

Host-gene expression across human tissues reveals cargo-specific intronic outcomes of L1-mediated retrotransposition

Min-Gyu Lee

leemg@korea.ac.kr

KOREA university

Research Article

Keywords: LINE-1, Alu, SVA, processed pseudogene, retrotransposition, intronic sequence burden, host-gene expression, GTEx, genome evolution

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10975043/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

LINE-1 (L1) machinery mobilizes L1, Alu, SVA and cellular mRNAs, but whether these cargoes show similar relationships with host-gene transcription remains unclear.

Results

Using the Genotype-Tissue Expression (GTEx) atlas as a reference for gene expression in adult human tissues, we grouped host genes into expression quartiles within each of 51 anatomical tissues. We analysed the same catalogue of inherited intronic insertions across these quartiles. We compared these patterns with those observed for reference-retained, somatic and experimentally recovered insertions. Across tissues, the median ratios of event density in the highest versus lowest expression quartiles (Q4/Q1) were 0.539 for L1, 1.112 for Alu and 2.255 for SVA. The corresponding ratios for inserted-sequence burden were 0.504, 1.127 and 2.452. Adjustment for intronic opportunity, host architecture and native intronic GC reduced L1 event depletion to a borderline association. L1 sequence burden remained depleted, and direct cargo contrasts remained significant. Cargoes also differed in occupancy, multiplicity and orientation. Reference-genome analyses distinguished event persistence from sequence persistence and included processed pseudogenes. Somatic L1 and published recovered L1–Alu maps showed aspects of this divergence without germline transmission.

Conclusions

Using host-gene expression as a common reference clarifies how cargoes sharing L1-associated machinery differ in their intronic distributions. Integrating event density, sequence burden and orientation separates differences in insertion abundance from those in retained sequence, revealing population cargo contrasts that measured GC content and host architecture do not fully explain.

Background

LINE-1 (L1) is the only autonomous non-long-terminal-repeat retrotransposon active at substantial levels in the human genome. Its proteins preferentially mobilize their encoding L1 RNA *in cis*, but can also support retrotransposition of other RNAs *in trans*. Alu and SVA use L1-associated machinery, whereas reverse transcription of cellular mRNAs can generate processed pseudogenes (PPs) [1–4]. These substrates share access to L1-encoded activities, including the endonuclease and reverse-transcriptase activities of ORF2p, while differing in sequence composition, length, RNA structure and interactions with proteins. We use “cargo” for the RNA-derived sequence mobilized through this machinery, while recognizing that the associated ribonucleoprotein complexes and requirements for L1 proteins can differ.

The genomic distributions of these cargoes suggest different relationships with transcription. Highly expressed domains tend to be GC rich, SINE rich and LINE poor. Alu and L1 show different associations with expression breadth and level, and intragenic L1 content decreases with increasing host-gene expression [5–8]. Insertion-associated biases, recombination and post-insertion selection can all contribute to these distributions [9–13]. Orientation and sequence content provide additional constraints. Intronic transposable elements are distributed non-uniformly near exons, and their orientation biases vary with gene context [14, 15]. L1 sequence can disrupt elongation, promote premature polyadenylation and alter splicing, while Alu introduces distinct opportunities and liabilities for RNA processing [16–24]. De novo L1 mapping studies implicate target sequence, replication and DNA repair in the distribution of recovered insertions [25–28]. Polymorphic retrotransposons are associated with variation in gene expression [29–31], and functional studies show that individual SVA elements can regulate gene expression [32–34]. SVA also has distinctive mobilization and evolutionary properties [35–38], while cellular-mRNA retrocopies contribute inherited and somatic structural variation [39–42].

A central question is whether L1-mediated cargoes show similar intronic distributions when compared across the same levels of host-gene expression, with adjustment for intronic opportunity and host context. Event number alone is insufficient. A short retained fragment and a long insertion each contribute one event, although they expose the host transcript to different amounts of inserted sequence. A higher antisense fraction is also ambiguous because it can result from fewer sense events, more antisense events, or both. We therefore compared event density, sequence burden, host occupancy, the number of events per positive host and absolute densities in each orientation within the same expression strata. This design allowed us to distinguish these outcomes while examining how they varied together across cargoes.

Tissue-dependent expression provides a way to examine these relationships across different transcriptional contexts. An inherited insertion is present in the differentiated cells of its carrier, but its host gene may be weakly expressed in one tissue and highly expressed in another. The same sequence may therefore encounter different transcriptional and RNA-processing demands. Tissue-wide expression rankings can capture this variation, although present-day adult expression cannot reconstruct the developmental exposure or selection history of an individual insertion.

We used the Genotype-Tissue Expression (GTEx) atlas [43] as a reference for gene expression across adult human tissues. Within each of 51 anatomical tissues, we ranked host genes independently while holding a long-read 1000 Genomes Project (1KGP) insertion catalogue [44] fixed. This design describes how inherited insertions are distributed across tissue-specific expression strata. Because GTEx and 1KGP are not carrier matched, it does not measure insertion effects in carriers or directly observe transcriptional filtering. We also examined reference-retained L1, Alu, SVA and PPs, then compared the results with Pan-Cancer Analysis of Whole Genomes (PCAWG) somatic insertions [45] and published experimentally recovered L1 and Alu maps [25, 46]. The somatic and recovered maps allowed us to ask whether aspects of cargo divergence occur without germline transmission and subsequent organism-wide development. All analyses used previously generated public data.

Results

Inherited cargoes differ in their distributions across host-gene expression levels

The primary population catalogue comprised 1,923 L1, 11,069 Alu and 1,576 SVA insertions represented in the tissue-quartile analyses (Additional file 1: Figs. S1 and S4, Additional file 2: Tables S1 and S3). We retained only unambiguous cargo assignments within protein-coding gene introns. Genes were ranked independently within each of 51 GTEx anatomical tissues and divided into expression quartiles Q1–Q4. Insertion coordinates remained fixed, and event counts and inserted base pairs were normalized to eligible protein-coding intronic sequence in each stratum. Each tissue therefore provided a different expression ranking of the same insertion catalogue.

Across tissues, the median ratios of event density in Q4 versus Q1 were 0.539 for L1, 1.112 for Alu and 2.255 for SVA. The corresponding ratios for inserted-sequence burden were 0.504, 1.127 and 2.452 (Fig. 1a,b, Additional file 1: Fig. S4g). L1 density was lower in Q4 than Q1 in all 51 tissues, whereas Alu and SVA densities were higher in all 51 (Additional file 1: Fig. S1). Inserted sequence per event changed less than event abundance (Fig. 1c), indicating that the raw population patterns were driven mainly by event abundance, with a smaller contribution from length. Models treating quartiles as ordered groups and directly comparing expression associations between cargoes supported the same separation (Fig. 1d, Additional file 1: Fig. S4h). Tissue-specific Q4/Q1 distributions are shown in Additional file 1: Fig. S4a–f. Sensitivity analyses excluding first introns or intronic sequence within 10 kb of the TSS are shown in Additional file 1: Fig. S3a.

The expression assignments varied substantially across tissues. In the 18,236-gene GENCODE-matched background used for the descriptive matrix in Fig. 1e,f, 13.74% of genes remained in one quartile and 16.30% occupied all four. Among 5,918 mapped 1KGP host genes, 91.60% changed quartile at least once, including 96.21% of L1 hosts, 91.92% of Alu hosts and 91.58% of SVA hosts. When individual insertions were counted, 96.62% of mapped L1 insertions, 93.98% of Alu insertions and 92.26% of SVA insertions were linked to hosts occupying at least two quartiles (Additional file 1: Fig. S3e). Most hosts therefore occupied different expression strata across tissues. The repeated rankings assess consistency across these contexts, although they do not represent 51 independent insertion cohorts.

Differences between cargoes persist after adjustment, with stronger depletion of L1 sequence burden

Expression rank covaried with host sequence composition and architecture. Across tissues, the median intronic GC content, weighted by base pairs, increased from 40.25% in Q1 to 43.80% in Q4. The within-tissue Q4–Q1 difference was positive in all 51 tissues, with a median of 3.54 percentage points (Fig. 2a). Gene span, intron architecture and the amount of intronic sequence near the transcription start site

(TSS) also varied with expression. We fitted gene-level Poisson pseudo-maximum-likelihood (PPML) models with eligible intronic sequence as an exposure offset, adding host architecture and native intronic GC to the expression model.

After full adjustment, the estimated difference in L1 event count between Q4 and Q1 was smaller, with an incidence-rate ratio (IRR) of 0.871 (95% confidence interval [CI] 0.756–1.003, false discovery rate [FDR] = 0.054). The event-count association fell just outside the FDR < 0.05 threshold. L1 inserted-sequence burden nevertheless remained depleted (IRR = 0.763, 95% CI 0.639–0.909, FDR = 0.0031). Alu remained positively associated with expression for event count (IRR = 1.192, 95% CI 1.126–1.261) and inserted sequence (IRR = 1.202, 95% CI 1.133–1.274). SVA also showed positive associations for event count (IRR = 1.503, 95% CI 1.244–1.816) and inserted sequence (IRR = 1.615, 95% CI 1.285–2.029) (Fig. 2b). The reduction in the event-count effect thus left a stronger residual difference in L1 sequence burden.

Direct cargo contrasts remained significant after adjustment. For event count, the estimated Q4/Q1 rate ratio was 1.369-fold greater for Alu than L1, 1.726-fold greater for SVA than L1 and 1.261-fold greater for SVA than Alu. For inserted sequence, the corresponding ratios of rate ratios were 1.576, 2.118 and 1.344 (all FDR $\leq$ 0.023, Fig. 2c). The joint models allowed architecture and GC effects to vary separately for L1, Alu and SVA. Sensitivity analyses using grouped models preserved the ordering of cargo effects. This held after excluding first introns or intronic sequence within 10 kb of the TSS, across 51 refits that each omitted one tissue, and in 500 bootstrap resamples using chromosomes as blocks (Fig. 2d–f, Additional file 1: Fig. S3c).

Local sequence context varied by cargo, orientation and window size (Fig. 2g–j). In the broader 100-bp and 500-bp flanks of L1 insertion sites in the host reference genome, GC content generally increased toward Q4 as L1 abundance decreased. The immediate 10-bp and 20-bp profiles were less uniform and included non-monotonic trajectories. Alu showed rising GC and abundance, whereas SVA abundance increased despite variation in the local GC patterns. These profiles are consistent with the adjusted models, in which measured GC explains part of the pattern but leaves substantial cargo separation. Local target preferences remain a possible contributor.

Host occupancy and changes in sense and antisense density distinguish the cargoes

We separated host occupancy, the fraction of eligible background genes containing a focal cargo, from multiplicity, the number of events per cargo-positive host. Across tissues, the median ratios of occupancy and multiplicity in Q4 versus Q1 were 0.46 and 0.73 for L1, 1.13 and 0.63 for Alu, and 1.42 and 1.01 for SVA (Fig. 3a). L1 was represented in fewer hosts and by fewer events per positive host toward Q4. Alu occupied a broader fraction of hosts while multiplicity decreased, whereas SVA combined broader occupancy with approximately maintained multiplicity. Occupancy and multiplicity use gene-based denominators, whereas event density uses intronic base pairs. Additional analyses

examined hosts with high Alu multiplicity and measured density within introns already containing the focal cargo (Additional file 1: Figs. S8a–c and S9).

Across tissues, the median antisense fraction for L1 increased from 0.540 in Q1 to 0.607 in Q4, although both sense and antisense densities decreased. The decrease was larger for sense L1. Alu sense density remained close to its Q1 level while antisense density increased. SVA density increased in both orientations, more strongly in antisense (Fig. 3b–d). The rise in antisense fraction therefore reflected different absolute changes across cargoes. For L1, it accompanied a reduction in antisense insertions per unit intronic sequence.

In adjusted models of individual events, the odds ratios for antisense orientation in Q4 versus Q1 were 1.353 for L1, 1.282 for Alu and 1.530 for SVA (Fig. 3e). Direct differences between cargoes in these changes did not remain significant after FDR correction. Rates estimated separately for each orientation, with quartiles treated as ordered groups, are shown in Additional file 1: Fig. S5a. Models testing whether L1 orientation varied jointly with expression and proximity to an exon boundary also found no significant interaction after correction for multiple testing, including among insertions at least 6 kb long (Additional file 1: Fig. S5b–e, Additional file 2: Table S2). The absolute changes in sense and antisense density thus differed across cargoes, while direct comparisons of the change in antisense odds remained statistically unresolved.

Reference retention separates event representation from surviving sequence

We reconstructed reference-retained L1, Alu and SVA insertion events in GRCh38 and added GENCODE-annotated PPs as cellular-mRNA-derived retrocopies. The all-intronic reference sets contained 207,208 L1, 529,144 Alu and 2,080 SVA events, together with 3,256 annotated intronic PPs (Additional file 2: Tables S1 and S3). We used the same definition of current intronic sequence as the denominator for all cargoes. Across tissues, the median ratios of event density in Q4 versus Q1 were 0.87 for L1, 2.00 for Alu, 2.08 for SVA and 1.25 for PP. The corresponding ratios for retained-sequence burden were 0.60, 1.97, 2.27 and 1.02 (Fig. 4a,b, Additional file 1: Fig. S2). Retained sequence per L1 event decreased toward Q4, while antisense representation increased across all four cargoes (Fig. 4c–e).

PP event number, host occupancy and annotated sequence span showed different expression-associated patterns. Across tissues, the median Q4/Q1 ratios were 0.864 for the fraction of hosts containing a PP, 1.253 for event density, 1.017 for annotated-span burden and 0.836 for annotated span per event. After host-architecture adjustment, PP occupancy (odds ratio [OR] = 1.17) and event count (IRR = 1.12) did not remain significant after FDR correction (both FDR = 0.062). Annotated-span burden remained near parity (IRR = 0.94, FDR = 0.51), whereas antisense orientation remained increased (OR = 1.35, FDR = 0.021, Fig. 4f–h). These PP models adjusted for host architecture and tissue, but did not include adjustment for native intronic GC, which was included in the population analysis of the three TE cargoes. PP annotated span measures the present-day retained locus and may differ from the original reverse-transcribed mRNA length. Reference controls, denominator comparisons and PP heterogeneity are provided in Additional file 1: Figs. S3b and S6.

L1 evolutionary strata retain a distinction between events and sequence burden

The relationship between event density and host-gene expression differed among evolutionary groups of reference L1. Across tissues, the median ratios of event density in Q4 versus Q1 were 0.420 for L1HS/L1PA2, 0.396 for L1PA3–8, 0.380 for L1PA9–17, 0.476 for L1PB and 0.497 for L1MA, while older L1M approached parity (1.063). Retained-sequence burden remained below unity in every stratum, including older L1M (0.831), and ranged from 0.272 to 0.353 in the younger and intermediate groups (Fig. 5a,b,d,e). Older annotations therefore approached parity in number while retaining less L1 sequence in highly expressed hosts.

Antisense changes differed across L1 strata without a monotonic progression (Fig. 5c,f). Alu and SVA subfamilies also varied within their broader cargo classes (Additional file 1: Fig. S7), with additional antisense trajectories shown in Additional file 1: Fig. S8d,e. These comparisons concern surviving annotations with different evolutionary histories. They cannot estimate a rate of sequence loss or distinguish insertion-time truncation from later deletion and fragmentation.

Somatic and recovered insertions show cargo divergence without germline transmission

The PCAWG analysis linked somatic insertions to donor-matched tumor expression. Across Q1–Q4, 2,285 L1 events were represented, compared with 29 Alu and 39 source-designated PP events (Additional file 2: Tables S1, S3 and S4). L1 density decreased from 0.002906 events Mb⁻¹ in Q1 (942 events) to 0.000602 events Mb⁻¹ in Q4 (156 events), giving a raw Q4/Q1 ratio of 0.207. With cancer type included and covariance clustered by donor, the per-quartile IRR was 0.580 (95% CI 0.548–0.614), and the Q4/Q1 IRR was 0.225 (95% CI 0.187–0.269, $P = 8.76 \times 10^{-59}$, Fig. 6a,b). The estimate changed little after donor trimming, leave-one-cancer-type-out analysis and the prespecified event filters (Additional file 1: Fig. S10). Formal pan-cancer inference was restricted to L1 because the other cargoes were too sparse for an equivalent comparison.

Among somatic L1 events with reported lengths passing the sanity filter, 744 of 942 Q1 events and 117 of 156 Q4 events had resolved lengths, corresponding to 79.0% and 75.0% coverage. For this subset, the ratio of sequence burden in Q4 versus Q1 was 0.273, but mean resolved length per event was 1.391-fold higher in Q4. Lower somatic L1 burden therefore coexisted with longer mean insertions among Q4 events with resolved lengths (Additional file 1: Fig. S10). Incomplete length coverage prevents extension of this comparison to all events, and the reasons for missing lengths remain unknown.

Published recovered insertion maps provided an additional comparison using a common HeLa expression background and common intron annotation. The maps contained 7,946 HeLa L1 and 109 HeLa Alu events. An additional 8,535 PA-1 L1 events provided secondary support (Additional file 2: Tables S3 and S4). Recovered HeLa L1 density was lower in Q4 than Q1 (rate ratio = 0.687, 95% CI

0.641–0.737), whereas Alu density was 4.373-fold higher (95% CI 2.216–8.630, Fig. 6c). The decrease in recovered L1 was greater for sense events, while recovered Alu increased in both orientations (Fig. 6d). Dividing the Q4/Q1 rate ratio for Alu by that for L1 gave a ratio of 6.361 (95% CI 3.211–12.599, FDR = 3.36×10^{-7} , Fig. 6e).

Although the L1 and Alu maps were analyzed with a common expression background, they came from separate experiments with different recovery protocols. This comparison provides support across studies but does not constitute a controlled comparison of the two cargoes. The published tables did not report inserted lengths for individual events, so an equivalent sequence-burden analysis was not possible (Additional file 1: Fig. S11). We compared expression-associated effects within each layer rather than absolute densities between layers (Fig. 6f). The somatic and recovered maps show that aspects of cargo divergence can arise without germline transmission, although both remain subject to cell growth, selection and ascertainment.

Discussion

A shared mobilization route produces distinct intronic landscapes

Stratifying insertions by host-gene expression brought several aspects of cargo divergence into a common comparison. Within the same expression quartiles, we could distinguish how often a cargo occurred, how much sequence it contributed, how widely it was distributed among hosts and how its sense and antisense densities changed. These measures revealed different combinations of intronic outcomes for L1, Alu and SVA, even though all three use L1-associated machinery. Direct contrasts in event count and sequence burden persisted after adjustment for intronic opportunity, host architecture and native intronic GC. Reference-retained PPs extended the comparison to cellular-mRNA-derived cargo, and somatic and recovered-insertion datasets showed which density patterns recurred in other biological settings.

The 51 tissue-specific rankings extend earlier observations that expressed regions differ in repeat composition by comparing several insertion properties as host genes change expression groups across tissues. Most insertion hosts changed quartile across tissues, yet the ordering of the L1, Alu and SVA density associations persisted. The pattern therefore extended beyond a single expression summary or a fixed division of genes into expression groups. Examining the same catalogue under these different rankings connected cargo differences in abundance, sequence burden and orientation to a common host-expression context. GTEx was used here as a reference atlas of host-gene expression across adult human tissues, providing tissue-specific transcriptional contexts for inherited alleles. It cannot measure their developmental histories, effects in carriers or probabilities of elimination, and gene-level confounding remains possible.

GC contributes to the pattern but does not fully explain cargo separation

GC is closely associated with both gene expression and retrotransposon distribution. Insertion bias, selection and recombination have all been implicated in these relationships [5, 9–12, 25–28, 47]. Consistent with these contributions, adjustment for host architecture and GC substantially reduced the depletion of L1 events seen in the unadjusted analysis. The adjusted association no longer met the $FDR < 0.05$ threshold. However, L1 sequence burden remained depleted and direct cargo contrasts remained significant. The residual pattern therefore concerns both the amount of L1 sequence retained and its difference from other cargoes.

Insertion number and the amount of inserted sequence capture different biological outcomes. Expression strata with similar event densities can differ in insertion lengths and sequence composition. In the population analysis, L1 sequence burden showed a stronger adjusted association with expression than event count. Older reference L1M approached parity in event density but retained lower sequence burden. Among somatic events with resolved lengths, Q4 had fewer events per unit intronic sequence but longer mean reported insertions. Event count alone would miss these differences. Their origins remain uncertain because insertion-time truncation, later sequence loss and differential recovery can all affect the observed lengths.

The local GC profiles also differed with sequence scale. The overall composition of host introns and the sequence immediately flanking insertion sites followed different trajectories across cargoes and orientations. Measured GC and architecture therefore explain only part of the comparative pattern, leaving room for unmeasured sequence preferences and other host features. This conclusion applies to the adjusted population comparison. The PP models included host architecture but lacked an equivalent GC adjustment.

Cargo sequence and orientation may contribute to intronic compatibility

Sharing L1-associated machinery does not require cargoes to have identical insertion-site or retention distributions. L1 RNA is preferentially mobilized *in cis*, whereas Alu, SVA and cellular mRNAs access the machinery *in trans* through different RNA and protein interactions [1–4, 35–38]. RNA structure, poly(A)-associated features, ribonucleoprotein assembly and engagement with ORF2p may affect which intermediates form and are recovered. Such differences could contribute to cargo-dependent integration. The present data, however, cannot isolate RNA sequence as a cause of altered ORF2p target choice. The compared insertions also need not have originated from the same source L1 or identical protein complexes.

After integration, inserted sequence can affect host transcription and RNA processing. Long L1 sequences, premature polyadenylation signals and splice-related features can interfere with host

transcripts [14, 16–21]. Sense L1 density decreased more strongly than antisense L1 density, consistent with the effects discussed above. Thus, the higher antisense proportion reflected a greater reduction in sense L1 density, even though both orientations became less frequent. Alu and SVA also showed higher antisense proportions, but their underlying changes in sense and antisense densities differed. After correction, the expression-associated change in antisense odds did not differ significantly between cargoes. We also found no significant interaction between expression and proximity to an exon boundary. These results highlight the need to interpret orientation proportions alongside absolute insertion densities and host-gene context.

These analyses identify associations between host-gene expression and intronic insertion patterns but do not establish the underlying mechanisms. Linking individual insertions to transcript structure and following their persistence over time could help distinguish effects on integration from those acting after insertion.

Complementary insertion datasets capture different biological histories

Intronic distributions can reflect processes acting before, during and after integration. Heritable alleles also pass through development, transmission and population history, while reference sequence can be altered by deletion, fragmentation and other evolutionary processes. Some mobile-element fragments are conserved near developmental genes [48], indicating that retained sequence can have roles beyond neutrality or avoidance of damage. Here, “compatibility” describes the resulting representation of each cargo in a host context. It is a descriptive concept, not a measured fitness parameter or evidence that all retained insertions benefit their host.

PPs illustrate these distinctions because event representation, host occupancy, annotated span and orientation followed different trajectories. GENCODE PP loci are present-day annotations whose boundaries may differ from the original cDNA insertion. Subsequent duplication can also add retrocopy loci without additional retrotransposition events. PPs thus provide a useful cellular-mRNA-derived comparison, but their annotation and event definitions differ from those of reconstructed reference TEs. Annotated PP counts cannot recover the total number of historical mRNA retrotransposition events.

Somatic and recovered insertions show that some of the divergence occurs without germline transmission. PCAWG showed pronounced L1 depletion with increasing donor-matched tumor expression, while the recovered HeLa maps showed opposing L1 and Alu density changes. These results complement studies of the heterogeneous activity and evolutionary consequences of L1 in cancer [45, 49, 50]. The timing of depletion remains uncertain. Tumors undergo clonal evolution, and recovery assays detect events after growth and selection under their particular protocols. A common HeLa annotation and expression background also cannot remove differences in source constructs, experimental procedures or detection sensitivity between the L1 and Alu studies.

The population, reference, somatic and recovered datasets differ in event age, genome assembly, expression source and detection method. They are complementary observations rather than a time series of the same insertions. Somatic and recovery datasets cover more restricted biological histories, but they do not provide a shared clock or an estimated rate of change. Their main contribution is to identify patterns that recur across settings, alongside the longer histories represented by inherited and reference-retained insertions.

Scope and remaining limitations

GTEx provided a common reference for comparing insertions across host-gene expression levels in adult human tissues. These expression profiles do not capture the developmental history of individual insertions, and retained insertions may themselves influence present-day expression [29–34]. Because the same insertion catalogue was evaluated across tissues, the tissue-specific results describe consistency across expression contexts rather than independent replication. Models accounting for repeated observations and sensitivity analyses using chromosome resampling and tissue omission supported the main comparisons.

Differences in annotation and data coverage also affect comparisons between insertion datasets. TE and PP loci were assigned using different criteria, while sparse non-L1 somatic events and incomplete length information limited some analyses. Adjustment for host architecture and intronic GC addressed measured differences among hosts, although other genomic features may contribute to the observed associations. Within this scope, comparing cargoes across the same expression strata revealed differences in event density, sequence burden and orientation that would be less apparent from any single measure.

Conclusions

Using GTEx as a reference for gene expression across adult human tissues, we found that stratifying intronic insertions by host-gene expression revealed a consistent separation of L1, Alu and SVA across tissue-specific rankings. Comparing event density, sequence burden, host occupancy, multiplicity and orientation within the same expression quartiles showed distinct combinations of intronic outcomes among cargoes that share L1-associated machinery. This joint comparison exposed differences that event counts or antisense fractions alone would miss. In the population analysis, direct cargo contrasts persisted after adjustment for measured intronic GC and host architecture, and L1 sequence burden remained depleted when its event-count association was borderline. Reference-retained insertions further separated event representation from sequence burden, while somatic and recovered maps showed aspects of density divergence without germline transmission. These public-data analyses do not establish the causal role of transcription or distinguish integration bias from post-insertion selection. They show that host-expression stratification brings cargo-specific differences in insertion abundance, sequence burden and orientation into an integrated view of intronic compatibility.

Methods

Study design and data sources

We reanalyzed public insertion catalogues, genome annotations and gene-expression resources to compare L1-mediated cargoes across groups of host genes defined independently by expression level. For inherited and reference-retained insertions, GTEx supplied the reference expression profiles of adult human tissues used to rank host genes. These profiles defined the tissue-specific expression context of the human host genome for the insertion comparisons. Outcomes included event number, inserted or retained sequence, host occupancy and orientation. No new sequencing, cell-culture or animal experiments were performed, and no participants were recruited. Sample sizes were determined by the eligible observations available in each resource, without a prospective sample-size or power calculation. Dataset roles and analysis sizes are summarized in Additional file 2: Tables S1, S3 and S4.

Population-polymorphic events were obtained from the long-read 1KGP structural-variation resource [44], using the SVAN v1.3 annotated unphased variant call format (VCF) file in the native T2T-CHM13v2.0 (hs1) assembly and its corresponding RefSeq gene annotation. Reference-retained elements were analyzed in GRCh38 with GENCODE v48. PCAWG events [45] and published recovered L1 and Alu maps [25, 46] were analyzed in GRCh37 with GENCODE v19. Coordinates were not pooled across genome builds. Comparisons between datasets used the associations with expression estimated separately within each dataset for protein-coding introns.

GTEx v10 bulk RNA-seq expression and sample annotations defined the tissue-specific strata for population and reference analyses [43, 51]. Annotations from the corresponding GENCODE releases [52, 53] and UCSC genome and repeat resources [54, 55] supplied the corresponding host and reference-repeat definitions. Donor-matched PCAWG tumor expression was obtained from ArrayExpress/Expression Atlas accession E-MTAB-5200 [56]. Human Protein Atlas (HPA) cell-line expression supplied the external expression background for the recovered-insertion maps [57]. The population sequence resource is available through the International Genome Sample Resource (IGSR) and European Nucleotide Archive accession PRJEB89727 [44, 58].

Protein-coding genes, exon-union introns and eligible opportunity

Protein-coding gene models were reconstructed independently for each assembly. Exons from all annotated transcripts of a gene were merged into a non-overlapping exon union, which was subtracted from the gene span to define eligible intronic sequence. Genes with no remaining intronic sequence were excluded. This conservative definition removes any region annotated as exonic in at least one transcript, so it does not represent the introns of a specific isoform. Source coordinates were converted before annotation, and interval operations used zero-based, half-open coordinates.

The gene models supplied gene span, the mean length and number of introns remaining after exon union, the number of merged exon intervals, the first-intron fraction of total intronic sequence and the fraction of intronic sequence within 10 kb of the TSS. Unless specified otherwise, the background for each expression stratum comprised all eligible protein-coding genes with an assigned expression quartile, including genes with no focal insertion. The denominator was the sum of eligible intronic base pairs contributed by those genes. This denominator represents available intronic sequence, independent of the number of sampled cells, chromosomes or carriers.

GTEX tissue policy, expression quartiles and host re-ranking

GTEX v10 supplied 54 RNA-seq tissue/cell categories in the analyzed resource. A prespecified minimum of 20 samples per category excluded Kidney - Medulla (11 samples), leaving 53 eligible categories. Excluding Cells - Cultured fibroblasts and Cells - EBV-transformed lymphocytes then yielded the 51 anatomical tissues used in all primary GTEX-based analyses. The pipeline records category counts and exclusion reasons for this 54→53→51 selection.

Within each retained tissue, gene expression was summarized as median transcripts per million (TPM) across samples and transformed as $\log_2(\text{TPM} + 0.1)$. Eligible protein-coding genes were ranked independently within tissue and divided into four approximately equal groups. Tied values were ranked by their first occurrence before quartile assignment. Q1 and Q4 denote the lowest and highest relative expression strata within each tissue, with tissue-specific TPM boundaries. A sensitivity analysis defined quartiles only among genes harboring the focal polymorphic cargo.

Insertion coordinates and identities were held fixed across tissue rankings. The primary hs1-matched population background comprised 18,357 protein-coding genes. For each gene, we recorded the number of distinct quartiles occupied across 51 tissues and whether it changed quartile. We then used mapped 1KGP events to summarize quartile changes for all insertion hosts, for hosts of each cargo and for individual insertion events. The descriptive matrix and bar for all genes in Fig. 1e,f instead use the parallel 18,236-gene GENCODE-matched background from the GRCh38 branch. The 1KGP-host bars use hs1-matched host sets. We distinguished the minimum-to-maximum quartile span from the number of distinct quartiles visited. Both describe changes in host-expression classification across tissues rather than independent insertion replicates.

Because GTEX and 1KGP are not carrier matched, the tissue rankings describe reference host-expression contexts rather than insertion-time expression or allele-specific effects. Transcription-associated filtering was considered as a possible explanation for differential representation, but it was not directly measured.

Population insertion classification and abundance measures

The primary population analysis retained exact, unambiguous SVAN family assignments to L1, Alu or SVA. Composite family calls were retained in audit outputs rather than reassigned to a focal cargo. Events were assigned to hs1 protein-coding exon-union introns using the source insertion coordinates and the workflow's host-assignment rules. SVAN INS_LEN supplied the primary inserted-sequence length. RT_LEN was analyzed separately as a robustness measure. Orientation was defined relative to host-gene transcription when insertion strand was resolved. Unresolved orientations remained eligible for total event and sequence summaries but were omitted from orientation-specific analyses.

For each cargo, tissue and quartile, event density was the number of intronic events divided by eligible background intronic base pairs, multiplied by 10^6 to give events Mb^{-1} . Inserted-sequence burden was the sum of reported inserted base pairs divided by the same denominator, also multiplied by 10^6 . Inserted base pairs per event were summarized separately. Events were counted as catalogue loci rather than weighted by allele frequency. Q4/Q1 ratios were calculated within each tissue and then summarized by their median, which can differ from the ratio of the corresponding medians.

Positive-host occupancy was the number of genes containing at least one focal event divided by the number of eligible background genes. Multiplicity was the event count divided by the number of hosts containing the focal cargo. Their product gives events per background gene, whereas event density additionally accounts for intronic base pairs per gene. The local architectural analysis used only the unique introns containing the focal cargo as its denominator (Additional file 1: Fig. S9). This conditional density was analyzed separately from the primary analysis of all background introns.

Population abundance models and cargo interactions

PPML models were fitted to event counts and summed inserted base pairs, using the logarithm of eligible intronic base pairs as an offset. For summed base pairs, the log-link model estimates a sequence-burden ratio without treating individual bases as independent events. Inference used robust sandwich covariance. Expression was modeled either as ordered Q1–Q4, giving an effect per quartile step, or as the contrast between Q1 and Q4 after restriction to those groups. The two estimates were reported separately.

Gene-by-tissue analyses included genes with zero focal events and tissue fixed effects. We log10-transformed four architecture covariates: gene span, mean intron length, intron number and number of merged exon intervals. These were all derived from the exon-union gene models. We also included the fractions of intronic sequence in the first intron and within 10 kb of the TSS. Continuous covariates were standardized. Native hs1 intronic GC was computed from the exon-union intronic sequence as the percentage of G or C among A, C, G and T bases. The fully adjusted population model included this GC measure alongside architecture covariates. Chromatin accessibility and callability were not comprehensively modeled.

Within-cargo models estimated expression-associated rate ratios. Joint models stacked the three cargo outcomes and included cargo indicators, expression-by-cargo interactions, tissue effects and cargo-

specific slopes for the adjustment covariates. Models used complete cases for the specified covariates and required positive intronic opportunity. Missing covariates were not coded as zeros. Covariance was clustered by host gene to account for repeated appearances across tissues and cargoes. Exponentiated linear contrasts yielded within-cargo IRRs and the Alu-versus-L1, SVA-versus-L1 and SVA-versus-Alu ratios of rate ratios.

Sensitivity analyses and local host-reference sequence

Grouped abundance models evaluated exclusion of first introns and exclusion of intronic sequence within 10 kb of the TSS, recalculating events and eligible opportunity for each restriction. The canonical-only sensitivity further restricted the SVAN classification. Tissue dependence was assessed by omitting each of the 51 tissues in turn and refitting the grouped cargo-interaction model. Chromosome-block bootstrap analyses resampled chromosomes with replacement, carrying their counts and intronic opportunity together, with 500 resamples per expression contrast and a fixed random seed of 20260809. These checks evaluated cargo ordering in grouped models, separately from the fully adjusted gene-level analyses (Fig. 2d–f, Additional file 1: Fig. S3).

Sequence context was extracted from the hs1 host reference around population insertion coordinates, without adding the non-reference inserted sequence. Symmetric 100-bp and 500-bp flanks characterized broader local GC. For orientation-resolved insertions, 10-bp, 20-bp, 100-bp and 500-bp windows were also represented relative to TE orientation, retaining distinct 5' and 3' contexts in the underlying summaries. Figure 2g–j shows the corresponding profiles for each cargo and orientation relative to the host gene. Additional analyses of local sequence context masked the central bases nearest the target to distinguish immediate sequence from broader flanks. Coordinate conversions were checked during extraction.

Local cargo comparisons included host intronic GC, local intron length and chromosome as specified for each model. GC percentages used A/C/G/T bases as the denominator and excluded ambiguous bases. These profiles describe the sequence in the present-day host reference genome around observed insertion coordinates. They do not reconstruct ancestral pre-insertion sequence or exact target-site-duplication boundaries and cannot establish a causal targeting preference.

Orientation, insertion length and exon-boundary models

Sense and antisense event densities used the same eligible intronic denominator as total density. Antisense fraction was calculated only among orientation-resolved events. We analyzed absolute rates for each orientation separately from the relative antisense fraction. Event-level binomial models tested antisense status against expression, insertion length, local intron length, distance to the nearest exon boundary, intronic GC and tissue. Length and distance covariates were log transformed, with distance represented as $\log_{10}(\text{distance} + 1)$, and continuous covariates standardized. In the population orientation models, covariance was clustered by insertion event.

Length-stratified analyses used INS_LEN as the primary population measure and RT_LEN as a robustness measure, with corresponding Alu and SVA summaries. The formal L1 boundary model included expression, an indicator of whether the nearest exon boundary was ≤ 1 kb away and their interaction, together with insertion length, local intron length, intronic GC and tissue. The comparison was repeated for L1 insertions at least 6 kb long. With quartiles treated as ordered groups, the analysis of all L1 insertions included 1,831 clusters defined by insertion event, and the analysis of long L1 insertions included 626. Analyses restricted to Q1 and Q4 involved fewer clusters (Additional file 2: Table S3). Boundary effects were assessed through the interaction term, rather than differences in stratum-specific P values.

Reconstruction of reference-retained TE events and sequence

GRCh38 reference L1, Alu and SVA were reconstructed from full RepeatMasker identifiers in the UCSC rmskOutBaseline resource and the release-compatible nestedRepeats table [54]. Full identifiers were matched within chromosome to identify fragments belonging to one interrupted element. A group was collapsed only if the expected member structure was completely recovered and class assignments were concordant. Incomplete or discordant groups were set aside. Unique identifiers outside interrupted groups were retained as singleton events. Fragments were not joined on the basis of genomic proximity alone. The optional rmskJoinedBaseline table was used to check source consistency and did not define primary events.

An event entered the primary intronic set only when its complete outer span was contained within one exon-union intron of an unambiguous GENCODE v48 protein-coding host. A genome-wide protein-coding exon mask additionally excluded overlap with any protein-coding exon, including loci intronic to one gene but exonic to another. For interrupted events, retained TE sequence was measured from the aligned RepeatMasker member blocks, excluding intervening non-TE sequence. Outer span and aligned retained sequence were therefore distinct quantities.

To define intronic opportunity in the primary reference-TE analysis, we subtracted the aligned blocks of reconstructed TE events from the current intronic sequence of each gene, as defined by exon union. Event density and aligned-sequence burden were then normalized to this self-excluded background. A conventional current-intron denominator was calculated in parallel and used for the harmonized comparison of L1, Alu, SVA and PP in Fig. 4. Each contrast used a single denominator consistently. Denominator calibration and broad repeat controls are shown in Additional file 1: Figs. S3b and S6a,c.

L1 was grouped into L1HS/L1PA2, L1PA3–8, L1PA9–17, L1PB, L1MA and older L1M strata. Alu was grouped into AluJ, AluS and AluY, and SVA into SVA_A–SVA_F. These groups represent relative evolutionary strata of surviving annotations rather than exact insertion ages. Event density, retained-sequence burden and orientation were evaluated within each group across the same expression quartiles (Fig. 5, Additional file 1: Figs. S7 and S8d,e).

Processed-pseudogene definitions and models

PPs were identified from GENCODE v48 processed_pseudogene, transcribed_processed_pseudogene and translated_processed_pseudogene annotations. They were treated as cellular-mRNA-derived retrocopies rather than transposable elements. Each annotated locus contributed one retained retrocopy and was assigned to a protein-coding exon-union intron by its genomic midpoint. This PP assignment rule differs from the full-span containment rule for reconstructed reference TEs. Orientation was relative to the assigned host. Sequence burden used the current GENCODE PP span, whose length may differ from the original cDNA insertion.

PP event density, annotated-span burden, occupancy, multiplicity, antisense fraction and span per event were evaluated across the 51 tissue rankings with the current-intron denominator. Gene-by-tissue models used binomial regression for positive-host status and offset PPML models for event number and annotated-span burden, adjusting for host architecture and tissue with covariance clustered by host gene. The PP branch did not include native intronic GC adjustment. Its event-level antisense model included expression, standardized log₁₀ annotated PP span, their interaction, host architecture and tissue, again clustering by host gene. The reported OR for the association between PP orientation and expression is evaluated at the mean standardized log span and can vary with PP length through the interaction term. Supporting estimates and tissue heterogeneity are in Additional file 1: Fig. S6 and Additional file 2: Table S2.

PCAWG somatic events and donor-matched expression

Somatic calls were obtained from the published PCAWG retrotransposition analysis [45]. Events were assigned to GENCODE v19 protein-coding introns. L1, Alu and SVA identities followed the source mobile-element labels. Source-designated processed-pseudogene events were retained as PP. Tumor RNA from E-MTAB-5200 was linked using donor metadata. Where multiple eligible runs were available for one donor, median reads per kilobase per million mapped reads (RPKM) were used. Genes were ranked within donor after $\log_2(\text{RPKM} + 0.1)$ transformation and divided into Q1–Q4.

For each quartile, the denominator summed the eligible protein-coding intronic sequence across donors. Donor-quartile combinations with zero events were retained to preserve their intronic opportunity. Formal pan-cancer inference was restricted to L1 and used PPML with fixed effects for cancer type and covariance clustered by donor. Ordered-Q and Q4/Q1 effects were fitted separately. Sensitivity analyses trimmed donors with high L1 burden, omitted cancer types and applied event restrictions. Sparse Alu and PP events provided descriptive or lower-powered support. SVA was not modeled inferentially.

Sequence burden was calculated only for events with explicitly reported retrotransposed lengths. Missing values were not imputed, and values greater than 100 kb were excluded by the prespecified sanity filter. We reported burden, mean length and the fraction of events with resolved lengths to make

the subset coverage explicit. Source-designated PP lengths were not reconstructed from other fields (Additional file 1: Fig. S10).

Reanalysis of published recovered L1 and Alu insertions

Published L1 [25] and Alu [46] recovery maps were reannotated against a common GENCODE v19 protein-coding intron background. Primary analyses required an unambiguous intronic host and no protein-coding exonic assignment. HPA expression ranked eligible protein-coding genes within the corresponding cell line. The primary comparison used a common HeLa expression stratification and intronic background. Expression-matched PA-1 L1 data provided secondary support. HPA provides an external cell-line expression reference and does not measure expression immediately before insertion in the source experiments.

Recovered event density was the number of eligible insertions per megabase of background intronic sequence in each expression stratum. Q4/Q1 rate ratios used the corresponding event counts and exposures. Confidence intervals were calculated on the log-rate-ratio scale with Poisson count-based standard errors. To compare the Q4/Q1 effects directly, we divided the rate ratio for Alu by that for L1. The common stratum denominators cancel in this direct contrast. The workflow applied a 0.5 continuity correction only when a count required for the relevant contrast was zero. Orientation-specific summaries used the same host-relative strand definition and opportunity background.

The published call tables did not report explicit inserted lengths for individual events. We therefore did not estimate sequence burden from coordinate span, poly(A)-tail length, read length or assumed construct length. These maps contain insertions recovered after growth, survival, selection and detection. Since the L1 and Alu maps came from separate studies, their comparison was treated as support across studies rather than a controlled experiment within the present work (Fig. 6c–f, Additional file 1: Fig. S11).

Statistical reporting and computational reproducibility

Tests were two sided, with 95% confidence intervals and Benjamini–Hochberg correction within the specified families of related comparisons. The significance threshold was $FDR < 0.05$ where FDR-adjusted values were reported. Covariance was clustered by host gene for gene-level abundance and PP orientation models, by insertion event for population orientation models, by tissue for grouped tissue comparisons and by donor for PCAWG models. Medians and distributions across tissues were treated as descriptive summaries because the same insertions recur across tissues.

For population gene-level models, correction families were defined by model specification, outcome and expression contrast. Joint cargo and orientation contrasts were grouped by model identifier. PP burden families were grouped by model and expression contrast, and PP span–orientation contrasts by expression contrast. Recovered-insertion comparisons used the corresponding contrast families recorded by the workflow. Outputs recorded model convergence, usable observations, clusters and included covariates. Non-converged estimates were excluded from inference.

The custom Python workflow records input resources, assembly-specific annotation rules, tissue exclusions, model definitions and source-length checks. It uses NumPy, pandas, SciPy and spreadsheet readers, with version constraints specified in the requirements file. Python 3.11 is recommended. Synthetic checks cover critical annotation and tissue-selection logic but do not reproduce estimates from the original large input datasets. Analysis counts and selected statistical results are provided in Additional file 2: Tables S2 and S3. Access to the analysis code and derived data is described in the Availability of data and materials section.

Use of AI-assisted tools

OpenAI's ChatGPT was used to assist with code development. The author reviewed and verified all code and analysis outputs and takes full responsibility for the analyses and results.

Abbreviations

1KGP (1000 Genomes Project), CI (confidence interval), FDR (false discovery rate), GC (guanine–cytosine content), GTEx (Genotype-Tissue Expression), HPA (Human Protein Atlas), hs1 (T2T-CHM13v2.0 human genome assembly), IGSR (International Genome Sample Resource), IRR (incidence-rate ratio), L1 (long interspersed nuclear element-1), LTR (long terminal repeat), OR (odds ratio), ORF2p (L1 open reading frame 2 protein), PCAWG (Pan-Cancer Analysis of Whole Genomes), PP (processed pseudogene), PPML (Poisson pseudo-maximum likelihood), RPKM (reads per kilobase per million mapped reads), SINE (short interspersed nuclear element), SVA (SINE–VNTR–Alu), TE (transposable element), TPM (transcripts per million), TSS (transcription start site), VNTR (variable number tandem repeat).

Declarations

Ethics approval and consent to participate

This study reanalyzed publicly available genomic annotations, insertion catalogues and expression data. No participants were recruited, no new human or animal samples were collected, and no identifiable participant information was used. Ethics approval and participant-consent arrangements for the original data collections are described in the source studies. The present study involved no new experimental procedures in humans or animals.

Consent for publication

Not applicable. The manuscript contains no identifiable individual participant information, images or case descriptions.

Availability of data and materials

The primary datasets analyzed in this study are publicly available from their original resources. GTEx v10 expression data and sample annotations are available from the GTEx Portal [51]. The 1KGP long-read resource is described by Schloissnig et al. [44], is distributed through IGSR as 1KG_ONT_VIENNA, and is associated with ENA accession PRJEB89727 [58]. GENCODE v48 and v19 annotations are available from their release pages [52,53]. GRCh38 RepeatMasker baseline and nested-repeat annotations and hs1 resources are available from UCSC [54,55]. PCAWG insertion calls are provided with the pan-cancer retrotransposition study and its data release, including Synapse accession syn21052009 [45]. Donor-matched tumor expression is available under E-MTAB-5200 [56], and cell-line expression from HPA [57]. Recovered L1 and Alu insertion maps are available with the source publications [25,46].

Supplementary figures and summary tables are provided as Additional files 1 and 2. Custom analysis scripts and the derived data underlying the reported results will be provided to the editors and reviewers during peer review. These materials will be deposited in a public repository and made openly available no later than publication, with a permanent repository identifier included in the final manuscript. No new sequencing or experimental data were generated.

Competing interests

The author declares no competing interests.

Funding

Not applicable. This work received no specific funding from public, commercial or not-for-profit funding agencies.

Authors' contributions

M.-G.L. conceived the study, designed the analytical framework, curated and harmonized the public datasets, developed the computational workflows, performed the analyses, interpreted the results, prepared the figures and wrote and revised the manuscript. M.-G.L. is the sole author and is responsible for the final manuscript.

Acknowledgements

The author acknowledges the investigators and consortia who generated and made available the public datasets reanalyzed in this study. The use of AI-assisted tools is described in Methods.

References

1. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, et al. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol*. 2001;21:1429–39. 10.1128/mcb.21.4.1429-1439.2001.

2. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet.* 2003;35:41–8. 10.1038/ng1223.
3. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet.* 2011;20:3386–400. 10.1093/hmg/ddr245.
4. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet.* 2000;24:363–7. 10.1038/74184.
5. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 2003;13:1998–2004. 10.1101/gr.1649303.
6. Kim TM, Jung YC, Rhyu MG, Alu. L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. *J Korean Med Sci.* 2004;19:783–92. 10.3346/jkms.2004.19.6.783.
7. Jjingo D, Huda A, Gundapuneni M, Mariño-Ramírez L, Jordan IK. Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol Evol.* 2011;3:259–71. 10.1093/gbe/evr015.
8. Lerat E, Sémon M. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene.* 2007;396:303–11. 10.1016/j.gene.2007.04.002.
9. Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, Marahrens Y. Repetitive sequence environment distinguishes housekeeping genes. *Gene.* 2007;390:153–65. 10.1016/j.gene.2006.09.018.
10. Urrutia AO, Ocaña LB, Hurst LD. Do Alu repeats drive the evolution of the primate transcriptome? *Genome Biol.* 2008;9:R25. 10.1186/gb-2008-9-2-r25.
11. Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res.* 2002;12:1483–95. 10.1101/gr.388902.
12. Graham T, Boissinot S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J Biomed Biotechnol.* 2006;2006:75327. 10.1155/jbb/2006/75327.
13. Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol.* 2001;18:926–35. 10.1093/oxfordjournals.molbev.a003893.
14. Zhang Y, Romanish MT, Mager DL. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol.* 2011;7:e1002046. 10.1371/journal.pcbi.1002046.
15. Linker S, Hedges DJ. Linear decay of retrotransposon antisense bias across genes is contingent upon tissue specificity. *PLoS ONE.* 2013;8:e79402. 10.1371/journal.pone.0079402.
16. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature.* 2004;429:268–74. 10.1038/nature02536.

17. Perepelitsa-Belancio V, Deininger P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet.* 2003;35:363–6. 10.1038/ng1269.
18. Belancio VP, Hedges DJ, Deininger P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* 2006;34:1512–21. 10.1093/nar/gkl027.
19. Ustyugova SV, Lebedev YB, Sverdlov ED. Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. *Genetica.* 2006;128:261–72. 10.1007/s10709-005-5967-2.
20. Xiong F, Wang R, Lee JH, Li S, Chen SF, Liao Z, et al. RNA m6A modification orchestrates a LINE-1-host interaction that facilitates retrotransposition and contributes to long gene vulnerability. *Cell Res.* 2021;31:861–85. 10.1038/s41422-021-00515-8.
21. Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, et al. Human retroelements may introduce intragenic polyadenylation signals. *Cytogenet Genome Res.* 2005;110:365–71. 10.1159/000084968.
22. Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, et al. Minimal conditions for exonization of intronic sequences: 5' splice site formation in Alu exons. *Mol Cell.* 2004;14:221–31. 10.1016/s1097-2765(04)00181-9.
23. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell.* 2013;152:453–66. 10.1016/j.cell.2012.12.023.
24. Lebedev YB, Amosova AL, Mamedov IZ, Fisunov GY, Sverdlov ED. Most recent AluY insertions in human gene introns reduce the content of the primary transcripts in a cell type specific manner. *Gene.* 2007;390:122–9. 10.1016/j.gene.2006.09.031.
25. Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, et al. Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. *Cell.* 2019;177:837–e85128. 10.1016/j.cell.2019.02.050.
26. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, El Zine A, et al. The landscape of L1 retrotransposons in the human genome is shaped by pre-insertion sequence biases and post-insertion selection. *Mol Cell.* 2019;74:555–e5707. 10.1016/j.molcel.2019.02.036.
27. Gasior SL, Preston G, Hedges DJ, Gilbert N, Moran JV, Deininger PL. Characterization of pre-insertion loci of de novo L1 insertions. *Gene.* 2007;390:190–8. 10.1016/j.gene.2006.08.024.
28. Servant G, Strevia VA, Deininger PL. Transcription coupled repair and biased insertion of human retrotransposon L1 in transcribed genes. *Mob DNA.* 2017;8:18. 10.1186/s13100-017-0100-5.
29. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Res.* 2017;45:2318–28. 10.1093/nar/gkw1286.
30. Spirito G, Mangoni D, Sanges R, Gustincich S. Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *BMC Bioinformatics.* 2019;20:495. 10.1186/s12859-019-3113-x.

31. Hughes LS, Fröhlich A, Pfaff AL, Bubb VJ, Quinn JP, Köks S. Exploring SVA insertion polymorphisms in shaping differential gene expressions in the central nervous system. *Biomolecules*. 2024;14:358. 10.3390/biom14030358.
32. Savage AL, Bubb VJ, Breen G, Quinn JP. Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol Biol*. 2013;13:101. 10.1186/1471-2148-13-101.
33. Price E, Gianfrancesco O, Harrison PT, Frank B, Bubb VJ, Quinn JP. CRISPR deletion of a SVA retrotransposon demonstrates function as a cis-regulatory element at the TRPV1/TRPV3 intergenic region. *Int J Mol Sci*. 2021;22:1911. 10.3390/ijms22041911.
34. Hall A, Middlehurst B, Cadogan MAM, Reed X, Billingsley KJ, Bubb VJ, et al. A SINE-VNTR-Alu at the LRIG2 locus is associated with proximal and distal gene expression in CRISPR and population models. *Sci Rep*. 2024;14:792. 10.1038/s41598-023-50307-w.
35. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, et al. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res*. 2012;40:1666–83. 10.1093/nar/gkr863.
36. Chu C, Lin EW, Tran A, Jin H, Ho NI, Veit A, et al. The landscape of human SVA retrotransposons. *Nucleic Acids Res*. 2023;51:11453–65. 10.1093/nar/gkad821.
37. Kirby AE, Loftus M, Golba EC, Human Genome Structural Variation Consortium, Konkel MK. Structural and transduction patterns of human-specific polymorphic SVA insertions. *Mob DNA*. 2025;16:42. 10.1186/s13100-025-00373-w.
38. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA elements: a hominid-specific retroposon family. *J Mol Biol*. 2005;354:994–1007. 10.1016/j.jmb.2005.09.085.
39. Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol*. 2013;14:R22. 10.1186/gb-2013-14-3-r22
40. Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res*. 2013;23:2042–52. 10.1101/gr.154625.113.
41. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JM, et al. Processed pseudogenes acquired somatically during cancer development. *Nat Commun*. 2014;5:3644. 10.1038/ncomms4644.
42. Feng X, Li H. Higher rates of processed pseudogene acquisition in humans and three great apes revealed by long-read assemblies. *Mol Biol Evol*. 2021;38:2958–66. 10.1093/molbev/msab062.
43. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369:1318–30. 10.1126/science.aaz1776.
44. Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, et al. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature*. 2025;644:442–52. 10.1038/s41586-025-09290-7.

45. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet.* 2020;52:306–19. 10.1038/s41588-019-0562-0.
46. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, Makova KD, et al. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS Genet.* 2012;8:e1002842. 10.1371/journal.pgen.1002842.
47. Hackenberg M, Bernaola-Galván P, Carpena P, Oliver JL. The biased distribution of Alus in human isochores might be driven by recombination. *J Mol Evol.* 2005;60:365–77. 10.1007/s00239-004-0197-2.
48. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A.* 2007;104:8005–10. 10.1073/pnas.0611223104.
49. Solovyov A, Behr JM, Hoyos D, Banks E, Drong AW, Thornlow B, et al. Pan-cancer multi-omic model of LINE-1 activity reveals locus heterogeneity of retrotransposition efficiency. *Nat Commun.* 2025;16:2049. 10.1038/s41467-025-57271-1.
50. Zhang T, Zhao W, Wirth C, Díaz-Gay M, Yin J, Cecati M, et al. Uncovering the role of LINE-1 in the evolution of lung adenocarcinoma. *Nature.* 2026;650:230–41. 10.1038/s41586-025-09825-y.
51. GTEx Consortium. GTEx Portal: version 10 bulk-tissue expression datasets and sample annotations. <https://www.gtexportal.org/home/downloads/adult-gtex>. Accessed 8 Sep 2026.
52. GENCODE. Human release 48 (GRCh38.p14). https://www.gencodegenes.org/human/release_48.html. Accessed 8 Sep 2026.
53. GENCODE. Human release 19 (GRCh37.p13). https://www.gencodegenes.org/human/release_19.html. Accessed 8 Sep 2026.
54. UCSC Genome Browser. GRCh38/hg38 annotation database: rmskOutBaseline, nestedRepeats and associated repeat resources. <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/>. Accessed 8 Sep 2026.
55. Genome Browser. T2T-CHM13v2.0 (hs1) assembly and NCBI RefSeq gene annotations. <https://hgdownload.soe.ucsc.edu/goldenPath/hs1/bigZips/>. Accessed 8 Sep 2026.
56. EMBL-EBI Expression Atlas. The International Cancer Genome project: Pan-Cancer Analysis of Whole Genomes (PCAWG). Dataset E-MTAB-5200. <https://www.ebi.ac.uk/gxa/experiments/E-MTAB-5200>. Accessed 8 Sep 2026.
57. Human Protein Atlas. Downloadable data: cell-line RNA expression resource. <https://www.proteinatlas.org/about/download>. Accessed 8 Sep 2026.
58. European Nucleotide Archive. 1KGP long-read sequencing study, accession PRJEB89727. <https://www.ebi.ac.uk/ena/browser/view/PRJEB89727>. Accessed 8 Sep 2026.

Figures

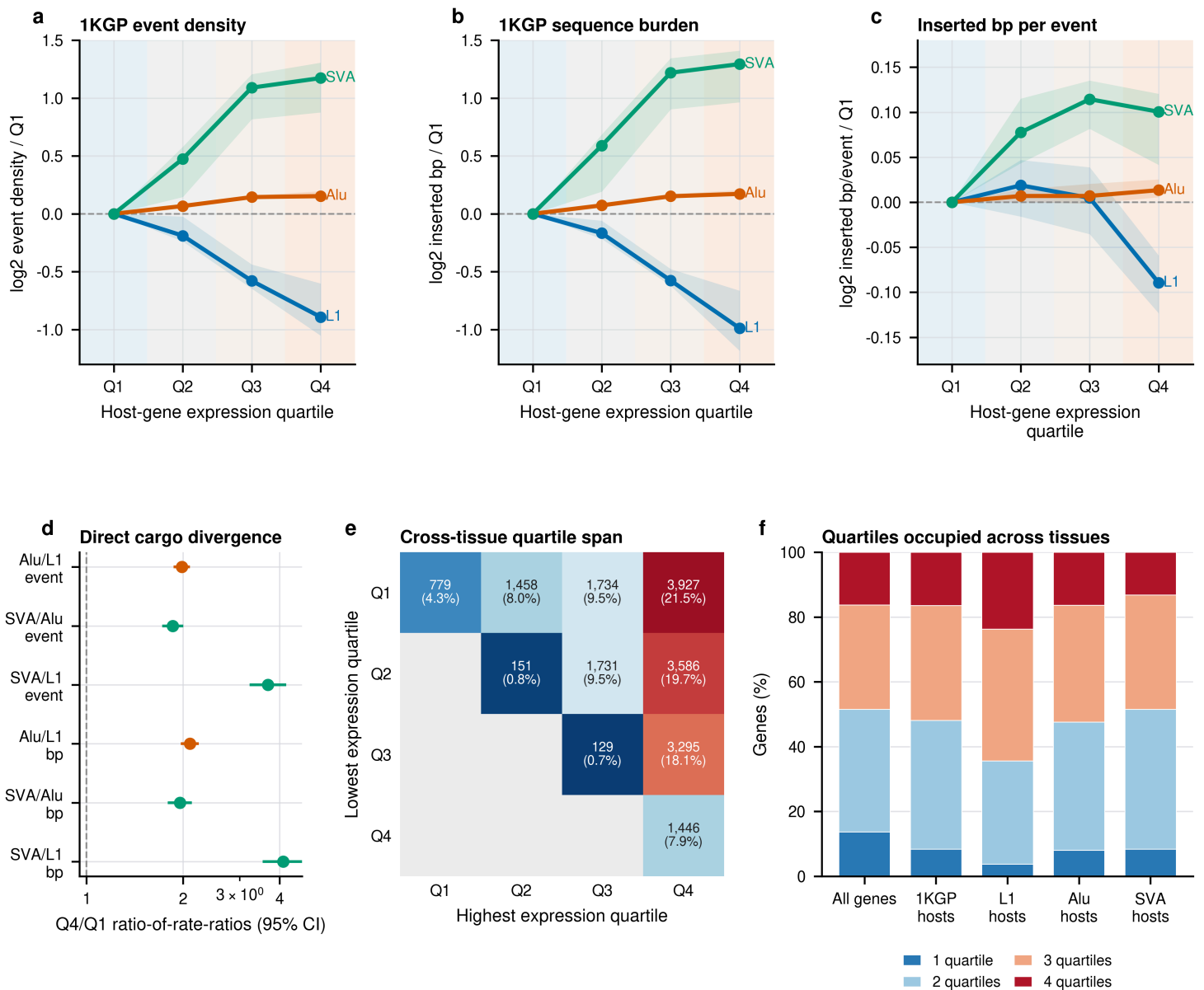


Figure 1

1KGP cargo divergence persists across tissue-specific host-gene re-ranking

a,b, Q1–Q4 1KGP polymorphic event density and inserted-sequence burden for intronic L1, Alu and SVA across 51 GTEx anatomical tissues, expressed as log2 ratios to Q1. Lines show tissue medians and shaded bands summarize the across-tissue distribution. c, Inserted base pairs per event relative to Q1. d, Direct Q4-versus-Q1 cargo-by-expression rate-ratio contrasts for event count and inserted sequence, with 95% CIs. e, Minimum-versus-maximum quartile matrix for the descriptive 18,236-gene GENCODE-matched GTEx background. f, Number of distinct quartiles occupied by that all-gene background and by the hs1-matched 1KGP host-gene union and cargo-specific host sets. Panel e records the minimum and maximum quartiles even if intermediate quartiles were not visited. Panel f counts distinct quartiles visited. Event count and inserted sequence are normalized to eligible protein-coding intronic sequence.

The insertion catalogue is fixed across tissues. Tissue-wise endpoints and ordered-Q models are in Additional file 1: Fig. S4. Event-weighted re-ranking is in Additional file 1: Fig. S3e.

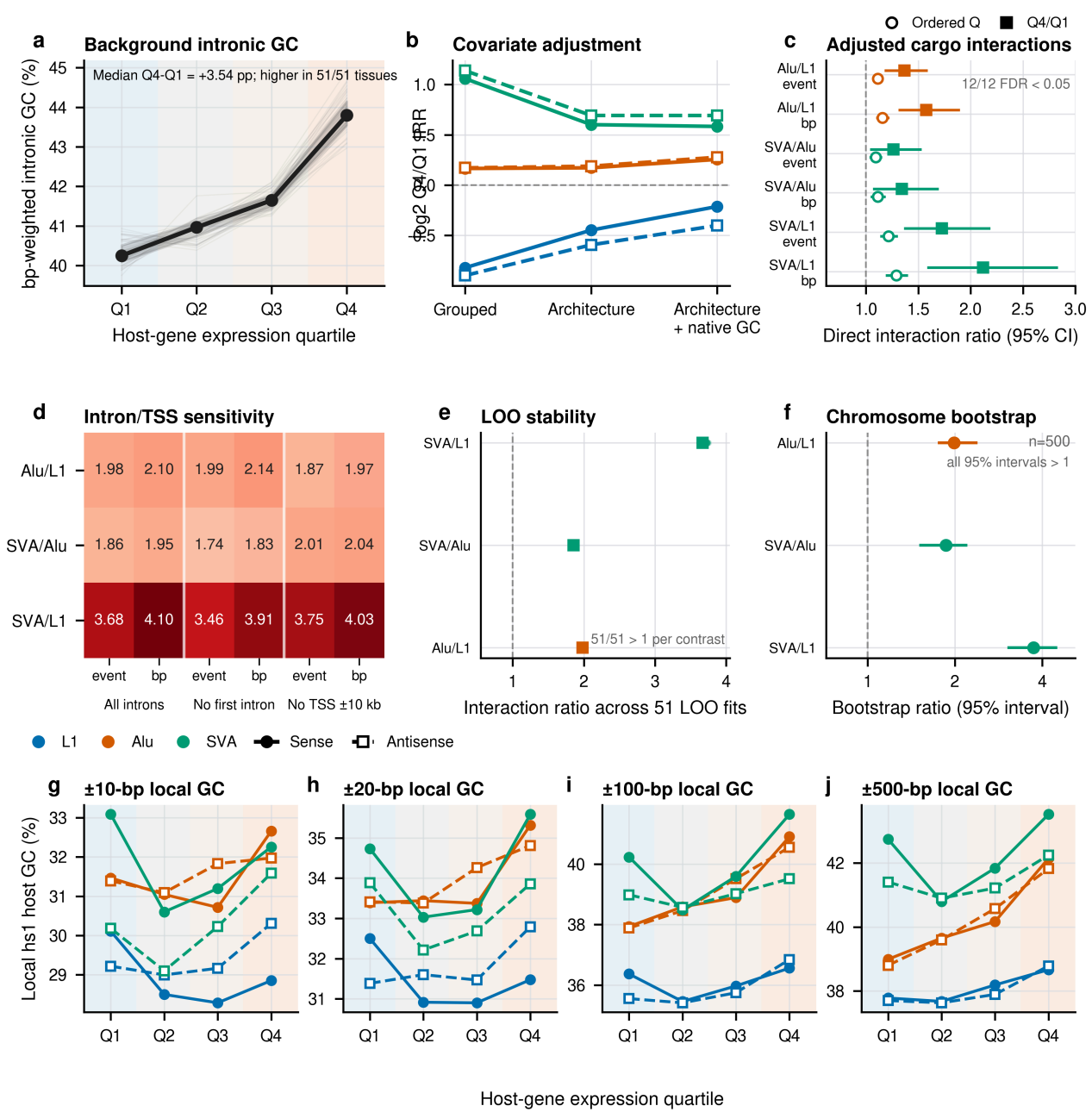


Figure 2

Cargo divergence persists after host-context adjustment amid distinct local GC environments

a, Background protein-coding intronic GC across Q1–Q4 in 51 tissues. The black line shows the tissue median. b, Within-cargo Q4/Q1 IRRs from grouped, host-architecture-adjusted and architecture-plus-native-GC-adjusted PPML models, displayed on a log2 scale. Solid filled-marker trajectories show event-

count effects and dashed open-marker trajectories show inserted-sequence effects in this panel. Eligible intronic sequence is the exposure. c, Fully adjusted direct cargo-by-expression interactions for ordered-Q and Q4-versus-Q1 effects, with 95% CIs. All 12 displayed direct contrasts are FDR-significant. d, Grouped-model direct-interaction sensitivity after excluding first introns or intronic sequence within 10 kb of the TSS. e, Grouped-model estimates from 51 leave-one-tissue-out refits. f, Estimates from chromosome-block bootstrap analyses with 500 resamples. Intervals are bootstrap 95% intervals. g–j, Median host-reference GC around the same 1KGP coordinates at ± 10 , ± 20 , ± 100 and ± 500 bp, stratified by cargo and host-relative orientation. For g–j, solid circles and dashed squares denote sense and antisense, respectively. Full adjustment yields L1 event-count Q4/Q1 FDR = 0.054, whereas L1 sequence-burden Q4/Q1 FDR = 0.0031.

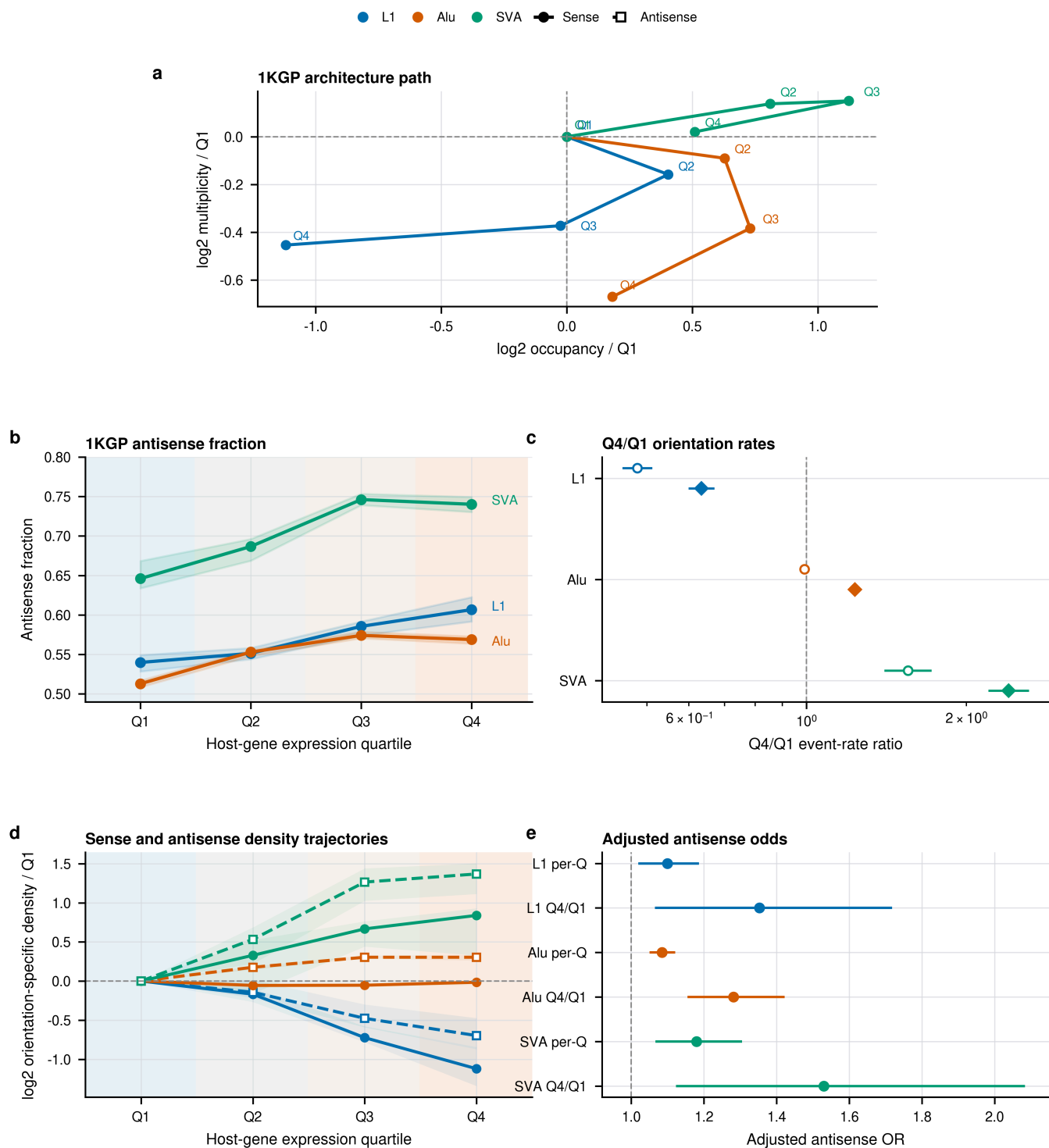


Figure 3

Cargo-specific host architecture and orientation diverge along the expression axis

a, Median positive-host occupancy and within-positive-host multiplicity across Q1–Q4 for L1, Alu and SVA in 51 tissues. b, Median antisense fraction among orientation-resolved events. c, Median tissue-wise Q4/Q1 ratios for absolute sense and antisense event density. d, Orientation-specific event-density

trajectories relative to Q1. e, Adjusted antisense odds ratios from event-level binomial models for ordered-Q and Q4-versus-Q1 effects. Points show odds ratios and lines show 95% CIs. Models adjust for insertion length, local intron length, nearest-exon distance, intronic GC and tissue with covariance clustered by insertion event. Exon-boundary analyses are in Additional file 1: Fig. S5b–e.

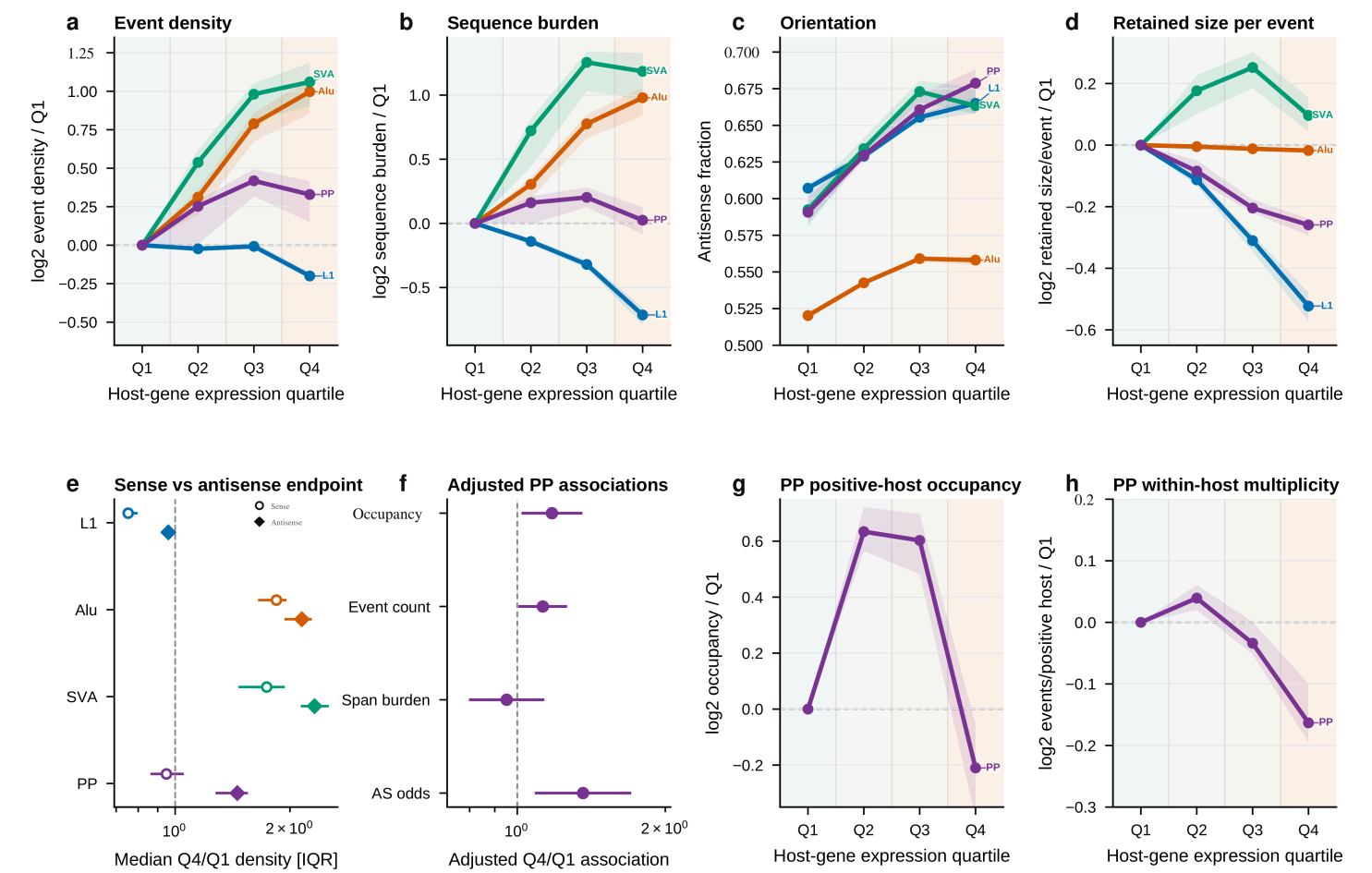


Figure 4

The human reference genome extends cargo divergence to processed pseudogenes

a–d, Reference-retained L1, Alu, SVA and processed-pseudogene trajectories across 51 tissue expression rankings for event density, retained-sequence burden, antisense fraction and retained sequence per event. Lines show tissue medians. e, Median tissue-wise Q4/Q1 sense- and antisense-specific event-density ratios. f, Adjusted Q4-versus-Q1 PP associations for positive-host occupancy, event count, annotated-span burden and antisense orientation. Points show ORs or IRRs and lines show 95% CIs. g,h, PP positive-host occupancy and multiplicity relative to Q1. The four-cargo display uses a harmonized current-intron denominator. PPs are cellular-mRNA-derived retrocopies. Their current annotated spans measure retained loci and may differ from the original cDNA insertions. Reference-denominator calibration is in Additional file 1: Figs. S3b and S6c.

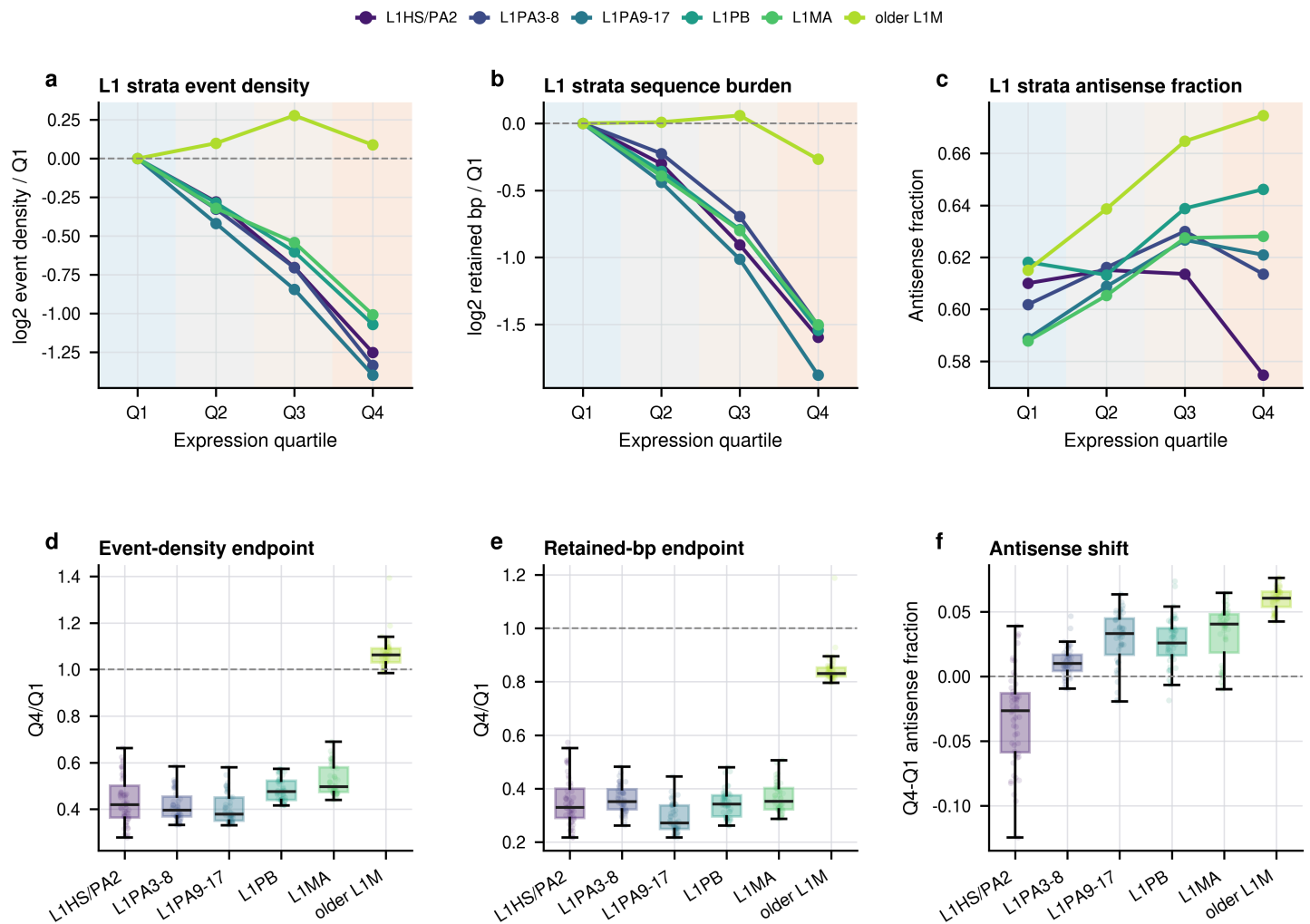


Figure 5

L1 evolutionary strata resolve distinct retention trajectories

a–c, Median Q1–Q4 trajectories across 51 tissues for L1HS/L1PA2, L1PA3–8, L1PA9–17, L1PB, L1MA and older L1M, showing event density, retained-sequence burden and antisense fraction. d,e, Tissue-wise Q4/Q1 endpoint distributions for event density and retained-sequence burden. Boxes and whiskers summarize the 51 tissue-wise endpoint values. f, Tissue-wise Q4-minus-Q1 antisense-fraction change. RepeatMasker groups are treated as evolutionary proxies for surviving annotations rather than exact insertion dates. Alu and SVA subfamily analyses are in Additional file 1: Figs. S7 and S8d,e.

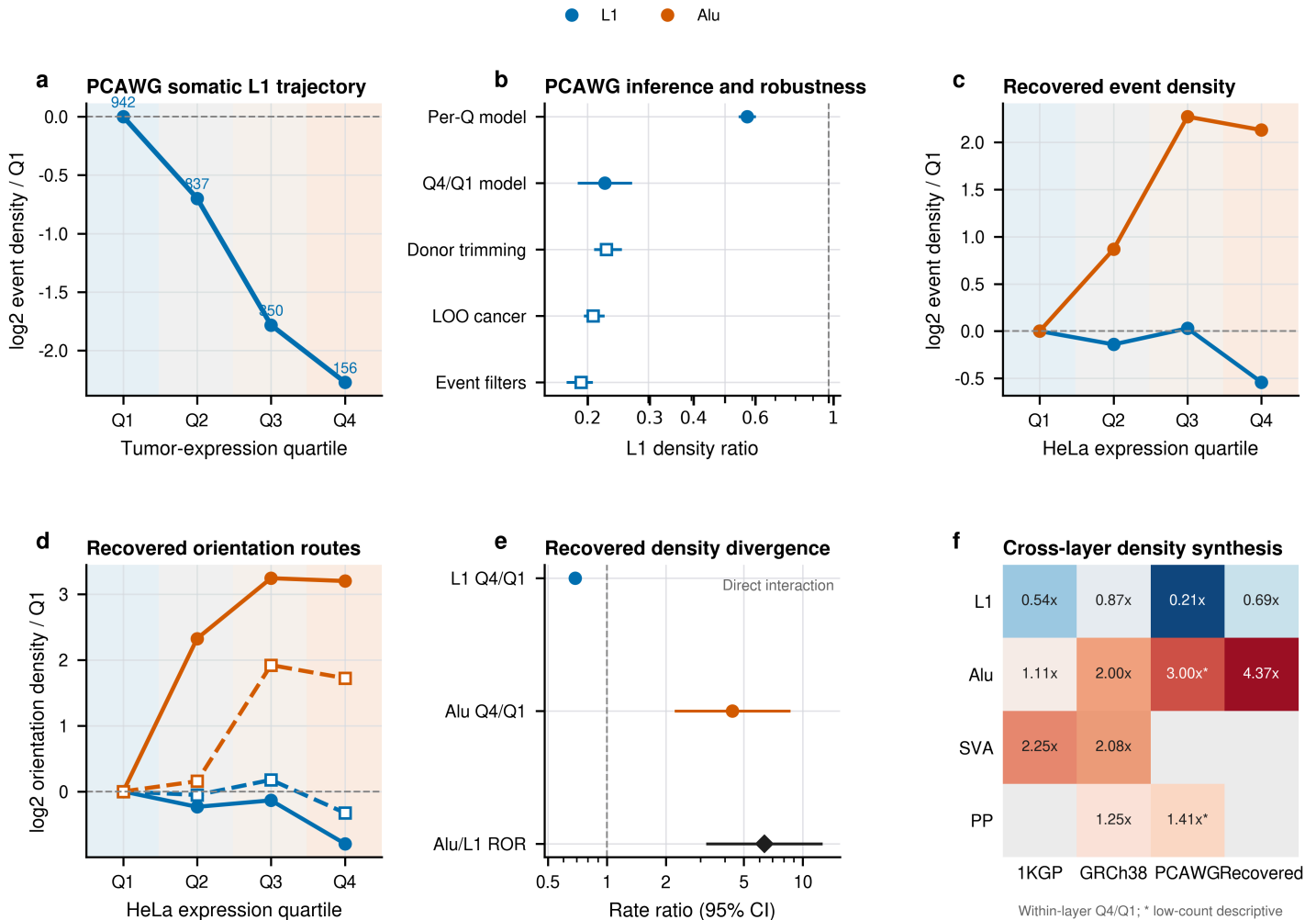


Figure 6

Somatic and recovered insertion layers support cargo-specific density divergence

a, PCAWG somatic L1 event density across donor-specific tumor-expression quartiles. Numbers indicate matched L1 events (Q1, n = 942, Q4, n = 156). b, Cancer-adjusted L1 ordered-Q and Q4/Q1 IRRs with 95% CIs, together with robustness estimates from donor trimming, leave-one-cancer-type-out analysis and event filters. c,d, Total and orientation-specific density of recovered L1 and Alu insertions across matched HeLa expression quartiles. e, Q4/Q1 within-cargo rate ratios and the direct Alu-versus-L1 ratio of rate ratios. Points show estimates and lines show 95% CIs. f, Descriptive cross-layer summary of within-layer Q4/Q1 event-density ratios. Formal PCAWG inference is restricted to L1. Sparse Alu and PP values are descriptive. Recovered maps reflect growth, survival, selection and ascertainment after insertion. The L1 and Alu recovery maps originate from separate studies and are harmonized by annotation and expression background. No new experiments were performed in the present study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFiguresS1S11.pdf](#)
- [SupplementaryTablesS1S4.xlsx](#)