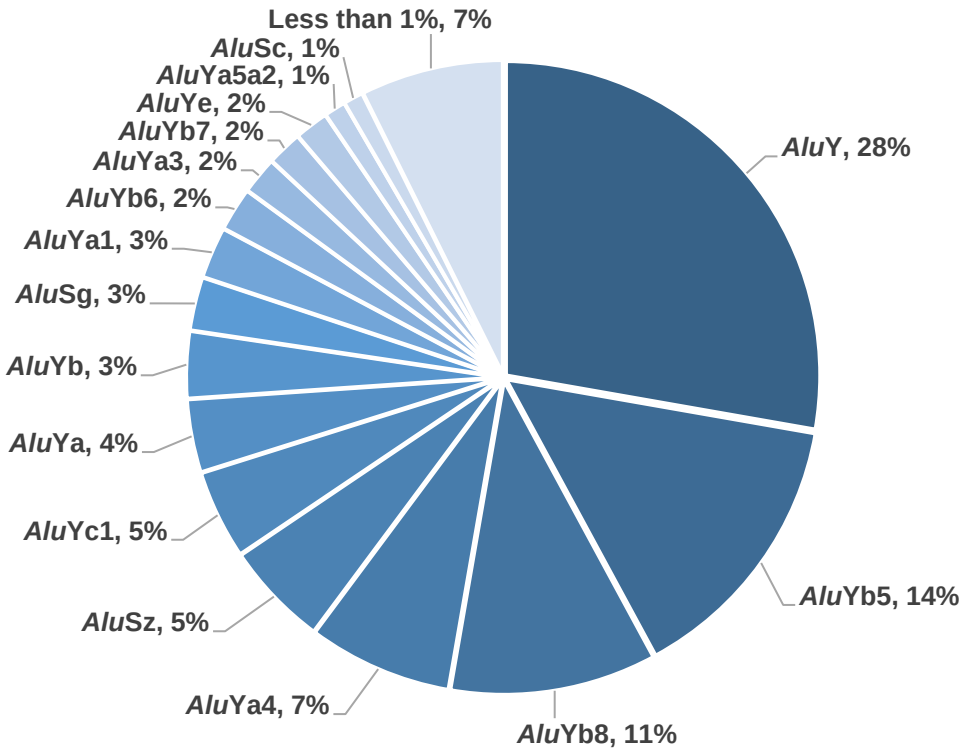
Supplemental_Fig_S5. Comparisons of germline and somatic L1 MEIs. The human germline L1 MEIs that we identified in this study (blue circle) were compared with three studies that discovered somatic L1 insertions in human cancers (Lee et al. 2012—orange circle; Tubio et al. 2014—light red circle; and Rodriguez-Martin et al. 2020—green circle). Note that we found minimal overlap between the germline and somatic studies (only two MEIs overlap between our study and Tubio et al. and 14 MEIs overlap between our study and Rodriguez-Martin et al.). These could be false positive MEIs in the somatic studies (i.e., not found in the normal tissue, but actually present). Alternatively, they could represent MEIs at the same site that occurred independently in the germline and somatic tissues.

Supplemental_Fig_S6.



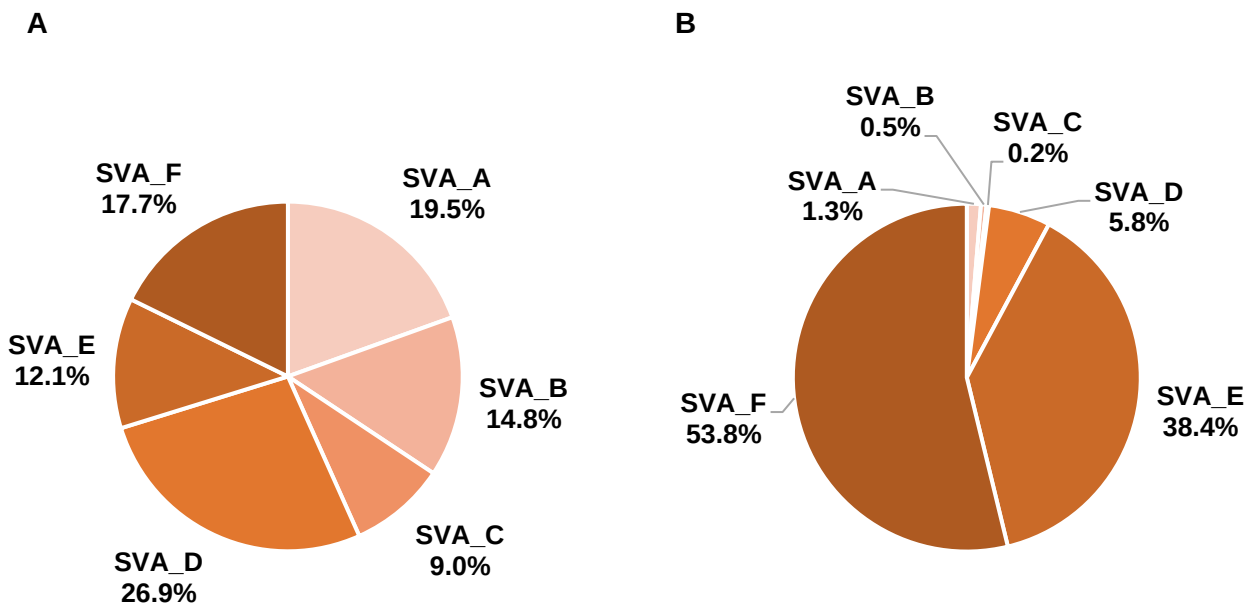
Supplemental_Fig_S6. MEI counts in TOPMed and UKBB individuals. A-C. MEI counts (L1=light blue; *Alu*=light green; SVA=light brown; HERV-K=light red) are plotted for each individual in the three populations. The dark lines of the same colors indicate the boundaries of each MEI type. The dark purple lines indicate the fraction of non-REF FL-L1Hs elements (plotted for whole genome data only). Note that the Jackson Heart Study (**B**), which is African Americans, has increased MEI counts compared to the Amish population (**A**). These results are consistent with a European bottleneck associated with the Amish population (Pollin et al. 2008) compared to the African American Jackson Heart population. Note also that the UKBB individuals (**C**) have fewer MEIs compared to the TOPMed Jackson Heart and Amish populations, consistent with the fact that only the exome portions of the genome were analyzed in the UKBB individuals. **D, E.** Sharing analysis across the three populations. Note that only 23 MEIs were shared by all three populations (**D**, lower right blue area). In contrast to the 1KGP populations, where MEI discovery was performed in ~100 individuals from each of the 26 diverse populations, MEI discovery was performed in 1,112 Amish genomes, 3331 JHS genomes, and 49,635 UKBB exomes. Thus, we discovered more unique variation in the Amish, JHS, and UKBB samples.

Supplemental_Fig_S7.



Supplemental_Fig_S7. Subfamily analysis of *Alu* MEIs. *Alu* subfamilies were annotated using the CAlu package of MELT (Gardner et al. 2017). Note that the majority of subfamilies that we identified are young *Alu*Y subfamilies that are known to be polymorphic and active in humans (e.g., 81% of the elements belong to *Alu*Y and the *Alu*Ya, *Alu*Yb, and *Alu*Yc lineages similar to what has been observed in previous studies; Gardner et al. 2017). Nine percent of the elements belong to three *Alu*S subfamilies (*Alu*Sz, *Alu*Sg, *Alu*Sc), which are known to be polymorphic in humans as well (Mills et al. 2007).

Supplemental_Fig_S8.



Supplemental_Fig_S8. Subfamily analysis of SVA MEIs. SVA subfamilies were annotated using the approach outlined in Wang et al. 2005 using 3' SINE sequences. **A.** Classification of SVA subfamilies in the build GRCh38 REF genome and **B.** the non-REF SVAs identified in this study. **B.** Note that the youngest SVA_E and SVA_F elements have undergone a recent expansion from 29.8% in the REF genome (**A**) to 92.2% of the non-REF SVAs (**B**).

Supplemental Discussion

Because *Alu* elements are relatively short (~280 bp), we often recover the full interior sequences of these elements and this allows us to classify *Alu*s according to their known subfamilies (Supplemental_Fig_S7). As outlined in Figure 2, we had to resort to long PCR and sequencing to recover the interior sequences of FL-L1Hs elements because we only discover a limited amount of interior sequence information near the insertion junctions when using MELT for MEI discovery. SVA and HERV-K elements also can be up to several kb in length, and we likewise recover only a limited amount of interior sequences near the junctions. However, we were able to leverage internal sequences that we recovered with MELT near the 3' end of the SVA insertions to determine the subfamily status for these elements (Supplemental_Fig_S8).

In principle, the SVA and HERV-K sites that we have identified might serve as a foundation for future studies that also use long PCR and sequencing to more fully recover the interior sequences of these elements. Such studies may allow us to identify additional SVA and HERV-K subfamilies and other useful internal sequence variation. In contrast to the three active classes of mobile elements in humans (i.e., *Alu*, L1, and SVA elements), HERV-K elements are thought to be extinct, as no intact, active copies of HERV-K have been identified in the human genome (Dewannieux et al. 2006). Thus, it is unlikely that this would lead to the identification of any functionally intact HERV-K copies. However, our HERV-K (and SVA) sites could be used to examine genes located near these sites to determine whether these polymorphic MEIs impact the expression or regulation of nearby genes.

Supplemental References

Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T. 2006. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res* **16**: 1548-1556.

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.

Gardner EJ, Lam VK, Harris D, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 Genomes Project Consortium, Devine SE. 2017. The Mobile Element Locator Tool (MELT): Population-scale MEI discovery and biology. *Genome Res* **27**: 1916-1929.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23**: 183-189.

Pollin TI, McBride DJ, Agarwala R, Schäffer, AA, Shuldiner AR, Mitchell BD, O'Connell JR. 2008. Investigations of the Y Chromosome, Male Founder Structure and YSTR Mutation Rates in the Old Order Amish. *Hum Hered* **65**: 91–104.

Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, Santamarina M, Ju YS, Temes J, Garcia-Souto D, et al. 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nature Genetics* **52**: 306-319.

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.

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.