Multi-ancestry GWAS reveals loci linked to human variation in LINE-1- and Alu-copy numbers

Juan I. Bravo[a], Lucia Zhang[a,b], Bérénice A. Benayoun[a,c,d,e,f,*]

[a] Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA 90089, USA.

[b] Quantitative and Computational Biology Department, USC Dornsife College of Letters, Arts and Sciences, Los Angeles, California, USA.

[c] Molecular and Computational Biology Department, USC Dornsife College of Letters, Arts and Sciences, Los Angeles, CA 90089, USA.

[d] Biochemistry and Molecular Medicine Department, USC Keck School of Medicine, Los Angeles, CA 90089, USA.

[e] USC Norris Comprehensive Cancer Center, Epigenetics and Gene Regulation, Los Angeles, CA 90089, USA.

[f] USC Stem Cell Initiative, Los Angeles, CA 90089, USA.

[*] Corresponding author. Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA, 90089, USA.

E-mail address: berenice.benayoun@usc.edu (B.A. Benayoun)

ABSTRACT

Long INterspersed Element-1 (LINE-1; L1) and Alu are two families of transposable elements (TEs) occupying ~17% and ~11% of the human genome, respectively. Though only a small fraction of L1 copies is able to produce the machinery to mobilize autonomously, Alu elements and degenerate L1 copies can hijack their functional machinery and mobilize *in trans*. The expression and subsequent copy number expansion of L1 and Alu can exert pathological effects on their hosts, promoting genome instability, inflammation, and cell cycle alterations. These features have made L1 and Alu promising focus subjects in studies of aging and aging diseases where they can become active. However, the mechanisms regulating variation in their expression and copy number remain incompletely characterized. Moreover, the relevance of known mechanisms to diverse human populations remains unclear, as mechanisms are often characterized in isogenic cell culture models. To address these gaps, we leveraged genomic data from the 1000 Genomes Project to carry out a trans-ethnic GWAS of L1 and Alu insertion global singletons. These singletons are rare insertions observed only once in a population, potentially reflecting recently acquired L1 and Alu integrants or structural variants, and which we used as proxies for L1/Alu-associated copy number variation. Our computational approach identified single nucleotide variants in genomic regions containing genes with potential and known TE regulatory properties, and it enriched for single nucleotide variants in regions containing known regulators of L1 expression. Moreover, we identified many reference TE copies and polymorphic structural variants that were associated with L1/Alu singletons, suggesting their potential contribution to TE copy number variation through transposition-dependent or transposition-independent mechanisms. Finally, a transcriptional analysis of lymphoblastoid cells highlighted potential cell cycle alterations in a subset of samples harboring L1/Alu singletons. Collectively, our results (i) suggest that known TE regulatory mechanisms may also play regulatory roles in diverse human populations, (ii) expand the list of genic and repetitive genomic loci implicated in TE copy number variation, and (iii) reinforce the links between TEs and disease.

KEYWORDS: LINE-1, Alu, transposons, GWAS, regulators, copy number

1. INTRODUCTION

In the human genome, the two most abundant families of transposable elements (TEs) are Long INterspersed Element-1 (LINE-1; L1) and Alu, which account for ~16–17% and ~9–11% of the genome, respectively [1, 2]. Full-length L1 elements span ~6 kilobases and produce bicistronic messenger ribonucleic acids (mRNAs) encoding two polypeptides, ORF1p and ORF2p, necessary for L1 transposition (reviewed in [3]). The L1 family can be segregated into 3 subfamilies depending on the evolutionary age of the copy: the L1M (mammalian-wide) lineage is the oldest, the L1P (primate-specific) lineage is of intermediate age, and the L1PA lineage is the youngest. Importantly, only the L1PA1/L1Hs subfamily contains ~80-100 actively mobile copies in the average human genome [4], with the remaining ~500,000 L1 copies being rendered non-autonomous due to the presence of loss-of-function mutations or truncations [1]. In contrast to L1 elements, Alu elements are short (~300 bp) non-autonomous retrotransposons that rely on functional L1 machinery for their mobilization [5-7]. Alu retrotransposons can also be segregated by evolutionary age into the following subfamilies: AluJ is the oldest lineage and is likely completely inactive in humans, AluS is the middle-aged lineage and contains mobile copies, and AluY is the youngest lineage and contains the largest number of functionally intact elements [8].

For a transposition-dependent expansion of L1 copy number to occur, L1 must undergo a multi-step lifecycle. This lifecycle includes (i) transcription of an active, full-length L1 copy, (ii) potential RNA processing of the L1 transcript, (iii) nuclear export of the transcript to the cytoplasm, (iv) translation of the two open reading frames (ORFs), (v) potential post-translational modifications of ORF1p/ORF2p and binding of those proteins to the transcripts that produced them (*cis* preference) to form ribonucleoprotein (RNP) complexes, (vi) entry into the nucleus, and finally (vii) reverse transcription and integration by a process called target primed reverse transcription (TPRT) (reviewed in [3]). Importantly, though neither Alu elements or degenerate L1 copies can mobilize autonomously, they can hijack proteins from transposition-competent L1s and mobilize *in trans* [6, 7, 9]. Though not traditionally considered part of the L1/Alu lifecycles, other transposition-independent, but homology-dependent, mechanisms can further

contribute to TE copy number variation. This includes repeat-mediated deletion (RMD) events whereby two repetitive elements (often Alu elements) on the same chromatid can recombine to cause a deletion rearrangement, potentially resulting in the deletion of one of the repeats as well as the intervening sequence, which may include additional repeats [10, 11]. More broadly, TE-mediated and TE-independent non-allelic homologous recombination (NAHR) events ([12-15] and reviewed in [16, 17]) can directly generate large chromosomal deletions and duplications, which may include repetitive sequences. From an evolutionarily perspective, the burst of Alu transposition in primates is hypothesized to have sensitized the ancestral genome to Alu-mediated recombination events, which may have played a role in the emergence and expansion of segmental duplications, which would provide additional substrates for NAHR [18]. Ultimately, TE copy number is shaped by a combination of *de novo* insertions resulting from their lifecycle and genomic structural remodeling that can expand or retract the copy number.

Characterizing the mechanisms governing L1 and Alu transcriptional and copy number control will be important, given their associations with, and potential contributions to, aging and aging-associated diseases like cancer (discussed in [19-21]). Fundamentally, L1 and/or Alu can alter several hallmarks of aging [22], such as genomic instability, cellular senescence, and inflammation. Though the origin of the signal is unclear (genomic, extra-chromosomal, or cytosolic), an increase in L1 copy number has been observed with chronological aging [23] and during cellular senescence [24]. Moreover, a key feature of cellular senescence is the senescence-associated secretory phenotype (SASP) whereby cells secrete an amalgamation of pro-inflammatory factors [25] that may contribute to chronic, low-grade, sterile inflammation with chronological age (a phenomenon referred to as “inflamm-aging”) [25, 26]. L1 can induce a senescent-like state *in vitro* in several cell lines [27, 28] and its cytoplasmic complementary DNA (cDNA) is implicated in the maturation of the SASP response and the establishment of deep senescence through the production of interferons [29]. Similarly, Alu RNA can upregulate senescence markers in retinal pigment epithelium (RPE) cells from human eyes with geographic atrophy [30], and knockdown of Alu

transcripts were reported to promote senescence exit in adult adipose-derived mesenchymal stem cells [31]. These findings highlight the relevance of L1 and Alu retrotransposons in pathological, age-associated features.

To maintain homeostasis, it is imperative that host cells tightly regulate TE activity (reviewed in [32, 33]). These mechanisms, however, remain incompletely characterized due to the cell-specific and transposon-specific nature of TE regulatory mechanisms. Indeed, no systematic, genome-wide screen for regulators of Alu expression or copy number has been carried out thus far, to our knowledge. To address these gaps, a number of *in vitro* and *in silico* approaches have been developed to scan for novel regulators of TE expression or copy number. *In vitro* approaches have relied on clustered regularly interspaced short palindromic repeats (CRISPR)-based and small RNA-based tools to decipher L1 regulation in several types of cancer cells [34-38]. These approaches, however, can be technically challenging to implement in non-cancerous cells, like primary cells, which may not tolerate hyper-elevated transposon activity or that may resist genetic perturbations. To complement these methods, a number of *in silico* approaches have been developed that utilize chromatin immunoprecipitation followed by sequencing (ChIP-seq) [39], gene co-expression networks [40], or copy number-expression correlations [41] to explore L1 regulation without external manipulations. More recently, we screened for candidate regulators of L1 RNA levels in lymphoblastoid cell lines (LCLs) using trans-expression quantitative trait locus (trans-eQTL) analysis [42]. These tools highlight the need for, and usefulness of, alternative approaches that utilize increasingly available, large '-omic' datasets to identify potentially novel mechanisms of TE control.

In this study, we develop a genome-wide association study (GWAS) pipeline to identify genomic loci associated with global L1 and Alu insertion singletons in diverse human populations. Global singleton insertions are rare insertions observed only once in a population [43], potentially reflecting recently acquired L1 and Alu integrants or structural variants. Thus, we used insertion singletons as proxies for L1/Alu-associated copy number variation, which can arise through *de novo* transposition events or

alternative mechanisms. We demonstrate that our GWAS approach captures, and enriches, genomic regions containing known and potential regulators of TE activity. We observe that this approach also captures reference insertions and polymorphic structural variants that may influence L1 or Alu copy number variation through transposition-dependent or -independent mechanisms. Finally, we note that associated loci fall into a few genes with clinical relevance, strengthening the association between TEs and disease.

2. RESULTS

2.1 Identification of genomic loci associated with L1/Alu singletons in diverse human populations

To unbiasedly identify potential genetic sources of L1 and Alu copy number variation in human populations, we leveraged a publicly available human "omic" dataset with thoroughly characterized genetic information. For this analysis, we utilized 2503 multi-ethnic samples from the 1000 Genomes Project for which both single nucleotide variant (SNV) and structural variant (SV) data were available. Specifically, this included individuals from 5 super-populations: 660 African (AFR), 504 East Asian (EAS), 503 European (EUR), 489 South Asian (SAS), and 347 Admixed American (AMR) individuals who declared themselves healthy at the time of sample collection (**Figure 1A**). As a quality control step, we checked whether the combined SNV and SV data segregated samples by population following principal component analysis (PCA). These analyses demonstrated that the top four principal components segregated population groups within each super-population (**Figure S1A-S1E**).

We then carried out an integration of the available multi-ethnic SNV and SV genomic data (**Figure 1B**). For the phenotype, we focused on global singleton SVs, which are rare SVs that occur exactly once in a study [43], for L1 and Alu insertions. Given their rarity, these insertion singletons may reflect recently acquired L1 and Alu integrants or structural variants. Thus, we hypothesized that these global singletons could serve as proxies for elevated L1/Alu-associated copy number variation, which may arise from *de novo* transposition events or alternative mechanisms. Importantly, Alu is dependent on L1 machinery for its mobilization [6, 7] and both L1 and Alu can contribute to structural variation through recombination-based mechanisms [12, 14, 15], so variation in L1 and Alu copy numbers are likely to have overlapping regulatory mechanisms. Thus, we first split our samples into cases and controls, depending on whether or not they contained an L1 and/or an Alu insertion global singleton (**Table S1A**). Second, we carried out a GWAS within each super-population to identify common, polymorphic SNVs and SVs associated with case-control status. Third, to maximize statistical power and identify shared, trans-ethnic sources of TE singleton

number variation, we meta-analyzed our GWAS results across the 5 super-populations using a random effects statistical model [44, 45]. Interestingly, several hundred L1/Alu global singleton insertions were detected in each super-population, ranging from 322 in the American cohort to 866 in the African cohort (**Figure 1C**). Though most case samples had 1-3 L1/Alu global singletons, several samples exhibited much more extreme TE singleton accumulation, especially within the African super-population (**Figure 1C**). Finally, we note that L1/Alu global singleton insertions were distributed across all autosomes in all 5 super-populations (**Figure 1C**).

As expected, GWAS in each super-population was generally underpowered (**Figure S2A-S2E**). Though we were able to identify many significant (FDR < 0.05) variants in the African cohort (**Figure S2A**), we could not identify significant variants in the East Asian (**Figure S2B**) and European (**Figure S2C**) cohorts, and we were only able to identify a handful of significant variants in the South Asian (**Figure S2D**) and Admixed American (**Figure S2E**) cohorts. In contrast, the GWAS meta-analysis integrating all super-populations identified 658 significant variants distributed across all 22 autosomes, though there was especially strong and recurrent signal on chromosome 21 (**Figure 1D, Table S1B**). To simplify functional annotation of significant variants and discard potential false positives, we omitted from downstream analyses significant variants overlapping the "ENCODE blacklist v2" [46]. This curated "blacklist" represents a set of genomic regions with anomalous signal in next-generation sequencing experiments independent of cell type and individual experiment [46], and SNVs overlapping these regions were significantly enriched in our significant SNV list compared to the background SNV list (Fisher's exact test, FDR = 1.33E-302). After filtering 188 blacklisted SNVs, we were left with 449 greenlisted SNVs and 21 greenlisted SVs that were significantly associated with case-control status (**Figure 1D, Table S1B**).

To assess the potential functions of greenlisted, significant variants, nearby genes were assigned to variants using the Genomic Regions Enrichment of Annotations Tool (GREAT) [47, 48] and over-representation analysis (ORA) was carried out using

clusterProfiler [49] (**Figure S3A**). Except for 51 greenlisted SNVs that were not linked to any gene, the remaining SNVs were linked to 1-2 genes and were found predominantly within 500 kilobases of a transcriptional start site (TSS) (**Figure S3B**). This observation highlights the association of intergenic and gene-proximal, rather than distal, genetic variation with L1/Alu copy number differences. Over-representation analysis of the associated genes using the Gene Ontology (GO) Biological Process gene set revealed an enrichment of terms related to heart development (such as 'regulation of heart growth', 'cardiac chamber morphogenesis', and 'positive regulation of cardiac muscle cell proliferation') and neuronal function (such as 'neuron recognition' and 'axonal fasciculation'; **Figure S3B**, **Table S1C**). Interestingly, genes related to 'reproduction' were also significantly over-represented among our list of associated genes (**Table S1C**). Similar to the SNVs, greenlisted SVs were all linked to 1-2 genes and were mostly within 500 kilobases of an annotated TSS (**Figure S3C**). Likely due to the low number of greenlisted SVs, and consequently low number of associated genes, we were unable to identify any significantly enriched GO Biological Process gene sets (**Table S1D**). Given the limited number of greenlisted SVs and the unavailability of SV sequences, we largely focused on greenlisted SNVs in downstream enrichment analyses.

As a complementary approach to GREAT, we also predicted the functional impact of significant variants using SnpEff [50] (**Table S1E**). Most variants were assigned a 'modifier' impact by SnpEff—this annotation describes non-coding variants where definitive functional predictions are not straightforward. One exception to this included an SNV (rs367696690) introducing a synonymous substitution (p.Asp1605Asp) with a low predicted impact in *IGFN1* (immunoglobulin like and fibronectin type III domain containing 1). It is worth noting that synonymous substitutions can still impact mRNA levels and protein function because of host preferences for specific codons (i.e. codon usage bias; reviewed in [51]). Another exception was a missense variant (rs1406034099) introducing a nonsynonymous substitution (p.Leu13639Phe or p.Leu478Phe) with a moderate predicted impact in *MUC16* (mucin 16, cell surface associated). Nonetheless, we highlight a few variants which overlapped clinically

relevant genes. For example, the most significant, greenlisted variant we identified was an inversion SV (INV_delly_INV00066128) residing in an intronic Alu copy within the *APP* (amyloid beta precursor protein) gene, an important biological marker for Alzheimer's disease (AD) [52]. Similarly, we identified several SNVs (rs61994687, rs1175403595, rs1343402870) in intronic or downstream regions of *PWRN1* within the Prader-Willi syndrome (PWS) region. To explore non-protein-coding roles greenlisted SNVs may play, we assigned them to ENCODE candidate cis-Regulatory Elements (cCREs) [53] (**Figure S3D**). Although about 8% (36/449) of greenlisted SNVs resided in an ENCODE cCRE, these were significantly depleted (FDR = 6.86E-15) in our greenlisted SNVs compared to background SNVs (**Figure S3D**). Ultimately, our results suggest that proximal, intergenic variation is associated with L1/Alu insertion singleton number variation.

To further explore the potential functional roles of SNV-associated genes, we leveraged expression data from the Genotype-Tissue Expression (GTEx) Portal [54-56] to assess the pattern of expression of genes linked to greenlisted SNVs across tissues (**Figure S3E**, **Table S1F**). In particular, expression patterns in the brain and gonads were of special interest, given that L1 activity tends to be more frequent in those tissues compared to others (discussed in [57]), and that L1/Alu integration events observed in our GWAS would have to occur in the germline for transmission across generations. Thus, we reasoned that if SNV-associated genes played roles in L1/Alu singleton number variation, they may be more abundantly expressed in the brain and in gonads. Interestingly, there was a cluster of genes that were very abundantly expressed in testes but not in other tissue types (including ovarian tissue), suggesting the existence of potential sex-specific mechanisms of *de novo* L1/Alu insertion transmission. Furthermore, there was also a cluster of genes that were abundantly expressed across brain regions and much less abundantly expressed across other tissue types. More generally, there were many SNV-linked genes that were abundantly expressed in more than ~50% of tissue types. These results highlight that significant SNV-associated genes have tissue-specific expression patterns, including some genes that are very abundant in tissues with documented L1 activity.

2.2 Significant SNVs are enriched near regulators of L1 expression

One of the primary motivations for carrying out this study was to identify novel, candidate regulators of L1 and/or Alu copy number. In particular, there is a gap in our understanding of Alu regulation, as, to our knowledge, no systematic screen for regulators of Alu activity has previously been carried out. To determine whether our approach captured genes with transposon regulatory potential, we assessed whether our list of greenlisted SNVs was enriched for 1) genes with known TE regulatory capabilities and 2) genes in broader pathways involved in TE regulation (**Figure 2A**). Previously, two CRISPR-based screens for regulators of L1 expression [38] and L1 transposition [34] were carried out in cancer cell lines. In addition, we also recently carried out an eQTL-based computational screen for candidate regulators of L1 RNA levels in lymphoblastoid cell lines [42]. Interestingly, our greenlisted SNVs were significantly (FDR = 3.59E-3) enriched in regions near known L1 expression regulators compared to the background list of all SNVs (**Figure 2B, 2C, 2E**). A few examples of these associations included rs1350516110 which was upstream of *RHOT1*, rs201619112 which was downstream of *XPR1*, and rs71475866 which was upstream of *PFKP*. Overall, we identified 24 greenlisted SNVs that were proximal to 10 genes previously annotated as capable of regulating L1 expression. Our greenlisted SNVs also captured genes involved in regulating L1 transposition, though there was no significant enrichment (FDR = 7.78E-1) (**Figure 2B, 2D, 2E**). A few examples of these associations included rs75237296 in the *PABPC1* 3'UTR, rs1288384419 in an intron of *RAD51B,* and rs1471205623 upstream of *MPHOSPH8*. Here, we identified 5 greenlisted SNVs near 3 genes capable of regulating L1 transposition. Importantly, *PABPC1* is a poly(A) binding protein that attaches to the poly(A) tail of L1, is important for the formation of L1 ribonucleoprotein particles (RNPs), and modules L1 and Alu transposition [58-60]. Additionally, *MPHOSPH8* is a component of the human silencing hub (HUSH) complex and is important for L1 repression, regulating both L1 expression and L1 transposition [34, 38, 61-63]. Finally, we checked the abundances of SNVs linked to candidate regulators of L1 RNA levels in lymphoblastoid cells [42]. However, we were not able to detect any greenlisted SNVs in regions containing candidate genes (**Figure 2B** and **2E**).

Nevertheless, these results highlight the ability of our approach to enrich for genomic regions containing known regulators of the retrotransposon lifecycle and suggest that these regulators may play important roles in diverse human populations.

We next repeated the above analyses using gene sets for broader pathways involved in TE regulation, including a gene set for “histone methyltransferase activity” (GO:0042054) and one for “RNA modifications” (GO:0009451). Though neither gene set was significantly enriched among our greenlisted SNVs (FDR = 1 for methyltransferase activity and FDR = 0.27 for RNA modification), there was some degree of overlap with each gene set (**Figure 2E**). We identified 4 greenlisted SNVs that were proximal to 2 genes with histone methyltransferase activity, including *EEF2KMT* and *PRDM7*. We also identified 8 greenlisted SNVs that were proximal to 3 genes with RNA modification capabilities, including *A1CF*, *ADARB2*, and *METTL14*. Importantly, ADARs (RNA-specific adenosine deaminases) are a family of double-stranded RNA (dsRNA)-binding proteins that modulate A-to-I editing events, including among Alu RNA species, which is important for preventing aberrant activation of innate immune signaling pathways [64]. Though *ADARB2* cannot catalyze A-to-I editing, it can negatively regulate the editing functions of other ADARs [64], making it a potential candidate regulator of Alu activity. These results demonstrate that our approach can capture genes implicated, but with uncharacterized roles, in TE regulation.

2.3 Significant SNVs are enriched in regions containing features that promote genome instability

After scanning for known and potential regulators of TE activity, we next explored the possibility that significant variants tagged genetically unstable TE loci (**Figure 3A**). Such loci could theoretically contribute to TE copy number variation through *de novo* transposition events. To probe this possibility, we assessed whether greenlisted SNVs were enriched for Alu and L1 loci belonging to subfamilies that have retained their ability to mobilize (**Figure 3B**). Specifically, Alu retrotransposons can be segregated into the old and inactive AluJ lineage, the middle-aged and active AluS lineage, and the young and active AluY lineage [8]. Interestingly, though there was no significant enrichment of

either AluJ- or AluY-overlapping SNVs (FDR = 0.76 and FDR = 1, respectively), SNVs overlapping AluS copies were significantly (FDR = 4.80E-6) enriched in our SNV greenlist compared to background (**Figure 3B**). Similarly, L1 retrotransposons can be segregated into the old L1M lineage, the middle-aged L1P lineage, and the young L1PA lineage. Importantly, the L1PA1/L1Hs subfamily within the L1PA lineage is the only actively mobile and autonomous subfamily within the human genome. Our enrichment analysis highlighted a significant (FDR = 7.29E-7) depletion of L1M-overlapping SNVs, a trending (FDR = 0.0786) enrichment of L1P-overlapping SNVs, and a significant (FDR = 3.46E-12) enrichment of L1PA-overlapping SNVs, all compared to background SNVs (**Figure 3B**). To obtain a higher resolution view of the transposition capabilities of overlapping L1PA copies, we checked the overlap of our greenlisted SNVs with L1 annotations on L1Base v2—a dedicated database of putatively active L1 insertions [65]. Surprisingly, active copies, with either an intact, full-length L1 or only an intact ORF2, were not significantly enriched/depleted in our greenlisted SNV list (**Figure 3C**). However, non-intact, full-length L1 copies—annotated for their regulatory potential—were significantly (FDR = 2.41E-6) enriched in our greenlisted SNV list compared to background. Though mutant L1s can be mobilized at very low frequencies by transposition-competent L1s *in trans* [9], these results suggest that most greenlisted SNV-overlapping L1 copies are limited in their ability to generate *de novo* insertions. This is potentially in contrast to greenlisted SNV-overlapping AluS copies, which may still be measurably active in the human genome. We must also consider the alternative interpretation that overlapping transposons may be directly involved in copy number variation through transposition-independent genomic remodeling, through processes such as repeat-mediated deletions and NAHR [12-15]. More generally, these results are consistent with the possibility that greenlisted SNVs tag reference Alu and L1 insertions that may contribute to TE copy number variation through transposition-dependent or transposition-independent mechanisms.

The TE enrichments we identified above were consistent with those identified using the Transposable Element Enrichment Analyzer (TEENA) [66] (**Table S1G**), which has the added advantage of analyzing other TE families, in addition to Alu and

L1, at subfamily-level resolution. The most significantly (FDR < 0.05) enriched Alu subfamilies included AluSg, AluSx3, AluSc, AluSx4, and AluSz6. Likewise, the most significantly enriched L1 subfamilies included L1PA3 and L1PA4. Unexpectedly, the most significantly enriched TE subfamily was HERVH-int of the ERV1 family. Interestingly, human specific endogenous retrovirus H (HERVH) is essential for maintenance of pluripotency in human stem cells (discussed in [67]). Finally, we observed a significant enrichment of other ERV1 subfamilies (PABL_A-int), ERVL subfamilies (HERVL-int), ERVK subfamilies (HERVK9-int, LTR13), and ERVL-MaLR subfamilies (THE1B-int). These results re-iterate the association between specific reference TE loci and variation in the number of L1/Alu singleton insertions.

We further explored the more general possibility that greenlisted SNVs tagged genomic regions that may be prone to transposition-independent structural alterations that may influence TE copy numbers. Such structural alterations may be facilitated by, but may not require, the presence of repetitive elements. In particular, extensive homology between segmental duplications, often in the vicinity of Alu elements [18], can facilitate NAHR and drive recurrent genomic rearrangements [68] that can help form SV hotspots [69, 70]. Noting the enrichment of AluS copies we observed among our greenlisted SNVs, we next assessed whether our greenlisted SNVs significantly overlapped regions of segmental duplication [71, 72] or regions characterized as SV hotspots [70] (**Figure 3D**). Consistent with the notion that SNVs may tag regions with potentially elevated rates of genome instability, our SNVs were very significantly enriched in regions of segmental duplication (FDR = 2.64E-99), as well as in regions harboring SV hotspots (FDR = 8.82E-7). These results further link variation in TE copy number to genomic loci where structural instability may arise through transposition-independent mechanisms.

Finally, we note that our GWAS analysis identified 21 polymorphic SVs that were significantly (FDR < 0.05) associated with the presence/absence of L1/Alu insertion singletons. These polymorphic SVs varied in nature and included inversions, Alu insertions, an L1 insertion, SINE-VNTR-Alu (SVA) insertions, and a multiallelic copy

number variant (**Figure S4A**). With the exception of the CN0 copy number variant which was associated with lower odds of carrying an L1/Alu insertion singleton (odds ratio = 0.28), all of the other structural variants were associated with higher odds of carrying an insertion singleton (odds ratio > 1). Since the sequences for these SVs were not available, it is unclear whether common, polymorphic L1/Alu insertion SVs may be directly increasing the singleton number through novel transposition events, or whether any of these polymorphic SVs may be influencing the L1/Alu singleton number through transposition-independent mechanisms. Indeed, polymorphic inversions, many of which are often flanked by retrotransposons, are associated with genetic instability [73]. Ultimately, these results suggest a tight link between common, polymorphic SVs of different types and L1/Alu singleton SVs, whereby having the former is generally associated with higher odds of having the latter.

2.4 Case samples exhibit elevated cell cycle-related gene expression profiles

To gain insight into the functional differences between controls and cases, we leveraged publicly available lymphoblastoid cell line mRNA-seq data generated by the GEUVADIS consortium for a subset of European and African samples in the 1000 Genomes Project [74] (**Figure 4A**). This included 358 Europeans samples from 4 populations (British, Finnish, Tuscan, and Utah residents with European ancestry) and 86 African samples from 1 population (Yoruba), which we recently used to quantify gene and TE expression profiles [42]. We utilized this expression data to construct consensus gene co-expression networks for both the European and African samples using the WGCNA [75] package. This approach led to the identification of 20 consensus modules and 1 module (MEgrey) containing genes that were not assigned to the consensus modules (**Table S1H**). We then ran a module-trait correlation analysis comparing the expression of these modules with the case/control status of the European and African samples (**Figure 4B**). Here, we used a stricter threshold of $p < 0.01$ to call significant correlations. We were not able to identify any significant module-phenotype correlations using the European network, which is potentially consistent with our difficulty in calling significant GWAS variants in this super-population at the available sample sizes (**Figure S2C**). In contrast, the MEroyalblue module was significantly ($p = 4.0E-4$) correlated with

African case/control status. To combine the results from each network, we utilized Fisher's method to meta-analyze the p-values for modules exhibiting correlations in the same direction. By meta-analysis, the MEroyalblue module was still significantly ($p = 0.002$) and positively correlated with case status. Finally, to functionally characterize this module, we ran over-representation analysis using the GO Biological Process gene set collection (**Figure 4C, Table S1I**). The top 10 over-represented gene sets were involved in cell cycle-related processes, including "mitotic cell cycle", "cell division", and "sister chromatid segregation". These findings are consistent with the biology of TE copy number expansion. Though L1 can mobilize in non-dividing cells [76, 77], L1 retrotransposition exhibits a cell cycle bias and peaks during the S phase [78]. Alternatively, chromosome segregation errors during mitosis or meiosis can generate cells with abnormal ploidy and either increased or decreased dosages of both genic and transposon content [79]. These results implicate cell cycle differences in cells from individuals with unique L1/Alu insertion singleton variation.

3. DISCUSSION

3.1 A new approach to identify loci implicated in L1 and Alu copy number variation

In this work, we developed a pipeline to computationally identify candidate loci involved in L1/Alu singleton number variation by GWAS analysis. Importantly, our study incorporates natural human genetic variation present in populations of different geographic origin via trans-ethnic GWAS meta-analysis to identify shared, candidate regulatory loci. Though several studies have begun to screen for regulators and potential regulators of L1 expression or transposition in cell culture models or across tissues [34-41], these can be limited in that the generalizability of these findings to different ethnic populations is unclear. Moreover, no systematic, genome-wide screen for candidate regulators of Alu activity has been carried out thus far, to our knowledge. To address these gaps, we previously utilized trans-eQTL analysis to identify potential regulators of L1 RNA levels in European and African populations [42]. Here, we utilized genomic data from samples originating from 5 super-populations to identify candidate loci modulating L1/Alu insertion singleton numbers.

TE copy number variation can arise through *de novo* transposition events or through transposition-independent mechanisms, including recombination-based mechanisms that can generate large deletions or duplications. We were particularly interested in identifying new candidate regulators of L1/Alu transposition. Consistent with the notion that greenlisted SNVs may play roles in the retrotransposon lifecycle, our approach enriched genomic regions containing genes that can regulate L1 expression levels. Though other known regulators of TE activity, and pathways involved in TE control, were not significantly enriched among our greenlisted SNVs, we nonetheless identified many SNVs in genomic regions containing these genes. This included, for example, *MPHOSPH8*—a component of the HUSH complex important for L1 repression, regulating both L1 expression and transposition [34, 38, 61-63]. As another example, we identified variants near *ADARB2*, a negative regulator of RNA A-to-I editing, including among Alu RNAs [64]. These results suggest that SNV-associated genes identified in this study hold TE regulatory potential and it may therefore be

informative to (i) test and validate these in future studies or (ii) use these to prioritize future, targeted studies of TE regulators.

Our approached also identified an enrichment of greenlisted SNVs in regions containing reference TE insertions, including AluS and full-length, non-intact L1PA copies. Though neither of these can mobilize autonomously, they can hijack machinery from transposition-competent L1s and mobilize *in trans* [6, 7, 9]. Thus, it is possible that greenlisted SNVs tag reference insertions contributing to L1/Alu singleton variation through transposition-dependent mechanisms. Of course, an alternative possibility is that these repetitive elements are directly involved in genomic remodeling involving transposition-independent mechanisms like repeat-mediated deletions or NAHR. We also note that greenlisted SNVs were enriched in regions containing segmental duplications and structural variation hotspots where recombination-based mechanisms, including NAHR, may lead to duplications or deletions of the local genomic architecture. Thus, it is also possible that greenlisted SNVs tag genomic regions prone to structural variation that can alter the L1/Alu copy number through recombination-dependent mechanisms.

Importantly, we also suggest the possibility that genome remodeling mechanisms (including recombination) may interact with gene-based mechanisms of TE regulation. Indeed, genes such as *BRCA1* are known to regulate L1 transposition [34] and are also known to undergo Alu-Alu recombination events that can give rise to new mutations in the gene [80-82]. Observations such as these highlight the possibility that TE insertions may modulate structural variation in genomic regions containing genes regulating retrotransposon lifecycles, which may facilitate an expansion of TE copy numbers through transposition-based mechanisms, which may influence further structural variation driving this whole process. This possibility is consistent with the enrichment of greenlisted SNVs in regions containing L1 expression regulators, Alu and L1 repeats, and other genomically unstable features like segmental duplications and structural variation hotspots. In the future, it may be informative to experimentally assess this

possibility in contexts where genome instability is a hallmark feature that is coupled with TE de-repression, such as aging [22] or aging-associated diseases like cancer [83, 84].

Finally, our approach also identified several common, polymorphic SVs that were significantly associated with L1/Alu insertion singletons. Overwhelmingly, the presence of polymorphic SVs of different types—inversions, Alu insertions, an L1 insertion, and SVA insertions—was associated with increased odds of a global L1/Alu insertion singleton. The exception to this was a multi-allelic copy number variant where 0 copies were present; this SV was associated with decreased odds of a global L1/Alu insertion singleton. Based on these results, we speculate that specific polymorphic SVs (i) may directly drive genome instability that can facilitate the acquisition of L1/Alu copies or (ii) may serve as markers for elevated risk of indirectly acquiring additional L1/Alu copies. Indeed, active donor L1 copies that can mobilize and generate *de novo* insertions are usually highly polymorphic in human populations (reviewed in [57]), and polymorphic inversions, many of which are often flanked by retrotransposons, are also associated with genetic instability and genomic disorders [73].

In summation, this study provides a list of variants that are associated with L1/Alu insertion singletons and includes (i) SNVs in regions containing regulators of TE activity, (ii) SNVs in regions containing features associated with genome instability, including retrotransposons, that may influence TE copy number variation through transposition-dependent or transposition-independent mechanisms, and (iii) common, polymorphic SVs that may also influence TE copy number variation through transposition-dependent or transposition-independent mechanisms.

3.2 L1/Alu insertion singleton-associated loci contain genes of clinical relevance

The most significant, greenlisted variant we identified was a polymorphic inversion SV (INV_delly_INV00066128, odds ratio = 4.38, FDR = 3.42E-17, chr21:26001780) residing in an intronic Alu copy within the *APP* gene, an important marker of Alzheimer's disease (AD). AD is characterized by (i) the accumulation of

amyloid β (Aβ) plaques derived from amyloidogenic *APP* processing and (ii) neurofibrillary tangles of hyperphosphorylated tau [85]. Importantly, tau protein can induce TE expression and there is speculation that TEs may mobilize in tauopathies [85]. Whether APP protein or Aβ plaques can similarly modulate TE expression or mobilization is an interesting area of potential future research; indeed, our findings are consistent with the possibility that *APP* products may act as regulators of TE activity. Of course, another possibility is that the genomic region containing *APP* may be a source of L1/Alu copy number variation independent of the functional properties of APP protein (i.e. genomic instability at that locus may be the driver of TE copy number variation). Nevertheless, our results offer another connection between transposable element regulation and Alzheimer's disease.

We also identified several SNVs proximal to *PWRN1*, which resides within the Prader-Willi syndrome region and is thought to play a role in PWS [86]. Prader-Willi is an imprinting disorder where genes in the chromosome 15q11-q13 region are maternally imprinted and paternal copies are not expressed [87]. This lack of paternal gene expression is predominantly caused by *de novo* paternally inherited deletions of the 15q11-q13 region, though, less frequently, inheritance of two maternal chromosome 15 copies is the cause [87]. Importantly, a feature of genomic disorders like PWS is the presence of segmental duplications that can serve as substrates for NAHR [88, 89]. Thus, we hypothesize that this particular region might be more prone to L1/Alu copy number variation as a consequence of recombination-based chromosomal alterations. Nevertheless, it is unclear (i) whether L1/Alu repeats are differentially active in PWS compared to healthy controls or, more specifically, (ii) whether genes like *PWRN1* can differentially express or mobilize L1/Alu transposons. Ultimately, the associations between L1/Alu singletons and both *APP* and *PWRN1* further implicate retrotransposons in disease.

3.3 Limitations and future considerations

In this study, we sought to identify new, candidate genes implicated in Alu and L1 copy number control. One specific mechanism of interest by which this can occur is target-primed reverse transcription (TPRT)-mediated transposition. L1 insertions generated by this method usually carry specific features, including short target site duplications (TSDs), a polyadenine (polyA) tail, and are found integrated at an L1 endonuclease motif (reviewed in [57]). Since insertion sequences for 1000 Genomes Project samples were not available, to our knowledge, it is difficult to assess to what degree TPRT is driving the associations we identified. In studies with larger cohorts where insertion sequences are available and insertions with TPRT features can be identified, our approach could theoretically be applied to explore the genetic basis of TPRT-specific copy number variation. Of course, our approach is generally restricted to identifying genomic loci where variation is common across human populations. We note, however, that significant variants were enriched in regions containing genes involved in L1 expression regulation. Since expression is one of the early steps of the L1 life cycle, our approach captured variants and genes with potential significance to TPRT-mediated transposition.

We also note that there are several variables that we are unable to control for in this study. To protect patient privacy, biological covariates such as chronological age were not available and therefore could not be corrected for in our analysis. Since increases in L1 expression and copy number have been observed with chronological aging [23], differences in copy number may reflect age differences rather than genetic differences. Importantly, however, samples were considered healthy at the time of sample collection, potentially mitigating health-related effects on copy number. The origin and developmental timing of the rare Alu and L1 insertions we utilized is also unclear. That is, it is unclear whether global singletons used in this study arose *de novo* in the germline, arose during life through somatic mutation, or even whether they just arose during the cultivation of the lymphoblastoid cells used to amplify each sample's DNA. Depending on when these insertions were acquired, the associations identified in this study may be relevant for either germ cell or somatic cell TE biology. A potential avenue of future research to address this question would be the incorporation of trio

parent-child genome sequencing and multi-generational genome sequencing to help identify *bona fide de novo* insertions and their developmental timing. Ultimately, future studies modulating genes identified with our approach will need to be carried out to validate any causal contributions to L1/Alu regulation.

3.4 Conclusions

In this study, we employed GWAS across human populations of different geographic origin to computationally identify genomic loci associated with variation in L1/Alu insertion singleton number, specifically, and L1/Alu-associated copy number variation, more generally. Our approach enriches for SNVs in genomic regions containing known regulators of L1 expression. This observation suggests that the TE-regulatory properties of these genes may extend beyond isogenic cell culture models to more diverse human populations. Moreover, this observation also suggests that our list of associated genes likely contains novel regulators of L1 or Alu activity that may be prioritized in future, validation studies. Our approach also identified reference insertions and non-reference, polymorphic SVs that may modulate L1/Alu copy numbers through transposition-dependent or transposition-independent mechanisms. Finally, the observation that some significant variants reside in genes of clinical relevance, like *APP* and *PWRN1*, reinforce accumulating evidence of biological associations between TE regulation and disease. Ultimately, our approach adds to the analytical toolkit that can be used to study the regulation of TE activity.

4. METHODS

4.1 Publicly available genomic data acquisition

The multi-ethnic GWAS analysis was carried out on 2503 individuals derived from 5 super-populations (African, East Asian, European, South Asian, and American) and for which paired single nucleotide variant and structural variant data were available from Phase 3 of the 1000 Genomes Project [90-92]. Specifically, Phase 3 autosomal SNVs called on the GRCh38 reference genome were obtained from The International Genome Sample Resource (IGSR) FTP site (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/). Structural variants, called against the GRCh37 reference genome and then lifted over to GRCh38, were also obtained from the IGSR FTP site (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/).

Human gene expression data across 54 non-diseased tissue sites was obtained from the GTEx Analysis v8 on the GTEx Portal [54-56]. Specifically, we downloaded the matrix containing the median gene-level transcripts per million (TPMs) by tissue, and we extracted the expression values for significant SNV-associated genes. After filtering out genes with no detectable expression (0 TPMs), we generated heatmaps using the pheatmap v1.0.12 package in R. Gene expression values were centered and scaled across tissues to visualize and compare the relative expression levels across tissues.

4.2 Aggregating and pre-processing genotype data for GWAS analysis

To define the phenotype of interest for GWAS, we first extracted global singleton Alu and L1 insertions. We utilized VCFtools v0.1.17 [93] to extract all autosomal SVs with no missing data (--max-missing 1) and an allele count of 1 across all samples (--non-ref-ac 1 --max-non-ref-ac 1), i.e. global singletons. From these, we extracted Alu- and L1-specific insertions using BCFtools v1.10.2 [94] to keep entries annotated with the ‘SVTYPE=“LINE1”’ and ‘SVTYPE=“ALU”’ tags. Finally, VCF files containing global

singleton L1 or Alu insertions were converted to PLINK BED format using PLINK v1.90b6.17 [95].

We note that SVs on sex chromosomes were not included in any part of the analysis since (i) Y chromosome SVs were not available, (ii) male genotypes on chromosome X were unknown, and (iii) association studies with X chromosome variants require unique algorithms and cannot easily be incorporated into traditional association pipelines [96, 97].

Secondly, we prepared polymorphic SVs for inclusion in the GWAS analysis. VCFtools was used to isolate SVs with the following properties in each individual super-population: possessed a minimum and maximum of two alleles (biallelic), possessed a minor allele frequency (MAF) of at least 1%, passed Hardy-Weinberg equilibrium thresholding at $p < 1e-6$, had no missing samples, and was located on an autosome. To focus on shared, trans-ethnic sources of genetic variation, we used BCFtools to identify and subset SVs that were shared across all 5 super-populations.

Third, we prepared polymorphic SNVs for inclusion in the GWAS analysis. All SNVs were first annotated with rsIDs from dbSNP build 155 using BCFtools. Within each super-population, VCFtools was used to remove indels and keep autosomal SNVs with the same parameters as the polymorphic SVs. We note that for similar reasons as with the polymorphic SVs, sex chromosome SNVs were also omitted from all analyses. We then used BCFtools to identify and subset SNVs that were shared across all 5 super-populations. Finally, we used BCFtools to generate the final genotype matrices by combining shared, polymorphic SNVs with shared, polymorphic SVs. VCF files containing the combined SNVs and SVs were then converted to PLINK BED format using PLINK, keeping the allele order. PLINK was also used to prune the combined SNV and SV matrices (--indep-pairwise 50 10 0.1) and to generate principal components (PCs) from the pruned genotypes, for inclusion as covariates in the GWAS.

4.3 Super-population-specific and trans-ethnic GWAS

We began by running GWAS within each super-population using PLINK v1.90b6.17 [95]. For the phenotype, we added the number of Alu and L1 global singleton insertions for each sample and segregated samples into cases and controls, depending on whether they contained or did not contain a global singleton. We ran GWAS analyses using a logistic regression model that included the following covariates: biological sex and the top 4 principal components from the pruned SNV and SV genotype matrices. Individual results from each super-population were combined via meta-analysis using PLINK. To help call significant variants, we generated a null distribution of p-values for each super-population by running 20 instances of the GWAS where the case-control status for each sample was randomly shuffled with the case-control status of a different sample. Each set of permutation results was meta-analyzed across super-populations to similarly obtain 20 random distributions of meta-analysis p-values. For the meta-analysis, we focused on the p-values and odds ratios generated using a random effects statistical model, as opposed to a fixed effects model, since 1) there may be heterogeneity across super-populations in response to different genetic variants, and 2) we were interested in enhancing the generalizability of our findings to facilitate future follow-up studies.

To limit false positives, the Benjamini-Hochberg (BH) false discovery rate (FDR) was calculated in each analysis, and we used the p-value corresponding to a BH FDR < 0.05 as the threshold for GWAS significance. As a secondary threshold, we used the permutation data to identify p-values corresponding to an average empirical FDR < 0.05. To note, we calculated the average empirical FDR at a given p-value p_i by (i) counting the total number of null points with $p \leq p_i$, (ii) dividing by the number of permutations, to obtain an average number of null points with $p \leq p_i$, and (iii) dividing the average number of null points with $p \leq p_i$ by the number of real points with $p \leq p_i$. GWAS variants were considered significant if they passed the stricter of the two thresholds in each analysis.

4.4 Annotation of variants and annotation enrichment analyses

We obtained BED files containing annotated genomic regions from various sources. We obtained the ENCODE blacklist v2 [46] for hg38 from https://github.com/Boyle-Lab/Blacklist/tree/master/lists. Segmental duplications [71, 72] and RepeatMasker annotations using the Repbase library [98] were obtained from the UCSC Genome Browser [99]. We obtained SV hotspot coordinates on hg19 from [70], and we used the online UCSC LiftOver tool to map coordinates to the hg38 genome assembly using the default settings. The BED tracks for full-length and intact L1s, only ORF2-intact L1s, and full-length non-intact L1s were obtained from L1Base v2 [65]. We obtained the Registry (version 4) of candidate cis-Regulatory Elements (cCREs) for hg38 from the Search Candidate Regulatory Elements by ENCODE (SCREEN) web interface [53] (http://screen-beta.wenglab.org). We used the 'intersect' command in BEDTools v2.31.1 [100] to assign genomic region annotations to all overlapping variants used in this study. We were also interested in assessing whether variants were linked to specific regulatory gene annotations. All variants used in the study were submitted to the GREAT v4.0.4 [47, 48] online platform with the default settings (basal plus extension, proximal with 5 kb upstream and 1 kb downstream, plus distal up to 1000 kb) to assign gene annotations to each variant. These gene annotations were then used to assess the number of variants linked to genes in several TE regulatory lists—including a list of genes that regulated L1 expression in a CRISPR screen using cancer cells [38], a list of genes that regulated L1 transposition in an independent CRISPR screen using cancer cells [34], a list of candidate genes influencing intronic, intergenic, or exonic L1 RNA levels in lymphoblastoid cell lines [42], a list of genes with histone methyltransferase activity (GO:0042054), and a list of genes with RNA modification activity (GO:0009451). The two GO gene sets were obtained on 2024-07-29 from the Molecular Signatures Database (MSigDB) v2023.2.Hs [101, 102], corresponding to GO release 2023-07-27.

Given the limited number of significant SVs and the unavailability of SV sequences, we largely focused on blacklist-filtered, significant SNVs in downstream enrichment analyses. For each of the above annotations, we assessed whether greenlisted SNVs were significantly enriched or depleted for that annotation compared

to background SNVs—all SNVs that were tested in the GWAS. The numbers of background and greenlisted SNVs overlapping or not overlapping a set of annotations were placed into a contingency table, and statistical significance was assessed using Fisher's exact test (with the fisher.test function in R v4.3.3). After all tests were carried out, p-values were FDR corrected using the p.adjust function in R. All enrichments/depletions with an FDR < 0.05 were considered significant. For the repeat enrichment analyses, we also analyzed our greenlisted SNVs using the TEENA web server [66] (on August 8, 2024), specifying the hg38 assembly and using all other default options.

4.5 RNA-seq and gene co-expression network analyses

For lymphoblastoid cell line transcriptional analyses, mRNA-sequencing was initially carried out by the GEUVADIS consortium [74] on LCLs from a small subset of European and African (Yoruban, specifically) samples from the 1000 Genomes Project. Recently, we re-processed this data to quantify gene and transposon expression levels [42]. Briefly, reads were trimmed using fastp v0.20.1 [103], trimmed reads were aligned to the GRCh38 human genome assembly using STAR v2.7.3a [104], and the TEtranscripts v2.1.4 [105] package was used to obtain gene and TE counts, using the GENCODE release 33 [106] annotations and a repeat GTF file provided on the Hammell lab website. To note, the EBV genome (GenBank ID V01555.2) was included as an additional contig in our reference genome, since LCLs are generated by infecting B-cells with Epstein-Barr virus (EBV).

Using these gene/TE count matrices, lowly expressed genes were filtered out if 50% of European or Yoruban samples did not have over 0.44 counts per million (cpm) or 0.43 cpm, respectively, which correspond to 10 reads in each cohort's median-length library. Since we were interested in building consensus co-expression networks between the European and Yoruban samples, we also removed genes that were not expressed in both groups. After, the filtered counts underwent a variance stabilizing transformation (vst) using DESeq2 v1.42.1 [107] and the following covariates were regressed out using the 'removeBatchEffect' function in Limma v3.58.1 [108]: biological

sex, sequencing lab, population category, principal components 1-2 of the pruned genotype matrices containing both SNVs and SVs, and EBV expression levels. The population category variable was omitted in the Yoruban batch correction since that did not vary.

The batch-corrected VST matrices were then used to perform weighted gene co-expression network analysis (WGCNA) [75] using the WGCNA v1.72-5 R package. We used the 'blockwiseConsensusModules' function to automate consensus network construction for both the European and African expression data, specifying these parameters: corType = "bicor", power = 12, networkType = "signed", maxPOutliers = 0.05, mergeCutHeight = 0.25, deepSplit = 2, minKMEtoStay = 0, pamRespectsDendro = FALSE, minModuleSize = 30, and randomSeed = 90280. Phenotype-module correlations, and the corresponding p-values, were calculated using WGCNA's 'cor' and 'corPvalueFisher' functions, respectively. The p-values for the European and Yoruban correlations were meta-analyzed using Fisher's method. For visualization purposes only, to show a correlation direction in the meta-analysis, we took the average of the European and Yoruban correlations. Modules showing opposite correlations across the two consensus networks were disregarded in the meta-analysis. Correlations with a p-value < 0.01 were considered significant.

4.6 Functional enrichment analyses

We used the over-representation analysis (ORA) paradigm as implemented in the R package clusterProfiler v4.10.1 [49]. Gene Ontology Biological Process gene sets were obtained from the R package msigdbr v7.5.1, an Ensembl ID-mapped collection of gene sets from the Molecular Signatures Database [101, 102]. For ORA with genes linked to greenlisted SNVs and SVs, we used the genes linked to the background SNVs and SVs, respectively, for the universe background to compute enrichment significance. For ORA analysis of co-expression network modules, we used all genes in the network for the universe background. All gene sets with an FDR < 0.05 were considered significant, and the top 10 significant gene sets, at most, were plotted. All enrichments results are included in **Table S1C, S1D,** and **S1I**.

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All code is available on the Benayoun lab GitHub (https://github.com/BenayounLaboratory/TE_GWAS). Analyses were conducted using R version 4.3.3. Code was re-run independently on R version 4.3.0 to check for reproducibility.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Science Foundation [https://www.nsf.gov/] Graduate Research Fellowship Program (NSF GRFP) DGE-1842487 (J.I.B.), the National Institute on Aging [https://www.nia.nih.gov/] T32 AG052374 (J.I.B.), the University of Southern California with a Provost Fellowship (J.I.B.), and the National Institute of General Medical Sciences [https://www.nigms.nih.gov/] R35 GM142395 (to B.A.B).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Juan I Bravo: Conceptualization, formal analysis, investigation, methodology, visualization, writing - original draft preparation, writing - review and editing. **Lucia Zhang:** Validation, formal analysis, writing - review and editing. **Bérénice A.**

Benayoun: Conceptualization, formal analysis, funding acquisition, methodology, supervision, visualization, writing - original draft preparation, writing - review and editing.

Acknowledgements

We would like to thank Prof. Rachel Brem at the University of California Berkeley for her feedback and insights on the GWAS analyses. We would also like to thank Dr. Heather C. Mefford at St. Jude Children's Research Hospital for referring us to publications that helped shape the analyses linking our variants to regions potentially prone to genomic instability and involved in disease. Finally, we would like to thank Dr. Minhoo Kim and Mr. Aaron Lemus for their comments on the manuscript.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The Sequence of the Human Genome.** *Science* 2001, **291:**1304-1351.
3. Ostertag EM, Kazazian HH, Jr.: **Biology of mammalian L1 retrotransposons.** *Annu Rev Genet* 2001, **35:**501-538.
4. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH: **Hot L1s account for the bulk of retrotransposition in the human population.** *Proceedings of the National Academy of Sciences* 2003, **100:**5280-5285.
5. Rubin CM, Houck CM, Deininger PL, Friedmann T, Schmid CW: **Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences.** *Nature* 1980, **284:**372-374.
6. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences.** *Nature Genetics* 2003, **35:**41-48.
7. Wallace N, Wagstaff BJ, Deininger PL, Roy-Engel AM: **LINE-1 ORF1 protein enhances Alu SINE retrotransposition.** *Gene* 2008, **419:**1-6.
8. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE: **Active Alu retrotransposons in the human genome.** *Genome Research* 2008, **18:**1875-1883.
9. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV: **Human L1 Retrotransposition: cisPreference versus trans Complementation.** *Molecular and Cellular Biology* 2001, **21:**1429-1439.
10. Mendez-Dorantes C, Bhargava R, Stark JM: **Repeat-mediated deletions can be induced by a chromosomal break far from a repeat, but multiple pathways suppress such rearrangements.** *Genes Dev* 2018, **32:**524-536.
11. Mendez-Dorantes C, Tsai LJ, Jahanshir E, Lopezcolorado FW, Stark JM: **BLM has Contrary Effects on Repeat-Mediated Deletions, based on the Distance of DNA DSBs to a Repeat and Repeat Divergence.** *Cell Reports* 2020, **30:**1342-1357.e1344.
12. Boone Philip M, Yuan B, Campbell Ian M, Scull Jennifer C, Withers Marjorie A, Baggett Brett C, Beck Christine R, Shaw Christine J, Stankiewicz P, Moretti P, et al: **The Alu-Rich Genomic Architecture of SPAST Predisposes to Diverse and Functionally Distinct Disease-Associated CNV Alleles.** *The American Journal of Human Genetics* 2014, **95:**143-161.
13. Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P, Patel A, Gambin A, Shaw CA, Rosenfeld JA, Stankiewicz P: **Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination.** *BMC Biology* 2014, **12:**74.
14. Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A: **Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination.** *Nucleic Acids Research* 2015, **43:**2188-2198.

893 15. Robberecht C, Voet T, Esteki MZ, Nowakowska BA, Vermeesch JR: **Nonallelic**
894 **homologous recombination between retrotransposable elements is a driver of de**
895 **novo unbalanced translocations.** *Genome Research* 2013, **23:**411-418.
896 16. Belancio VP, Deininger PL, Roy-Engel AM: **LINE dancing in the human genome:**
897 **transposable elements and disease.** *Genome Medicine* 2009, **1:**97.
898 17. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.**
899 *Nature Reviews Genetics* 2009, **10:**691-703.
900 18. Bailey JA, Liu G, Eichler EE: **An Alu Transposition Model for the Origin and**
901 **Expansion of Human Segmental Duplications.** *The American Journal of Human Genetics*
902 2003, **73:**823-834.
903 19. Kolomietz E, Meyn MS, Pandita A, Squire JA: **The role of Alu repeat clusters as**
904 **mediators of recurrent chromosomal aberrations in tumors.** *Genes, Chromosomes and*
905 *Cancer* 2002, **35:**97-112.
906 20. Bravo JI, Nozownik S, Danthi PS, Benayoun BA: **Transposable elements, circular RNAs**
907 **and mitochondrial transcription in age-related genomic regulation.** *Development* 2020,
908 **147**.
909 21. Zhang X, Zhang R, Yu J: **New Understanding of the Relevant Role of LINE-1**
910 **Retrotransposition in Human Disease and Immune Modulation.** *Frontiers in Cell and*
911 *Developmental Biology* 2020, **8**.
912 22. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G: **Hallmarks of aging: An**
913 **expanding universe.** *Cell* 2023, **186:**243-278.
914 23. De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA: **Transposable**
915 **elements become active and mobile in the genomes of aging mammalian somatic**
916 **tissues.** *Aging (Albany NY)* 2013, **5:**867-883.
917 24. De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J,
918 Peterson AL, Kreiling JA, Neretti N, Sedivy JM: **Genomes of replicatively senescent cells**
919 **undergo global epigenetic changes leading to gene silencing and activation of**
920 **transposable elements.** *Aging Cell* 2013, **12:**247-256.
921 25. Campisi J: **Aging, Cellular Senescence, and Cancer.** *Annual Review of Physiology* 2013,
922 **75:**685-705.
923 26. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A: **Inflammaging: a new**
924 **immune–metabolic viewpoint for age-related diseases.** *Nature Reviews Endocrinology*
925 2018, **14:**576-590.
926 27. Wallace NA, Belancio VP, Deininger PL: **L1 mobile element expression causes multiple**
927 **types of toxicity.** *Gene* 2008, **419:**75-81.
928 28. Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P: **Somatic expression of LINE-1**
929 **elements in human tissues.** *Nucleic acids research* 2010, **38:**3909-3922.
930 29. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G,
931 Adney EM, Boeke JD, et al: **L1 drives IFN in senescent cells and promotes age-**
932 **associated inflammation.** *Nature* 2019, **566:**73-78.
933 30. Yamada K, Kaneko H, Shimizu H, Suzumura A, Namba R, Takayama K, Ito S, Sugimoto M,
934 Terasaki H: **Lamivudine Inhibits Alu RNA-induced Retinal Pigment Epithelium**
935 **Degeneration via Anti-inflammatory and Anti-senescence Activities.** *Translational*
936 *Vision Science & Technology* 2020, **9:**1-1.

937 31. Wang J, Geesman GJ, Hostikka SL, Atallah M, Blackwell B, Lee E, Cook PJ, Pasaniuc B,
938 Shariat G, Halperin E, et al: **Inhibition of activated pericentromeric SINE/Alu repeat**
939 **transcription in senescent human adult stem cells reinstates self-renewal.** *Cell Cycle*
940 2011, **10:**3016-3030.
941 32. Levin HL, Moran JV: **Dynamic interactions between transposable elements and their**
942 **hosts.** *Nature Reviews Genetics* 2011, **12:**615-627.
943 33. Rebollo R, Romanish MT, Mager DL: **Transposable Elements: An Abundant and Natural**
944 **Source of Regulatory Sequences for Host Genes.** *Annual Review of Genetics* 2012,
945 **46:**21-42.
946 34. Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J: **Selective silencing of**
947 **euchromatic L1s revealed by genome-wide screens for L1 regulators.** *Nature* 2018,
948 **553:**228-232.
949 35. Tristan-Ramos P, Morell S, Sanchez L, Toledo B, Garcia-Perez JL, Heras SR: **sRNA/L1**
950 **retrotransposition: using siRNAs and miRNAs to expand the applications of the cell**
951 **culture-based LINE-1 retrotransposition assay.** *Philosophical Transactions of the Royal*
952 *Society B: Biological Sciences* 2020, **375:**20190346.
953 36. Mita P, Sun X, Fenyö D, Kahler DJ, Li D, Agmon N, Wudzinska A, Keegan S, Bader JS, Yun
954 C, Boeke JD: **BRCA1 and S phase DNA repair pathways restrict LINE-1**
955 **retrotransposition in human cells.** *Nat Struct Mol Biol* 2020, **27:**179-191.
956 37. Briggs EM, Mita P, Sun X, Ha S, Vasilyev N, Leopold ZR, Nudler E, Boeke JD, Logan SK:
957 **Unbiased proteomic mapping of the LINE-1 promoter using CRISPR Cas9.** *Mobile DNA*
958 2021, **12:**21.
959 38. Li X, Bie L, Wang Y, Hong Y, Zhou Z, Fan Y, Yan X, Tao Y, Huang C, Zhang Y, et al: **LINE-1**
960 **transcription activates long-range gene expression.** *Nature Genetics* 2024, **56:**1494-
961 1502.
962 39. Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, Mita P, Fenyö D, Boeke JD:
963 **Transcription factor profiling reveals molecular choreography and key regulators of**
964 **human retrotransposon expression.** *Proc Natl Acad Sci U S A* 2018, **115:**E5526-e5535.
965 40. Chung N, Jonaid GM, Quinton S, Ross A, Sexton CE, Alberto A, Clymer C, Churchill D,
966 Navarro Leija O, Han MV: **Transcriptome analyses of tumor-adjacent somatic tissues**
967 **reveal genes co-expressed with transposable elements.** *Mobile DNA* 2019, **10:**39.
968 41. Tristán-Ramos P, Rubio-Roldan A, Peris G, Sánchez L, Amador-Cubero S, Viollet S,
969 Cristofari G, Heras SR: **The tumor suppressor microRNA let-7 inhibits human LINE-1**
970 **retrotransposition.** *Nature Communications* 2020, **11:**5712.
971 42. Bravo JI, Mizrahi CR, Kim S, Zhang L, Suh Y, Benayoun BA: **An eQTL-based approach**
972 **reveals candidate regulators of LINE-1 RNA levels in lymphoblastoid cells.** *PLOS*
973 *Genetics* 2024, **20:**e1011311.
974 43. Johnston HR, Hu Y, Cutler DJ: **Population Genetics Identifies Challenges in Analyzing**
975 **Rare Variants.** *Genetic Epidemiology* 2015, **39:**145-148.
976 44. Evangelou E, Ioannidis JPA: **Meta-analysis methods for genome-wide association**
977 **studies and beyond.** *Nature Reviews Genetics* 2013, **14:**379-389.
978 45. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, Lam M,
979 Iyegbe C, Strawbridge RJ, Brick L, et al: **Genome-wide Association Studies in Ancestrally**

980 **Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations.** *Cell*
981 2019, **179:**589-603.
982 46. Amemiya HM, Kundaje A, Boyle AP: **The ENCODE Blacklist: Identification of Problematic**
983 **Regions of the Genome.** *Scientific Reports* 2019, **9:**9354.
984 47. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G:
985 **GREAT improves functional interpretation of cis-regulatory regions.** *Nature*
986 *Biotechnology* 2010, **28:**495-501.
987 48. Tanigawa Y, Dyer ES, Bejerano G: **WhichTF is functionally important in your open**
988 **chromatin data?** *PLOS Computational Biology* 2022, **18:**e1010378.
989 49. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al:
990 **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.** *The*
991 *Innovation* 2021, **2:**100141.
992 50. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A**
993 **program for annotating and predicting the effects of single nucleotide polymorphisms,**
994 **SnpEff.** *Fly* 2012, **6:**80-92.
995 51. Liu Y, Yang Q, Zhao F: **Synonymous but Not Silent: The Codon Usage Code for Gene**
996 **Expression and Protein Folding.** *Annual Review of Biochemistry* 2021, **90:**375-401.
997 52. Hampel H, Hardy J, Blennow K, Chen C, Perry G, Kim SH, Villemagne VL, Aisen P,
998 Vendruscolo M, Iwatsubo T, et al: **The Amyloid-β Pathway in Alzheimer's Disease.**
999 *Molecular Psychiatry* 2021, **26:**5481-5503.
1000 53. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, Aken B, Akiyama JA, Jammal
1001 OA, Amrhein H, et al: **Expanded encyclopaedias of DNA elements in the human and**
1002 **mouse genomes.** *Nature* 2020, **583:**699-710.
1003 54. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F,
1004 Young N, et al: **The Genotype-Tissue Expression (GTEx) project.** *Nature Genetics* 2013,
1005 **45:**580-585.
1006 55. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS,
1007 Peter-Demchok J, Gelfand ET, et al: **A Novel Approach to High-Quality Postmortem**
1008 **Tissue Procurement: The GTEx Project.** *Biopreservation and Biobanking* 2015, **13:**311-
1009 319.
1010 56. Consortium TG, Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, Graubert A, Hadley K,
1011 Handsaker RE, Huang KH, et al: **The GTEx Consortium atlas of genetic regulatory effects**
1012 **across human tissues.** *Science* 2020, **369:**1318-1330.
1013 57. Faulkner GJ, Billon V: **L1 retrotransposition in the soma: a field jumping ahead.** *Mobile*
1014 *DNA* 2018, **9:**22.
1015 58. Dai L, Taylor MS, O'Donnell KA, Boeke JD: **Poly(A) binding protein C1 is essential for**
1016 **efficient L1 retrotransposition and affects L1 RNP formation.** *Mol Cell Biol* 2012,
1017 **32:**4323-4336.
1018 59. Taylor Martin S, LaCava J, Mita P, Molloy Kelly R, Huang Cheng Ran L, Li D, Adney
1019 Emily M, Jiang H, Burns Kathleen H, Chait Brian T, et al: **Affinity Proteomics Reveals**
1020 **Human Host Factors Implicated in Discrete Stages of LINE-1 Retrotransposition.** *Cell*
1021 2013, **155:**1034-1048.

1022 60. Goodier JL, Cheung LE, Kazazian HH, Jr: **Mapping the LINE1 ORF1 protein interactome**
1023 **reveals associated inhibitors of human retrotransposition.** *Nucleic Acids Research*
1024 2013, **41:**7401-7419.
1025 61. Gu Z, Liu Y, Zhang Y, Cao H, Lyu J, Wang X, Wylie A, Newkirk SJ, Jones AE, Lee M, et al:
1026 **Silencing of LINE-1 retrotransposons is a selective dependency of myeloid leukemia.**
1027 *Nature Genetics* 2021, **53:**672-682.
1028 62. Müller I, Moroni AS, Shlyueva D, Sahadevan S, Schoof EM, Radzisheuskaya A, Højfeldt
1029 JW, Tatar T, Koche RP, Huang C, Helin K: **MPP8 is essential for sustaining self-renewal**
1030 **of ground-state pluripotent stem cells.** *Nature Communications* 2021, **12:**3034.
1031 63. Danac JMC, Matthews RE, Gungi A, Qin C, Parsons H, Antrobus R, Timms RT,
1032 Tchasovnikarova IA: **Competition between two HUSH complexes orchestrates the**
1033 **immune response to retroelement invasion.** *Molecular Cell* 2024, **84:**2870-2881.e2875.
1034 64. Raghava Kurup R, Oakes EK, Manning AC, Mukherjee P, Vadlamani P, Hundley HA: **RNA**
1035 **binding by ADAR3 inhibits adenosine-to-inosine editing and promotes expression of**
1036 **immune response protein MAVS.** *Journal of Biological Chemistry* 2022, **298**.
1037 65. Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, Zemojtel T:
1038 **L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes.** *Nucleic*
1039 *Acids Research* 2016, **45:**D68-D73.
1040 66. Li Y, Lyu R, Chen S, Wang Y, Sun M-a: **TEENA: an integrated web server for transposable**
1041 **element enrichment analysis in various model and non-model organisms.** *Nucleic Acids*
1042 *Research* 2024, **52:**W126-W131.
1043 67. Sexton CE, Tillett RL, Han MV: **The essential but enigmatic regulatory role of HERVH in**
1044 **pluripotency.** *Trends in Genetics* 2022, **38:**12-21.
1045 68. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA,
1046 Schwartz S, Segraves R, et al: **Segmental Duplications and Copy-Number Variation in**
1047 **the Human Genome.** *The American Journal of Human Genetics* 2005, **77:**78-88.
1048 69. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith
1049 C, Scherer SW, Eichler EE, et al: **Hotspots for copy number variation in chimpanzees**
1050 **and humans.** *Proc Natl Acad Sci U S A* 2006, **103:**8006-8011.
1051 70. Lin Y-L, Gokcumen O: **Fine-Scale Characterization of Genomic Structural Variation in**
1052 **the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots.** *Genome*
1053 *Biology and Evolution* 2019, **11:**1136-1151.
1054 71. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental Duplications:**
1055 **Organization and Impact Within the Current Human Genome Project Assembly.**
1056 *Genome Research* 2001, **11:**1005-1017.
1057 72. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li
1058 PW, Eichler EE: **Recent Segmental Duplications in the Human Genome.** *Science* 2002,
1059 **297:**1003-1007.
1060 73. Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P,
1061 Maria Maggiolini FA, Harvey WT, et al: **Recurrent inversion polymorphisms in humans**
1062 **associate with genetic instability and genomic disorders.** *Cell* 2022, **185:**1986-
1063 2005.e1926.
1064 74. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA,
1065 Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al: **Transcriptome and**

1066 **genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501:**506-
1067 511.
1068 75. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network**
1069 **analysis.** *BMC Bioinformatics* 2008, **9:**559.
1070 76. Kubo S, Seleme MdC, Soifer HS, Perez JLG, Moran JV, Kazazian HH, Kasahara N: **L1**
1071 **retrotransposition in nondividing and primary human somatic cells.** *Proceedings of the*
1072 *National Academy of Sciences* 2006, **103:**8036-8041.
1073 77. Macia A, Widmann TJ, Heras SR, Ayllon V, Sanchez L, Benkaddour-Boumzaouad M,
1074 Muñoz-Lopez M, Rubio A, Amador-Cubero S, Blanco-Jimenez E, et al: **Engineered LINE-1**
1075 **retrotransposition in nondividing human neurons.** *Genome Research* 2017, **27:**335-348.
1076 78. Mita P, Wudzinska A, Sun X, Andrade J, Nayak S, Kahler DJ, Badri S, LaCava J, Ueberheide
1077 B, Yun CY, et al: **LINE-1 protein localization and functional dynamics during the cell**
1078 **cycle.** *eLife* 2018, **7:**e30058.
1079 79. Potapova T, Gorbsky GJ: **The Consequences of Chromosome Segregation Errors in**
1080 **Mitosis and Meiosis.** *Biology* 2017, **6:**12.
1081 80. Mazoyer S: **Genomic rearrangements in the BRCA1 and BRCA2 genes.** *Human Mutation*
1082 2005, **25:**415-422.
1083 81. Peixoto A, Pinheiro M, Massena L, Santos C, Pinto P, Rocha P, Pinto C, Teixeira MR:
1084 **Genomic characterization of two large Alu-mediated rearrangements of the BRCA1**
1085 **gene.** *Journal of Human Genetics* 2013, **58:**78-83.
1086 82. Wang Y, Bernhardy AJ, Nacson J, Krais JJ, Tan Y-F, Nicolas E, Radke MR, Handorf E, Llop-
1087 Guevara A, Balmaña J, et al: **BRCA1 intronic Alu elements drive gene rearrangements**
1088 **and PARP inhibitor resistance.** *Nature Communications* 2019, **10:**5661.
1089 83. Hanahan D, Weinberg Robert A: **Hallmarks of Cancer: The Next Generation.** *Cell* 2011,
1090 **144:**646-674.
1091 84. Rodić N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, Hruban RH, Iacobuzio-
1092 Donahue CA, Maitra A, Torbenson MS, et al: **Long Interspersed Element-1 Protein**
1093 **Expression Is a Hallmark of Many Human Cancers.** *The American Journal of Pathology*
1094 2014, **184:**1280-1286.
1095 85. Evering TH, Marston JL, Gan L, Nixon DF: **Transposable elements and Alzheimer's**
1096 **disease pathogenesis.** *Trends Neurosci* 2023, **46:**170-172.
1097 86. Buiting K, Nazlican H, Galetzka D, Wawrzik M, Groβ S, Horsthemke B: **C15orf2 and a**
1098 **novel noncoding transcript from the Prader–Willi/Angelman syndrome region show**
1099 **monoallelic expression in fetal brain.** *Genomics* 2007, **89:**588-595.
1100 87. Angulo MA, Butler MG, Cataletto ME: **Prader-Willi syndrome: a review of clinical,**
1101 **genetic, and endocrine findings.** *J Endocrinol Invest* 2015, **38:**1249-1263.
1102 88. Mefford HC, Eichler EE: **Duplication hotspots, rare genomic disorders, and common**
1103 **disease.** *Current Opinion in Genetics & Development* 2009, **19:**196-204.
1104 89. Makoff AJ, Flomen RH: **Detailed analysis of 15q11-q14 sequence corrects errors and**
1105 **gaps in the public access sequence to fully reveal large segmental duplications at**
1106 **breakpoints for Prader-Willi, Angelman, and inv dup(15) syndromes.** *Genome Biology*
1107 2007, **8:**R114.

1108 90. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A,
1109 Clark AG, Donnelly P, Eichler EE, et al: **A global reference for human genetic variation.**
1110 *Nature* 2015, **526:**68-74.
1111 91. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P, null n: **Variant**
1112 **calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes**
1113 **Project [version 2; peer review: 2 approved].** *Wellcome Open Research* 2019, **4**.
1114 92. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye
1115 K, Jun G, Hsi-Yang Fritz M, et al: **An integrated map of structural variation in 2,504**
1116 **human genomes.** *Nature* 2015, **526:**75-81.
1117 93. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
1118 G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011,
1119 **27:**2156-2158.
1120 94. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
1121 McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *GigaScience*
1122 2021, **10**.
1123 95. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P,
1124 de Bakker PIW, Daly MJ, Sham PC: **PLINK: A Tool Set for Whole-Genome Association**
1125 **and Population-Based Linkage Analyses.** *The American Journal of Human Genetics*
1126 2007, **81:**559-575.
1127 96. Gao F, Chang D, Biddanda A, Ma L, Guo Y, Zhou Z, Keinan A: **XWAS: A Software Toolset**
1128 **for Genetic Data Analysis and Association Studies of the X Chromosome.** *J Hered* 2015,
1129 **106:**666-671.
1130 97. Keur N, Ricaño-Ponce I, Kumar V, Matzaraki V: **A systematic review of analytical**
1131 **methods used in genetic association analysis of the X-chromosome.** *Briefings in*
1132 *Bioinformatics* 2022, **23**.
1133 98. Jurka J: **Repbase Update: a database and an electronic journal of repetitive elements.**
1134 *Trends in Genetics* 2000, **16:**418-420.
1135 99. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C,
1136 Gonzalez JN, Hinrichs Angie S, Lee Brian T, et al: **The UCSC Genome Browser database:**
1137 **2023 update.** *Nucleic Acids Research* 2022, **51:**D1188-D1195.
1138 100. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
1139 **features.** *Bioinformatics* 2010, **26:**841-842.
1140 101. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A,
1141 Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A**
1142 **knowledge-based approach for interpreting genome-wide expression profiles.**
1143 *Proceedings of the National Academy of Sciences* 2005, **102:**15545-15550.
1144 102. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP:
1145 **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 2011, **27:**1739-1740.
1146 103. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor.**
1147 *Bioinformatics* 2018, **34:**i884-i890.
1148 104. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
1149 Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2012, **29:**15-21.

1150 105. Jin Y, Tam OH, Paniagua E, Hammell M: **TEtranscripts: a package for including**
1151 **transposable elements in differential expression analysis of RNA-seq datasets.**
1152 *Bioinformatics* 2015, **31:**3593-3599.
1153 106. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu
1154 C, Wright J, Armstrong J, et al: **GENCODE reference annotation for the human and**
1155 **mouse genomes.** *Nucleic Acids Research* 2018, **47:**D766-D773.
1156 107. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for**
1157 **RNA-seq data with DESeq2.** *Genome Biology* 2014, **15:**550.
1158 108. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers**
1159 **differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic*
1160 *Acids Research* 2015, **43:**e47-e47.
1161

1162

Legends to Figures

Fig 1. Overview of the pipeline to scan for genetic variants associated with L1/Alu global singletons.

(A) An illustration of the samples used in this study. SNV and SV genetic data was available for 2503 individuals from 5 super-populations, including 660 Africans (AFR), 504 East Asians (EAS), 503 Europeans (EUR), 489 South Asians (SAS), and 347 Admixed Americans (AMR). Males and females were approximately equally represented, with male-to-female ratios (M/F ratios) ranging from 0.91 to 1.14. **(B)** A schematic illustrating the trans-ethnic integration of available SNV and SV data to identify variants associated with L1/Alu insertion global singletons. Within each super-population, samples were segregated into cases and controls depending on whether or not they harbored a global Alu or L1 insertion singleton. GWAS was carried out within each super-population to identify polymorphic SNVs and SVs associated with case-control status. Finally, GWAS results from all 5 super-populations were meta-analyzed using a random effects statistical model, yielding a summary meta-analysis odds ratio and p-value for each variant. **(C)** The frequency of Alu and L1 insertion singletons in each super-population (*left panel*) or among cases within each super-population (*middle panel*). The distribution of insertion singletons across autosomes is also shown (*right panel*). **(D)** A Manhattan plot for the trans-ethnic GWAS meta-analysis. The dashed line at p = 1.40E-5 corresponds to an average empirical FDR < 0.05, based on 20 random permutations. One such permutation is shown in the bottom panel for illustrative purposes. The solid line at p = 6.00E-6 corresponds to a Benjamini-Hochberg FDR < 0.05. The stricter of the two thresholds, p = 6.00E-6, was used to define significant SNVs and SVs. Significant variants overlapping regions in the ENCODE blacklist v2 are shown in blue and were omitted from downstream analyses. FDR: False Discovery Rate.

Fig 2. Significant SNVs lie in genomic regions containing genes involved in transposon control.

(A) Scheme for assessing whether greenlisted SNVs were enriched in regions containing genes with TE regulatory potential. For a given gene set with regulatory potential (regulatory set A or B), the proportion of SNVs near genes in that gene set were calculated for the background and significant SNV lists, and statistical significance was assessed using Fisher's exact test. **(B)** Enrichment analysis of greenlisted SNVs near genes previously implicated in L1 expression control [38] or L1 transposition control [34] by CRISPR screening in cancer cell lines. Three specific examples of greenlisted SNVs near **(C)** genes controlling L1 expression and **(D)** genes controlling L1 transposition are shown. **(E)** A summary of the associations we identified with various TE regulatory gene sets, highlighting the number of associated SNVs and the regulatory genes those SNVs were proximal to. Though not exclusive regulators of TE activity *per se*, we included in our analysis gene sets involved in "histone methyltransferase activity" and "RNA modification" functions, since those processes have been implicated in transposon control. FDR: False Discovery Rate.

Fig 3. Significant SNVs are enriched in genomic regions containing features associated with genome instability.

(A) Scheme for assessing whether greenlisted SNVs are enriched in regions containing elements known for promoting genome instability. For a given set of potentially genetically unstable regions (unstable element set A or B), the proportion of SNVs overlapping regions in each set are calculated for the background and significant SNV lists, and statistical significance is assessed using Fisher's exact test. **(B)** Enrichment analysis of greenlisted SNVs overlapping evolutionary age-stratified Alu (*left column*) and L1 (*right column*) copies. **(C)** Enrichment analysis of greenlisted SNVs overlapping curated L1 loci in L1Base v2 [65]. This database contains putatively active L1 copies (with either full-length, fully intact L1 copies or L1 copies with only ORF2 intact), as well as non-autonomous, full-length, non-intact L1 copies with regulatory potential. **(D)** Enrichment analysis of greenlisted SNVs overlapping genomic regions containing segmental duplications [71, 72] or structural variation hotspots [70]. FDR: False Discovery Rate.

Fig 4. Alterations in the cell cycle are positively correlated with case status. **(A)** Scheme for characterizing transcriptomic differences between case and control samples. Gene expression profiles were quantified using mRNA-sequencing data from lymphoblastoid cells belonging to 358 European and 86 African individuals. To note, all African individuals here were from the Yoruban population. These gene expression profiles were used to construct consensus gene co-expression networks with WGCNA. We then quantified the correlations between each module in the network and the case-control status of all samples (encoded as 0 for controls and 1 for cases). Finally, over-representation analysis (ORA) using the Gene Ontology (GO) Biological Process gene set collection was used to assign functions to significantly correlated modules. **(B)** The correlations and correlation p-values between consensus network modules and case-control status in the European and African cohorts. Boxes were color-coded according to the strength of the correlation. A meta-analysis was also carried out to summarize statistical results by combining European and African correlation p-values using Fisher's method. For visualization purposes only, the average of the European and African correlations was assigned to the meta-analysis. Correlations with opposite trends in the two cohorts were disregarded in the meta-analysis and colored grey. Correlations with $p < 0.01$ were considered statistically significant and highlighted in bold. **(C)** The top 10 ORA results for the MEroyalblue module using the GO Biological Process gene set collection. The colors represent the gene ratio (i.e. the fraction of module genes from the listed gene set) and the sizes of the dots represent the Benjamini-Hochberg FDR. NA: Not Applicable. FDR: False Discovery Rate.

Legends to Supplementary Figures

S1 Fig. Quality control of combined SNV and SV 1000 Genomes Project data. PCA plots for pruned SNV and SV genotype data from **(A)** African, **(B)** East Asian, **(C)** European, **(D)** South Asian, and **(E)** Admixed American samples. Colors and shapes represent different populations within each super-population.

S2 Fig. GWAS in individual super-populations is underpowered. Manhattan plots for GWAS results in individual super-populations, including for the **(A)** African, **(B)** East Asian, **(C)** European, **(D)** South Asian, and **(E)** Admixed American cohorts. Solid lines correspond to a Benjamini-Hochberg FDR < 0.05 and dashed lines correspond to an average empirical FDR < 0.05, based on 20 random permutations. The Benjamini-Hochberg FDR and average empirical FDR, respectively, corresponded to the following p-values in each super-population: p = 1.18E-6 and p = 4.61E-6 in the African cohort, p = 9.06E-7 and p = 8.46E-7 in the South Asian cohort, and p = 3.53E-7 and p = 1.07E-6 in the Admixed American cohort. The stricter of the two thresholds in each super-population was used to define significant SNVs and SVs. No variant at an FDR < 0.05 was identified in the East Asian and European cohorts. Significant variants overlapping regions in the ENCODE blacklist v2 are shown in blue. FDR: False Discovery Rate.

S3 Fig. Functional annotation of significant variants. **(A)** Scheme for predicting functions of genomic regions containing significant variants. All SNVs and SVs used in this study were assigned genes using the GREAT [47] online platform. Significant SNV- and SV-associated genes were then tested for functional gene set enrichment by over-representation analysis (ORA) using clusterProfiler [49], specifying the corresponding background-associated genes as the universe. **(B)** The number of genes associated to greenlisted SNVs (*left*), the distance between greenlisted SNVs and the transcriptional start sites (TSS) of associated genes (*middle*), and the top 10 ORA results for associated genes using the GO Biological Process gene set collection (*right*). The colors represent the gene ratio (i.e. the fraction of significant

SNV-associated genes from the listed gene set) and the sizes of the dots represent the Benjamini-Hochberg FDR. **(C)** The number of genes associated to greenlisted SVs (*left*) and the distance between greenlisted SVs and the transcriptional start sites (TSS) of associated genes (*middle*). **(D)** Enrichment analysis of greenlisted SNVs overlapping genomic regions with candidate cis-Regulatory Elements (cCREs) from the ENCODE Registry v4 [53]. **(E)** Heatmap comparing the median expression levels of significant SNV-associated genes in each tissue included in the GTEx Analysis v8. FDR: False Discovery Rate.

S4 Fig. Polymorphic SVs of different classes are associated with L1/Alu insertion singletons.

(A) One example of each type of significant, polymorphic SV that was associated with L1/Alu singletons. These classes included inversions, Alu insertions, an L1 insertion, SINE-VNTR-Alu (SVA) insertions, and a multiallelic copy number variant. FDR: False Discovery Rate.

Inventory of Supplementary Tables

S1 Table. Sample singleton counts, significant variant annotations, and gene co-expression network results.

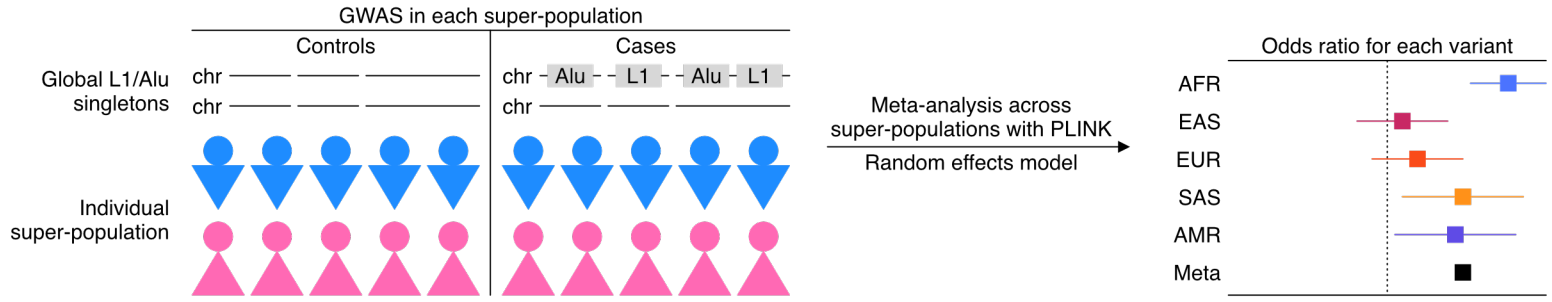
(A) Number of singletons for each sample. **(B)** All variants passing FDR < 0.05 in the GWAS meta-analysis. **(C)** Over-representation analysis of genes associated to greenlisted, significant SNVs using GO Biological Process gene sets. **(D)** Over-representation analysis of genes associated to greenlisted, significant SVs using GO Biological Process gene sets. **(E)** SnpEff annotations of significant variants. **(F)** Expression levels (median tissue-specific TPMs) of significant SNV-associated genes in the GTEx Analysis v8. The cluster of brain-associated genes is in blue, and the cluster of testes-associated genes is in orange. **(G)** TE enrichment analysis with TEENA. **(H)** Lymphoblastoid cell line WGCNA network gene-module assignments. **(I)** Over-representation analysis of MEroyalblue genes using GO Biological Process gene sets.

Figure 1

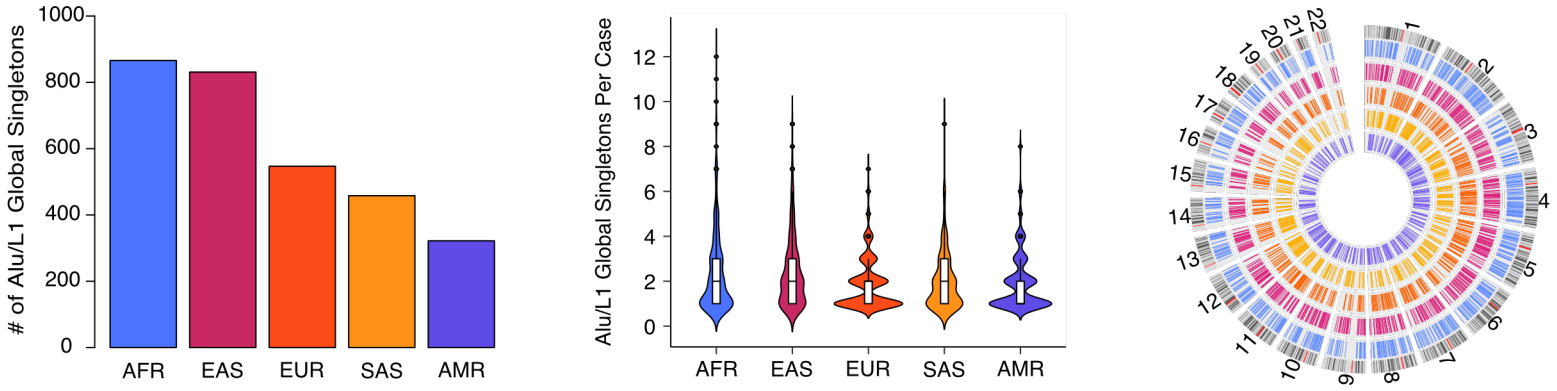
A The 2503 samples from the 1000Genomes Project used in this study

African (AFR)	East Asian (EAS)	European (EUR)	South Asian (SAS)	American (AMR)
Total = 660	Total = 504	Total = 503	Total = 489	Total = 347
M/F Ratio = 0.93	M/F Ratio = 0.94	M/F Ratio = 0.91	M/F Ratio = 1.14	M/F Ratio = 0.96
Controls = 295	Controls = 157	Controls = 211	Controls = 259	Controls = 180
Cases = 365	Cases = 347	Cases = 292	Cases = 230	Cases = 167

B Case-control study design for trans-ethnic GWAS



C Frequency and genomic distribution of L1/Alu global singletons across superpopulations



D Manhattan plot for the trans-ethnic GWAS meta-analysis associations

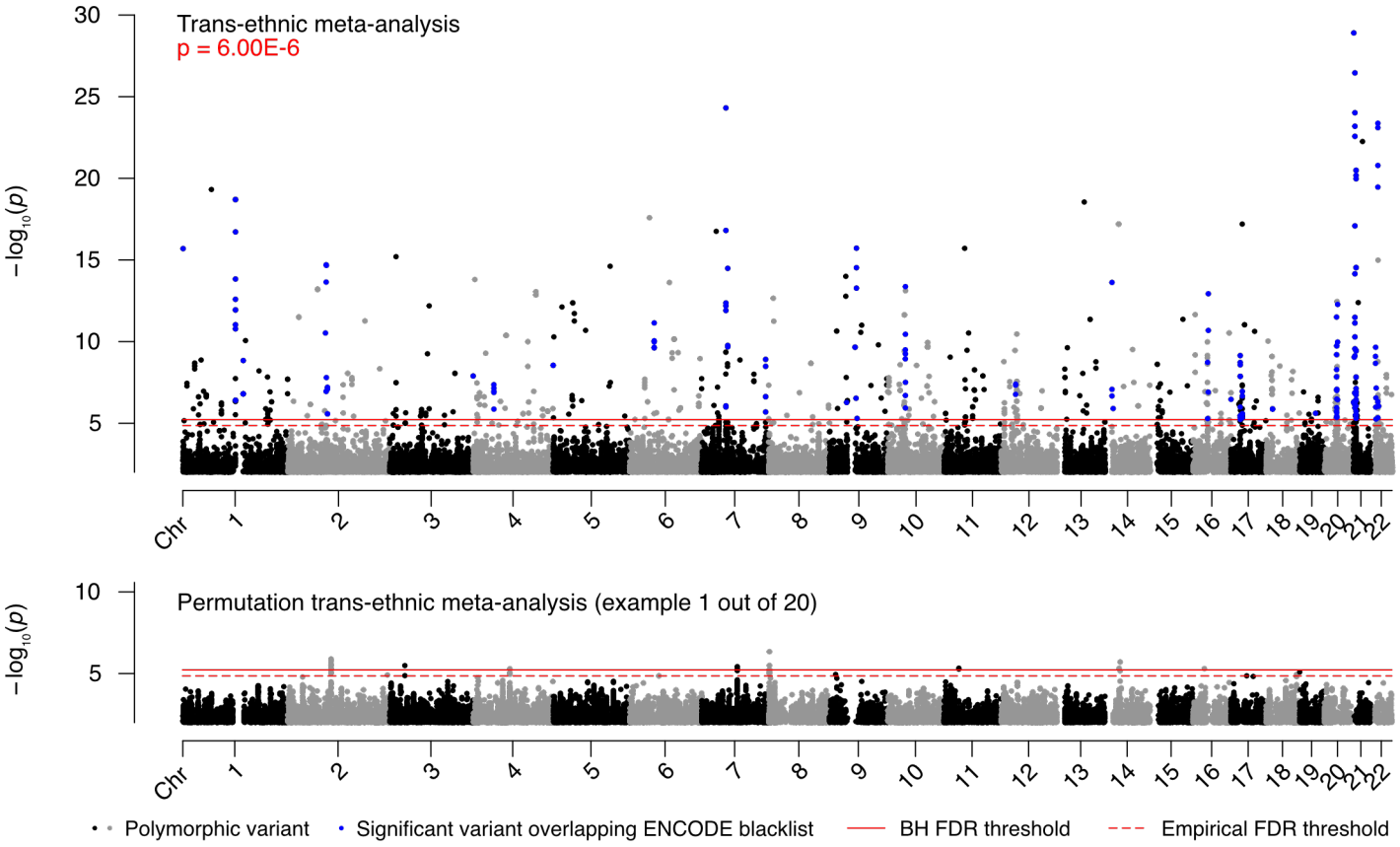
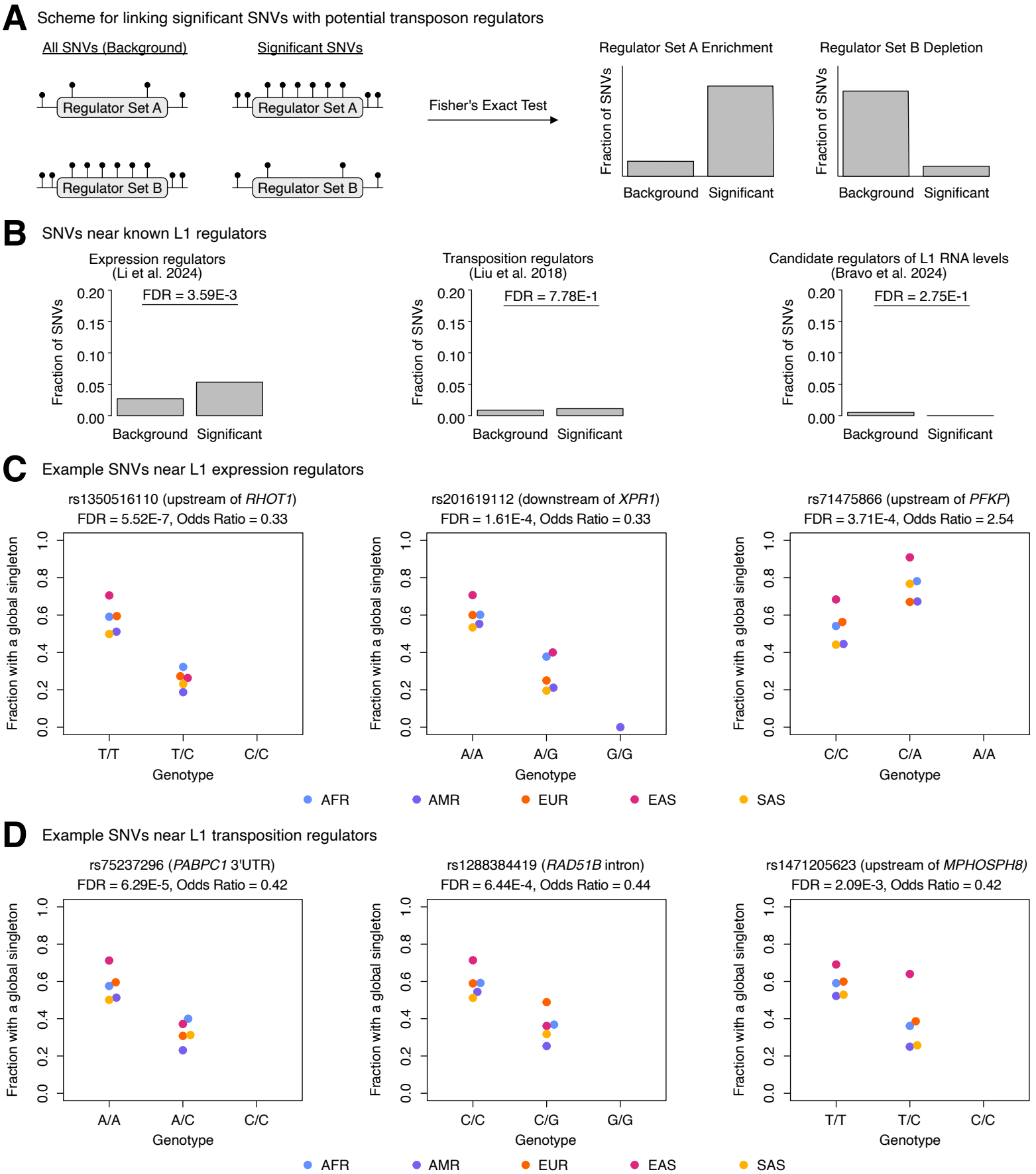


Figure 2



Regulatory gene set	Significant SNVs near genes	Genes near significant SNVs
L1 expression regulators (Li et al. 2024)	24	*IPO9, MAPK14, METTL14, MPHOSPH8, PFKP, PHF3, RBPJ, RHOT1, RRAGA, XPR1*
L1 transposition regulators (Liu et al. 2018)	5	*MPHOSPH8, PABPC1, RAD51B*
Candidate regulators of L1 RNA levels (Bravo et al. 2024)	0	
Histone methyltransferase activity GO:0042054	4	*EEF2KMT, PRDM7*
RNA modification GO:0009451	8	*A1CF, ADARB2, METTL14*

Figure 3

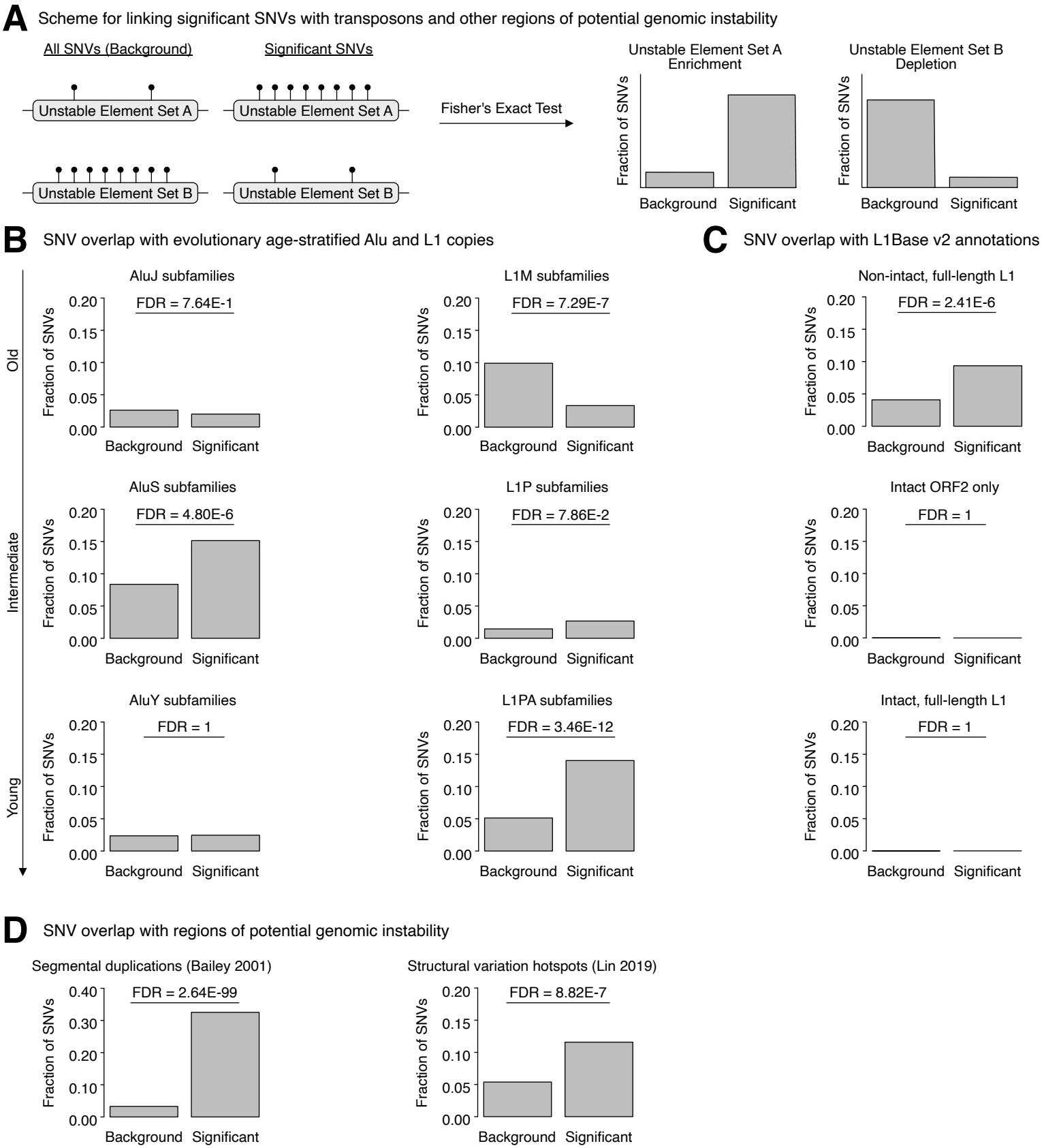
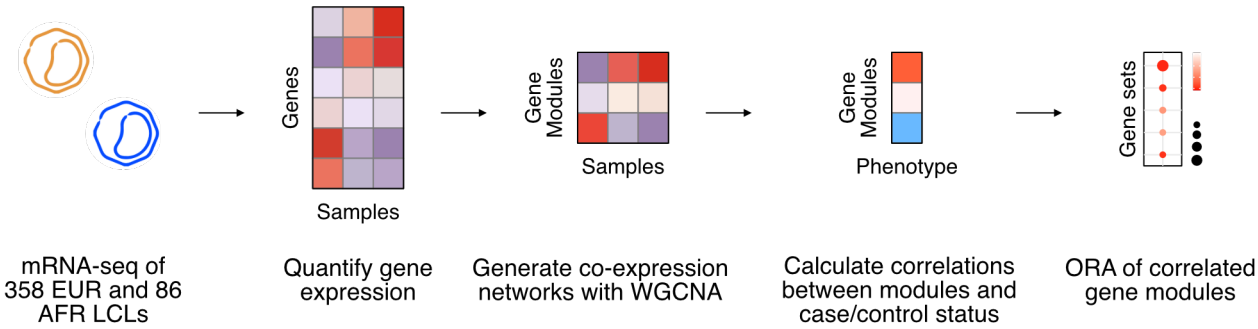

A
Scheme for linking significant SNVs with transposons and other regions of potential genomic instability
All SNVs (Background)
Significant SNVs
Unstable Element Set A
Unstable Element Set B
Fisher's Exact Test
Unstable Element Set A Enrichment
Unstable Element Set B Depletion
Fraction of SNVs
Background
Significant
B
SNV overlap with evolutionary age-stratified Alu and L1 copies
Old
Intermediate
Young
AluJ subfamilies
FDR = 7.64E-1
L1M subfamilies
FDR = 7.29E-7
AluS subfamilies
FDR = 4.80E-6
L1P subfamilies
FDR = 7.86E-2
AluY subfamilies
FDR = 1
L1PA subfamilies
FDR = 3.46E-12
C
SNV overlap with L1Base v2 annotations
Non-intact, full-length L1
FDR = 2.41E-6
Intact ORF2 only
FDR = 1
Intact, full-length L1
FDR = 1
D
SNV overlap with regions of potential genomic instability
Segmental duplications (Bailey 2001)
FDR = 2.64E-99
Structural variation hotspots (Lin 2019)
FDR = 8.82E-7

Figure 4

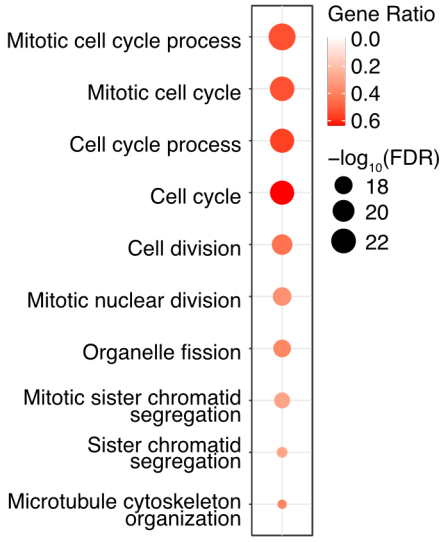
A Scheme for assessing transcriptomic differences between cases and controls



B Network module correlations with case/control status

	EUR case status	AFR case status	Meta-analysis
MEblack	0.077 (0.1)	0.056 (0.6)	0.066 (0.3)
MEpurple	0.083 (0.1)	0.057 (0.6)	0.07 (0.3)
MEgreenyellow	0.039 (0.5)	0.09 (0.4)	0.064 (0.5)
MEmidnightblue	−0.0024 (1)	0.12 (0.3)	NA (NA)
MEsalmon	−0.12 (0.03)	0.09 (0.4)	NA (NA)
MElightyellow	0.084 (0.1)	0.04 (0.7)	0.062 (0.3)
MEcyan	−0.039 (0.5)	0.11 (0.3)	NA (NA)
MEred	−0.057 (0.3)	0.23 (0.03)	NA (NA)
MElightcyan	0.017 (0.8)	−0.13 (0.2)	NA (NA)
MEroyalblue	0.035 (0.5)	**0.37 (4E−4)**	**0.2 (0.002)**
MEblue	0.036 (0.5)	0.18 (0.1)	0.11 (0.2)
MEbrown	−0.021 (0.7)	0.15 (0.2)	NA (NA)
MEturquoise	−0.087 (0.1)	−0.12 (0.3)	−0.1 (0.1)
MEgrey60	−0.022 (0.7)	0.08 (0.5)	NA (NA)
MEmagenta	−0.03 (0.6)	−0.14 (0.2)	−0.087 (0.3)
MEgreen	0.0054 (0.9)	−0.14 (0.2)	NA (NA)
MEtan	0.046 (0.4)	−0.0054 (1)	NA (NA)
MEyellow	0.048 (0.4)	−0.15 (0.2)	NA (NA)
MElightgreen	0.049 (0.4)	−0.048 (0.7)	NA (NA)
MEpink	0.074 (0.2)	−0.15 (0.2)	NA (NA)
MEgrey	−0.052 (0.3)	0.014 (0.9)	NA (NA)

Correlation (p-value)

Correlation: −1 0 1

C ORA of the MEroyalblue module with the GO Biological Process gene sets