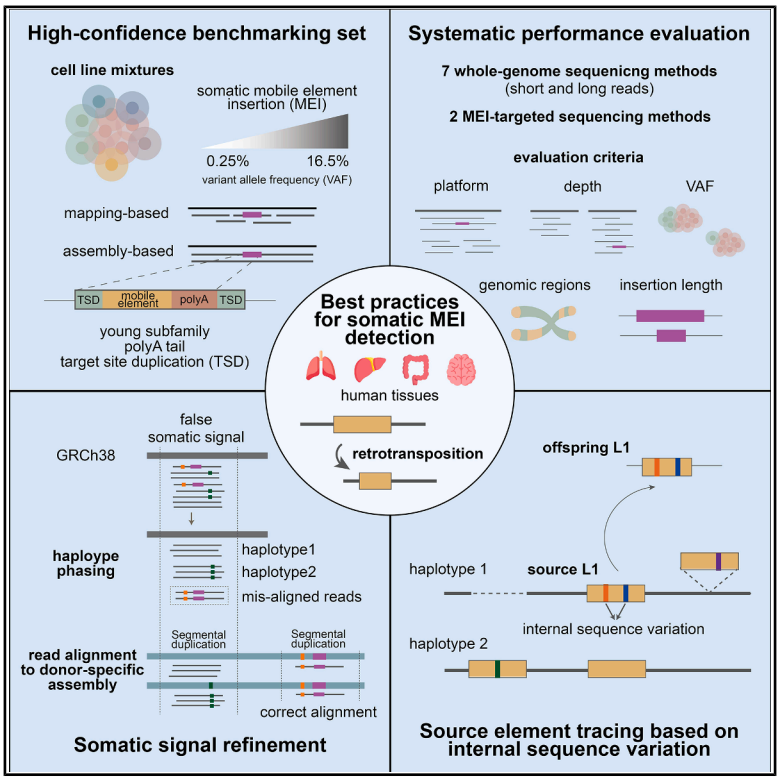
Cell Genomics

Article

Multi-platform framework for mapping somatic retrotransposition in human tissues

Graphical abstract



Highlights

- Systematic benchmarking of somatic MEI detection methods across platforms and VAFs
- Improved MEI detection through haplotype phasing and donor-specific assembly
- Development of a method for tracing source L1s using internal sequence variation
- Identification of 21 somatic L1 insertions in human liver, lung, colon, and brain

Authors

Seunghyun Wang, Mingyun Bae, Jinhao Wang, ..., Ryan E. Mills, Weichen Zhou, Eunjung Alice Lee

Correspondence

apboyle@umich.edu (A.P.B.), remills@umich.edu (R.E.M.), arthurz@umich.edu (W.Z.), ealice.lee@childrens.harvard.edu (E.A.L.)

In brief

Wang et al. establish a systematically benchmarked framework integrating long reads and genome assemblies for accurate detection and source tracing of somatic mobile element insertions. Applied to normal human tissues, this framework reveals rare somatic L1 insertions, enabling studies of how mobile elements reshape genomes during life.

Wang et al., 2026, Cell Genomics 6, 101418
November 11, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2026.101418

CellPress

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

Cell Genomics

CellPress
OPEN ACCESS

Article

Multi-platform framework for mapping somatic retrotransposition in human tissues

Seunghyun Wang,[1,2,18] Mingyun Bae,[1,2,18] Jinhao Wang,[3,18] Boxun Zhao,[1,2] Khue Nguyen,[1,2] Shayna Mallett,[1,2] Jessica A. Switzenberg,[3] Steven J. Losh,[3] Corinne E. Sexton,[4] Benpeng Miao,[5,6] Shihua Dong,[5,6] Xi Zeng,[1,2,7] Ziying Wang,[7] Torrin L. McDonald,[3] Camille Mumm,[3] Rohini K. Gadde,[1,2] Arnaz Maryam Tariq,[1,2] Zhuofu Chen,[8] William C. Feng,[4] Aidan Burn,[9] Junseok Park,[1,2] Chong Chu,[4] Hui Shen,[10] Ting Wang,[5,6,11] Alexander E. Urban,[12,13] Xiaowei Zhu,[8] Heng Li,[4,14,15] Kathleen H. Burns,[9] Hye-Jung E. Chun,[4] Peter J. Park,[4,16] SMaHT MEI Working Group, Alan P. Boyle,[3,17,*] Ryan E. Mills,[3,17,*] Weichen Zhou,[3,*] and Eunjung Alice Lee[1,2,19,*]

[1]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[3]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[5]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[6]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA
[7]College of Informatics, Huazhong Agricultural University, Hongshan District, Wuhan, China
[8]Department of Neuroscience, City University of Hong Kong, Hong Kong SAR, China
[9]Department of Pathology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
[10]Department of Epigenetics, Van Andel Research Institute, Grand Rapids, MI, USA
[11]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[12]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
[13]Department of Genetics, Stanford University, Stanford, CA, USA
[14]Department of Data Science, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
[15]Broad Institute of MIT and Harvard, Boston, MA, USA
[16]Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA
[17]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA
[18]These authors contributed equally
[19]Lead contact
*Correspondence: apboyle@umich.edu (A.P.B.), remills@umich.edu (R.E.M.), arthurz@umich.edu (W.Z.), ealice.lee@childrens.harvard.edu (E.A.L.)
https://doi.org/10.1016/j.xgen.2026.101418

SUMMARY

Mobile element insertions (MEIs) shape the human genome in both germline and somatic tissues. While inherited MEIs are well characterized, identifying somatic MEIs (sMEIs) in non-cancer tissues remains challenging due to their low allelic fraction and repetitive nature. Here, we present an integrative framework for sMEI mapping that leverages modern sequencing technologies and analytical innovations. We first benchmarked sMEI detection methods, demonstrating advantages of long-read and MEI-targeted sequencing for ultra-rare events, and established an effective cross-platform integrative approach using a mixture of well-established cell lines. Haplotype phasing and donor-specific genome assemblies further refined sMEI calls by distinguishing germline and false signals in *in silico* tumor-normal mixtures. We also developed a source-tracing approach based on sequence variation within MEIs, expanding the catalog of active source elements beyond traditional transduction-based inference. Applying this framework to benchmark donor tissues, we detected rare somatic L1 insertions in all four tissue types examined, revealing structural and source diversity. Our work provides a foundational framework and biological insight into somatic retrotransposition in the human body.

INTRODUCTION

Mobile elements are a major source of structural variation in the human germline, yet their activity within the human body remains unknown. Mobile element insertions (MEIs) are generated through an RNA-mediated insertion mechanism known as target-primed reverse transcription (TPRT).[1,2] In humans, only LINE-1 (L1) elements encode the proteins required for retrotransposition, enabling mobilization of not only L1s themselves but also other transposon-derived RNAs such as *Alu* and SVA (SINE-VNTR-Alu), as well as processed mRNAs.[3] Because their insertions endanger genomic integrity,[4–6] induce genetic and

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

epigenetic alterations,[7–9] and trigger immune responses,[10,11] mobile elements are kept under strict host surveillance through multiple layers of defense mechanisms, including epigenetic silencing of source elements and interfering RNAs.[12] Nevertheless, some elements evade this surveillance[13] and contribute to disease initiation and progression, such as in cancers[6,14–17] and neurological disorders.[18–20] Recent studies further indicate that somatic MEIs (sMEIs) occur even in non-diseased normal human tissues[19,21–25] and during aging,[12,26] raising important questions about their prevalence and biological impact throughout the body.

Identification of sMEIs and their parental source elements in tissues remains challenging, with most obstacles rooted in current short-read sequencing technologies and low variant allele frequency (VAF). For instance, correct read alignment in repetitive genomic regions, such as segmental duplications (SDs),[27,28] is particularly problematic with short reads and hampers accurate variant calling. The repetitive nature of MEIs themselves further complicates the delineation of true somatic insertions when they arise within or adjacent to existing MEIs with nearly identical sequences (i.e., repeat-in-repeat).[29–31] Short-read approaches are further limited by typically capturing only insertion junctions,[32] leaving complete insertion sequences unresolved, which often makes it difficult to confirm TPRT features and trace parental source elements. In cancer genomes, only 7.5%–22.8%[15,33] of sMEIs detected using short reads can be traced back to their sources, since source assignment relies on transduction sequences, which are unique flanking sequences occasionally co-mobilized with the source elements. Because most sMEIs are present in only a small subset of cells within bulk tissue samples, their VAF is typically very low, which poses a significant challenge.[34,35] Several studies have reported sMEIs occurring in the human brain at such low VAFs, ranging between 0.09% and 0.4%[21] or 0.53% and 1.2%.[23] When reads are unevenly distributed between the two haplotypes, it becomes especially difficult to distinguish somatic events from heterozygous germline MEIs.

Long-read whole-genome sequencing (WGS) combined with analytical innovations provides opportunities to map sMEIs and gain biological insights into their epigenetic regulation, as most platforms provide CpG methylation information.[36,37] Long reads that span large stretches of DNA—capturing entire insertions along with both junctions in a single molecule—can substantially improve sMEI detection. Yet even with long reads, reference bias persists when paralog content and sequences differ between the sample and the reference genome.[38] Personalized, haplotype-resolved, donor-specific genome assemblies (DSAs) overcome this limitation by accurately representing the individual's alleles and copy-number structure.[39,40] Aligning long reads directly to DSAs provides a subject-matched coordinate system that normalizes germline divergence in variant calling, reduces misalignment across paralogous loci, and improves read placement through SDs and other complex repeats. In particular, long reads link heterozygous germline variants and candidate sMEIs on the same molecule, enabling read-backed haplotype phasing against these assemblies[41]; this phasing information allows low-VAF signals to be correctly assigned to proper haplotypes and distinguished from noise. Recent studies have begun exploring the use of DSAs to phase read alignment and improve somatic mutation discovery in limited samples, such as cancer samples and individual brain tissues.[42,43] However, the application of DSAs to somatic discovery, especially for sMEIs within complex regions, and systematic evaluations across different samples are still lacking. Long-read-based MEI analysis also enables direct profiling of internal repeat sequence variation, which may allow more comprehensive tracing of a parental source locus without relying on transductions.[44]

For cost-effective MEI profiling, several targeted approaches have been developed.[19,25,45–47] These methods enrich DNA templates containing MEIs and their genomic flanks, followed by either short-[19,25,45,46] or long-read sequencing,[47] thereby enabling sensitive and scalable discovery of rare insertion events. Together, these advances establish numerous advantages of applying modern sequencing technologies and analytical approaches to investigate the ongoing mutagenesis in the human body caused by sMEIs. However, an integrative framework to fully leverage these technologies through systematic benchmarking is still lacking.

Here, we propose a multi-platform integrative framework for accurate sMEI detection and source tracing and demonstrate its applicability in normal human tissues from the Somatic Mosaicism across Human Tissues (SMaHT) Network.[48] We first weigh the strengths and the weaknesses of multiple sequencing platforms using a mixture of six highly characterized HapMap cell lines. For this, we report a high-confidence benchmarking callset for L1, *Alu*, and SVA, stratified into tiers based on the annotated TPRT features, and evaluate the detection performance of seven WGS-based computational methods and two MEI-targeted sequencing methods. We then propose an integration strategy whereby both insertion calls and raw signals from individual reads are combined for improved detection accuracy. We further demonstrate that haplotype phasing and DSAs enable more accurate refinement of sMEI candidates and that our new source-tracing method based on internal L1 sequence variation reveals allele-specific source L1 activities using *in silico* tumor-normal mixture experiments. We validate the applicability of our framework in normal tissue samples (liver, lung, colon, and brain) from four *postmortem* human donors and characterize sMEIs and their source diversity, including the discovery of a somatic chimeric insertion involving small nuclear RNA (snRNA) and L1HS. Together, our work provides a practical framework for accurate sMEI mapping and expands our understanding of sMEIs in normal human tissues.

RESULTS

Benchmark design for sMEI detection in HapMap mixture

Our aim is to establish a practical framework for sMEI analysis. As a first step, we sought to systematically benchmark sMEI detection approaches across sequencing platforms,

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

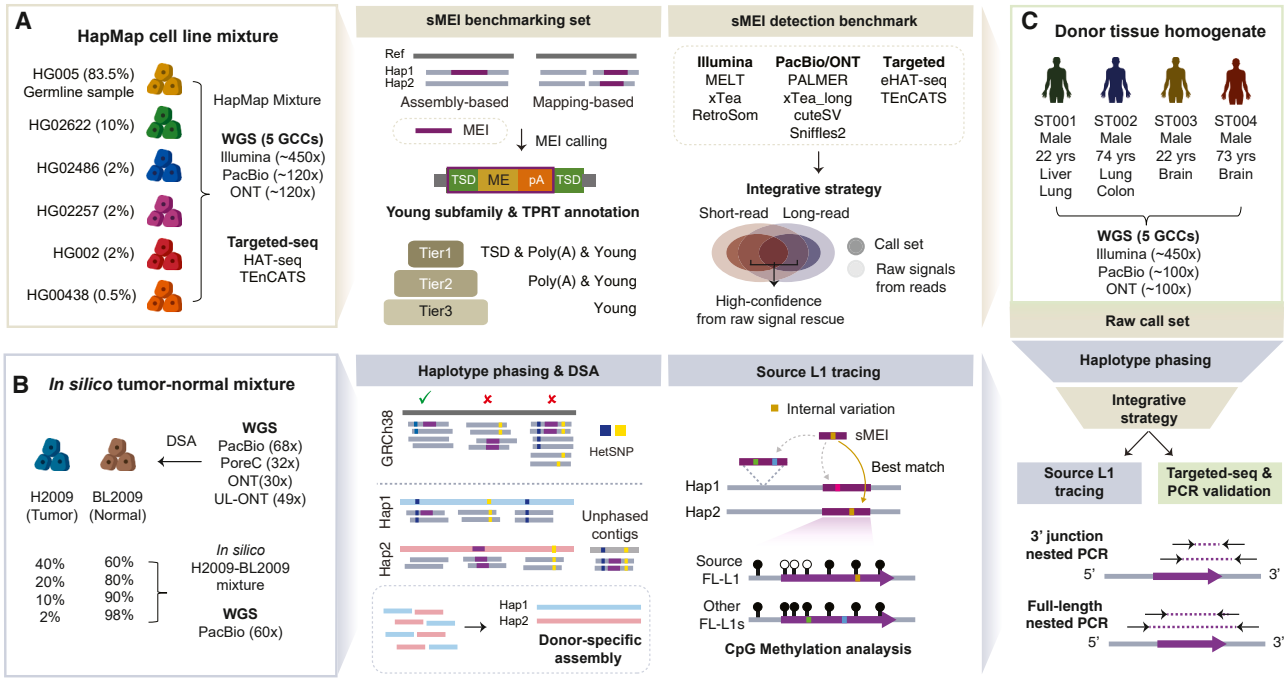


Figure 1. Overview of integrative framework for mapping sMEIs

(A) Benchmarking sMEI detection methods using HapMap mixture from the SMaHT network. A benchmarking set was generated from PacBio and haplotype assemblies and stratified into three tiers by MEI subfamily and TPRT features. Nine sMEI detection methods were evaluated across sequencing platforms, depths, callset concordance, genomic regions, and VAFs. Integrative strategy incorporates both calls and raw signals from the reads to expand high-confidence sMEI candidates. GCC stands for Genome Characterization Center in the SMaHT network.

(B) Refinement of sMEI detection using haplotype phasing and donor-specific assembly (DSA) and tracing of source L1s using internal sequence variation. The *in silico* tumor-normal mixture (CASTLE project) dataset was generated to simulate sMEIs at various VAFs. Source L1s were traced using internal L1 sequence variation, and their CpG methylation status was analyzed. FL-L1, full-length L1.

(C) The discovery of sMEIs in normal human tissues by applying the integrative framework for sMEI detection analysis. Each step established from HapMap mixture (sMEI detection method benchmark and multi-platform integration) and tumor-normal mixture (haplotype phasing/DSA and source L1 tracing) were integrated. The final sMEI candidates were confirmed by MEI-targeted sequencing and nested PCR.

sequencing depths, genomic contexts, and VAF levels. To enable such comparisons and to assess the strengths and the limitations of each platform and variant-calling method, we generated a high-confidence sMEI benchmarking set (Tables S1, S2, and S3). Specifically, we identified non-reference germline MEIs with TPRT signatures in six HapMap cell lines (HG005, HG02622, HG02486, HG02257, HG002, and HG00438) and mimicked sMEIs at varying VAFs by mixing these cell lines at different proportions (Figures 1A and S1A).[49] In this HapMap mixture design, non-reference MEIs present in HG005 (83.5%, backbone genome) represent germline insertions, whereas those present in the other five cell lines but absent in HG005 represent sMEIs (Figure S1B). The benchmarking set was stratified into three tiers by annotating MEI subfamilies and TPRT features, such as poly-A tail and target site duplication (TSD), ranging from the most conservative (tier 1) to the most permissive (tier 3) (STAR Methods and Figure S1A). The number of insertions in the benchmarking set varied by the class of mobile elements and their respective tiers, with *Alu* (n = 2,266, 2,431, and 2,543 in tiers 1–3, respectively) showing the highest counts, followed by L1 (n = 392, 466, and 522) and SVA (n = 171, 179, and 420) (Table S4; Figure S2). The VAF, calculated by the mixing ratios (0.5%–10%) and genotypes in individual cell lines, ranged from 0.25% to 16.5% across all three elements ("expected VAF"; STAR Methods and Figure S2D).

Most existing MEI detection tools were developed for identifying germline insertions or highly clonal sMEIs in cancer,[32] and their performance when applied to sMEIs in normal tissue has not been evaluated. We selected nine sMEI detection approaches, including seven WGS-based computational methods (STAR Methods) and two MEI-targeted sequencing methods for benchmarking. For methods originally developed for germline or highly clonal events (xTea_long,[31] cuteSV,[50] and Sniffles2[51]), we adjusted parameters by lowering supporting read counts; for general SV callers (cuteSV and Sniffles2), we additionally applied RepeatMasker[52] to identify sMEI candidates among their insertion calls. For the WGS-based methods, we used high-depth bulk WGS data of the HapMap mixture provided by the SMaHT Network, including 450× short-read (Illumina), 110× long-read (PacBio and ONT, with N50 > 1.6 kb and >2.6 kb, respectively), and their down-sampling data (50×, 100×, 200×, and 300× for short reads; 30×, 60×, and 90× for long reads) (Figure 1A; Table S5). Two MEI-targeted sequencing assays, enhanced human active transposon sequencing (eHAT-seq)[19] and TEnCATS,[47] were applied to the genomic DNA (gDNA) of the

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

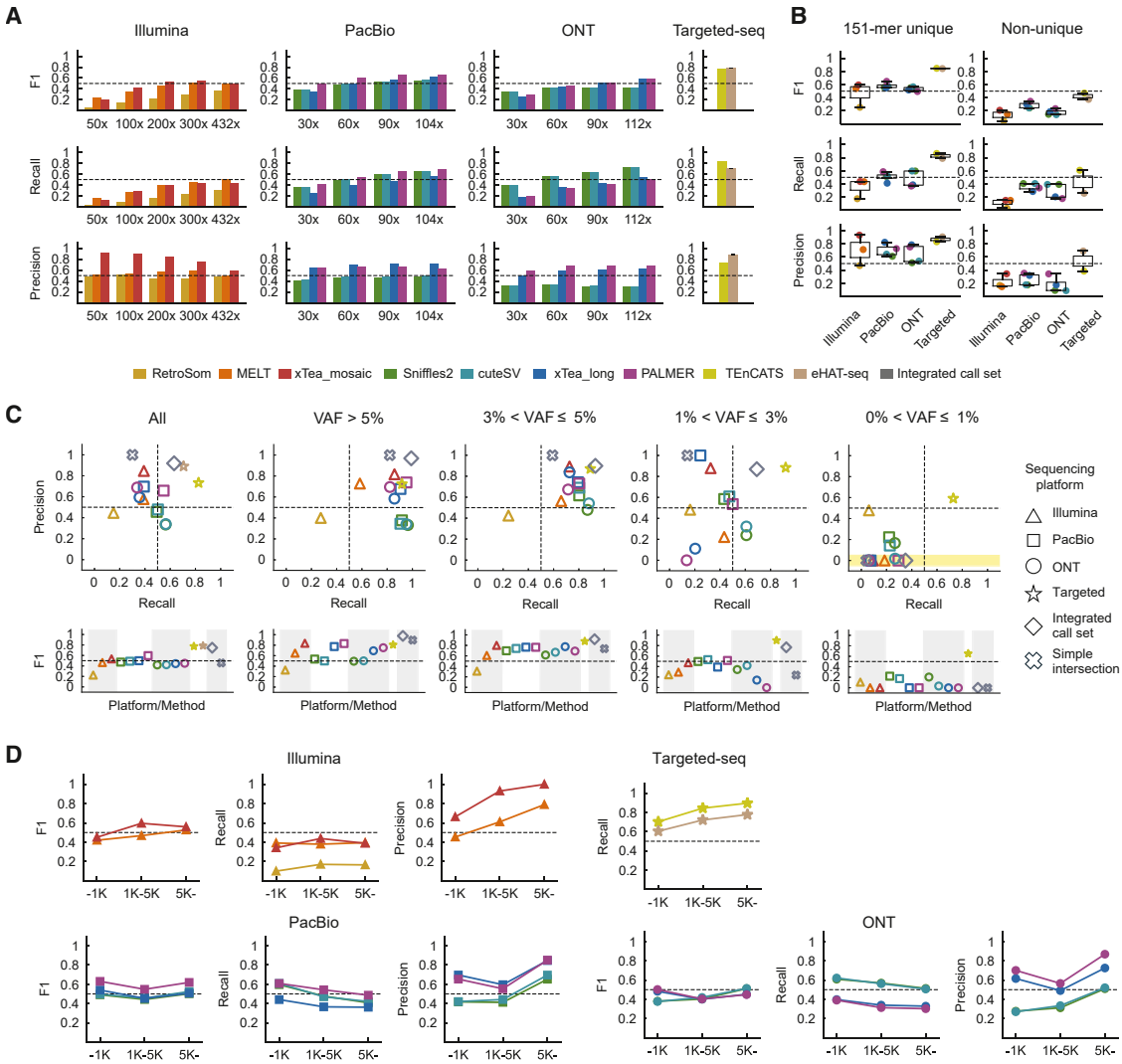


Figure 2. sMEI detection method benchmark using HapMap mixture

(A) F1 scores, recall, and precision across sequencing platforms (Illumina, PacBio, ONT, and MEI-targeted sequencing) at various sequencing depths (full-coverage WGS from BCM GCC and their downsampling data). eHAT-seq performance is the average of four experimental replicates, and the error bar was calculated based on the standard deviation.

(B) F1 scores, recall, and precision across 151-mer unique vs. non-unique regions.

(C) F1 scores, recall, and precision in each VAF bin. Platforms are indicated by shape. The dots located in the yellow area in the recall-precision scatterplot indicate that the corresponding method does not report the calls in the corresponding VAF bin. For eHAT-seq, only recall is shown, because caller-level VAFs are not available for precision calculation.

(D) F1 scores, recall, and precision across insertion length. For RetroSom, eHAT-seq, and TEnCATS, only recall is shown, because caller-level insertion lengths are not available for precision calculation.

(A–D) The same color key represents detection methods. (B–D) 200× and 60× callsets were used for short-read (Illumina) and long-read (PacBio and ONT) WGS data, respectively.

HapMap mixture as orthogonal technologies (STAR Methods and Table S6). We report our results mainly focusing on L1 with the middle-tier (tier 2) benchmarking sets, as L1s are the primary drivers of sMEIs in somatic tissues, as confirmed by our donor tissue analysis in this study. The results for the most conservative benchmarking set tier (tier 1, Figure S3) and other element types (*Alu* and SVA, Figure S4 and supplemental notes) are provided in the supplemental information.

Long-read and MEI-targeted sequencing improves sMEI detection at low VAFs and in challenging genomic contexts

We first evaluated the detection performance across sequencing platforms and depths using the benchmarking sMEI set (Figure 2A; Table S7). In the short-read data, we observed that the overall performance (F1 score) plateaued at ~200× (xTea_mosaic and MELT), marking it as the recommended depth that

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

balances cost and performance. The performance peaked at 300×, followed by a modest decline beyond this depth, such that the difference in performance between 200× and 300× was subtle (<0.02 in F1 score). In the long-read data, such F1 scores were achievable at a much lower depth of ~60×. For instance, xTea_mosaic showed the best performance (F1 score = 0.53) at 200× Illumina among short-read WGS-based methods, a level matched by PALMER using only 60× PacBio (F1 score = 0.60). The recall particularly improved from 30× to 60× across the long-read WGS-based methods (14.0% and 16.3% higher on average in PacBio and ONT, respectively). At the same depth, PacBio yielded 1.2-fold higher F1 scores than ONT across the sequencing depth and method on average. Both short-read (eHAT-seq) and long-read (TEnCATS) MEI-targeted sequencing methods showed higher performance than any WGS-based method.

We next examined whether the detection performance varied across genomic regions (Figure 2B) and used 200× Illumina and 60× PacBio/ONT as the representative sequencing depths (Figures 2B–2D). We used a list of pre-defined, sample-agnostic genomic easy-to-map regions based on 151-mer uniqueness in pan-genome assemblies.[53] In our benchmarking set, 84.1% sMEIs fell into the 151-unique region (Table S4). The recall dropped sharply in the non-unique regions for all short-read WGS methods (10.8% on average), whereas long-read WGS-based methods maintained substantially higher recall (3.3-fold in PacBio and 2.7-fold in ONT) than short-read methods. These results strongly demonstrate that improved mappability with long reads broadens the callable genomic regions for sMEIs.

Moreover, we assessed detection performance across different VAF ranges (Figure 2C). Detection patterns in variants with VAF > 3% were largely consistent with those observed in the overall callset. At the VAFs lower than 3%, however, long-read WGS-based methods and MEI-targeted sequencing methods outperformed short-read methods, particularly in recall. When the VAFs reach extremely low levels (≤1%), long-read-based methods achieved 1.6-fold higher recall than short-read-based methods on average. This observation was consistent in a different benchmarking set (Figure S3) and for *Alu* and SVA (Figure S4; supplemental notes; Tables S8 and S9). This underscores the critical role of long-read sequencing in detecting low-frequency insertions in normal human tissues. The MEI-targeted sequencing showed even higher recall in this VAF range (eHAT-seq: 65.2%, TEnCATS: 72.8%), demonstrating a cost-effective alternative to WGS-based approaches for detecting ultra-low VAF sMEIs.

Lastly, given that L1s have a relatively long full-length size (~6 kb) compared to other elements and that somatic L1 insertions are frequently truncated, we examined how insertion length influences detection performance (Figure 2D). In short-read WGS-based methods, recall did not show a significant dependence on insertion length. In contrast, long-read WGS-based methods exhibited a length-dependent decline in recall, with lower sensitivity for longer insertions. This trend was most pronounced in the 1% < VAF ≤ 3% range and was particularly evident for PacBio data, where recall for insertions shorter than 1 kb was 2.6-fold higher than for insertions longer than 5 kb (Figure S5). By contrast, both MEI-targeted sequencing methods showed the opposite pattern, with recall increasing with insertion length, especially for ultra-low-frequency insertions (VAF ≤ 1%) (Figure S5). Overall, longer insertions showed higher precision for both short-read and long-read WGS-based methods. These distinct performance characteristics should therefore be carefully considered when interpreting the observed length distribution of sMEIs across different sequencing platforms and analytical approaches.

Integrating callsets with raw read-alignment signals enhances high-confidence sMEI detection

To explore an effective integration strategy across the callsets, we evaluated the concordance of the true-positive (TP) and false-positive (FP) calls between the methods (Figure 3A). Except RetroSom, which shows relatively low recall (Figure 2A), combinations of two callsets show 37.1%–98.5% TP concordance. In contrast, FP concordance was markedly lower between the callsets derived from different platforms (0%–2.0%), suggesting that cross-platform integration can effectively filter out FPs. Indeed, simple intersection of short- and long-read callsets (S1) achieved near-perfect precision with no overlapping FPs but reduced recall (recall = 0.30) (Figure 3B; Table S10). Notably, 87.2% of calls in this intersection set had VAF above 3%, indicating that low-VAF calls and those with insufficient supporting reads were frequently missed (Figure 3C).

To mitigate the recall loss, we leveraged raw read-level signals to rescue calls detected by one platform but missed by the other, matching insertion sequences at the same breakpoint (STAR Methods). This rescue strategy (S1–S3 in Figure 3B) improved F1 by 1.5-fold (F1 = 0.70, Figure 3D), with more TPs supported by both platforms (recall = 0.54), while maintaining high precision of 0.98 (Table S10). Notably, the rescued set captured insertions across distinct VAF levels, insertion lengths, and genomic regions (Figure 3C). For instance, only 9.5% of TPs called by short reads and rescued by long-read alignment signals (S2) showed VAF ≤ 1%, whereas 42.2% of TPs called by long reads and rescued by short-read alignment signals (S3) showed VAF ≤ 1%. Additionally, the largest number of sMEIs in non-unique regions was called by long reads with short-read alignment signals (S3). In contrast, more long insertions (>5 kb) were called by short reads with long-read alignment signals (S2), reflecting the strength of short reads in detecting 3′ and 5′ insertion junctions regardless of insertion length. These patterns highlight that multi-platform integration, combined with raw read-level rescue, enables the expansion of high-confidence sMEIs, complementing the limitations of each platform.

We further examined single-platform-only calls (S4 and S5) to reduce FPs by designing additional filtering schemes (Figure S6). For short-read-based xTea_mosaic calls, many FPs originated from the non-unique region (Figures 2C and S6A), and restricting the calls to the unique region removed 66.7% of FPs with nearly no recall loss (only one TP excluded) (S4′ in Figure 3B). For PALMER, two filters were effective: the old subfamily filter removed 82.1% of FPs without excluding any TP, while the sole-supplementary alignment filter removed 56.0% of young subfamily FPs. Together, these filtered out 91.2% of FPs without losing any TPs (S5′ in Figures 3B, S6C, and S6D). Notably, 75% of S5′ TPs were ultra-low mosaic sMEIs with VAF ≤ 1%

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

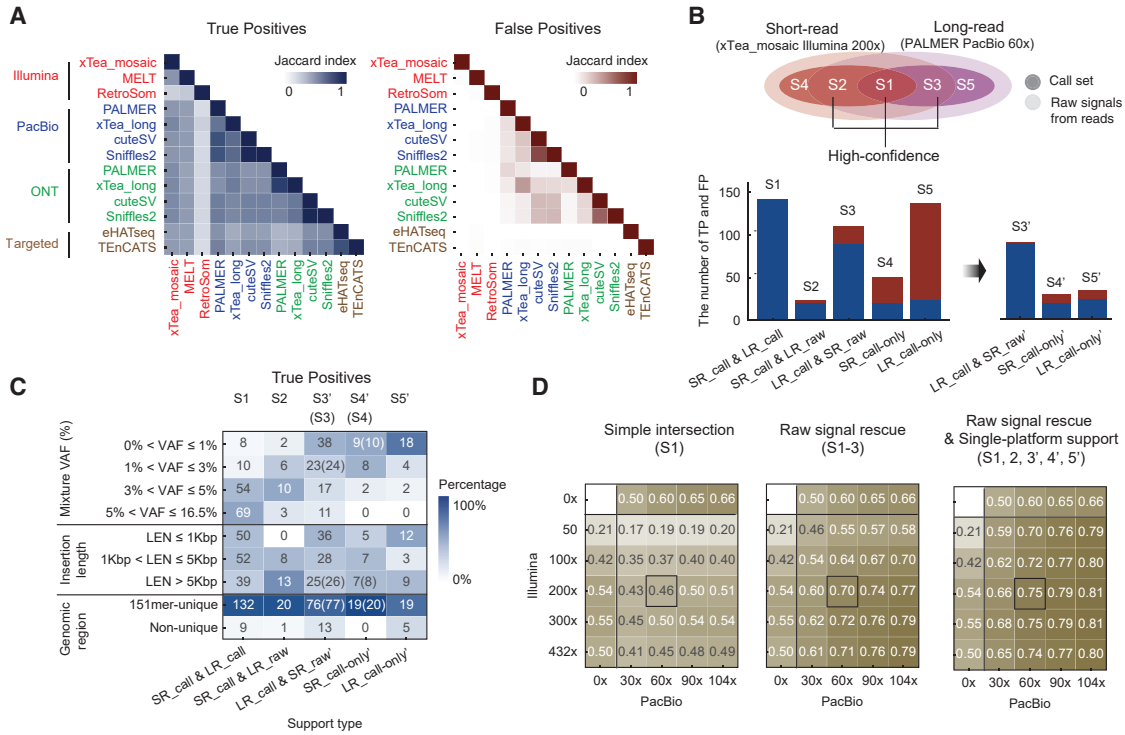


Figure 3. True-positive and false-positive similarity and integrative strategy (short and long read) using HapMap mixture

(A) Callset similarity (Jaccard index) of true positives (TPs) and false positives (FPs). Sequencing platforms are grouped by the same color (Illumina, red; PacBio, blue; ONT, green; MEI-targeted sequencing methods, brown). 200× and 60× callsets were used for short-read (Illumina) and long-read (PacBio and ONT) data, respectively.

(B) Schematic diagram of integrative strategy combined with raw signal rescues and the number of TP and FPs by each support type (S1–S5) before and after FP filtering (S3′–S5′). Support types are noted in the *x* axis.

(C) Percentages and counts of TPs across VAF bins, insertion length bins, and genomic regions, shown for each support type. Numbers of changed TP changed before filtering (S3 and S4) compared to after filtering (S3′ and S4′) are shown in parentheses. There was no difference in TPs between S5 and S5′.

(D) F1 scores of multi-platform integrated callsets across different Illumina and PacBio coverage combinations.

(Figure 3C). The same PALMER filters applied to the rescued callset S3 removed 95% of FPs (S3′).

This enhanced integration callsets (S1–S2 and S3′–S5′ in Figure 3B) achieved 1.6-fold higher detection performance (F1 = 0.75, recall = 0.63, and precision = 0.92) at Illumina 200× and PacBio 60×, compared to that of simple intersection between the two callsets (S1) (Figure 3D). Notably, the largest performance gain was observed in the 1% < VAF ≤ 3% range (F1 = 0.77), where recall particularly increased by 4.9-fold compared to the simple intersection (Figure 2C; Table S10). For ultra-low-frequency insertions (VAF ≤ 1%), recall increased by 8.8-fold (Figure 2C; Table S10). Compared to MEI-targeted sequencing, the integration callset shows comparable performance for insertions with VAF > 3%, whereas MEI-targeted sequencing retained higher recall at lower VAFs, particularly for ultra-low-frequency insertions (VAF ≤ 1%) with a 2.1-fold higher recall for TEnCATS (Figure 2C). When considering all combinations of Illumina and PacBio sequencing depths, the F1 scores were, on average, 2.1-fold higher than those obtained using simple intersections (Figure 3D; Table S10). A similar improvement was observed for the combinations of Illumina and ONT data, with an average 1.9-fold increase in F1 scores (Table S11). The performance of our integrative strategy was further validated using independent short-read and long-read WGS data (Table S12).

Phasing and mapping to DSA increase precision in sMEI detection

Accurate read mapping is essential for alignment-based variant detection; however, repetitive genomic regions such as SDs remain particularly challenging. For instance, inter-individual differences in SD copy number can cause reads originating from extra copies (absent in the reference genome) to misalign to existing loci, leading to FP somatic variant calls with low VAFs.[27] These artifacts can be effectively mitigated using long-read sequencing in combination with haplotype phasing and DSA (Figure 4A), which provides improved resolution in complex genomic regions.[43] The application of phasing and DSA to HapMap mixtures is inherently limited due to their synthetic nature, where multiple genomes are mixed. To address this, we generated *in silico* tumor-normal (H2009-BL2009) mixtures from the CASTLE project,[54,55] enabling the evaluation of phasing and DSA as refinement strategies for sMEI detection.

We first generated a high-quality haplotype-resolved DSA using the BL2009 normal cell line data (STAR Methods and

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

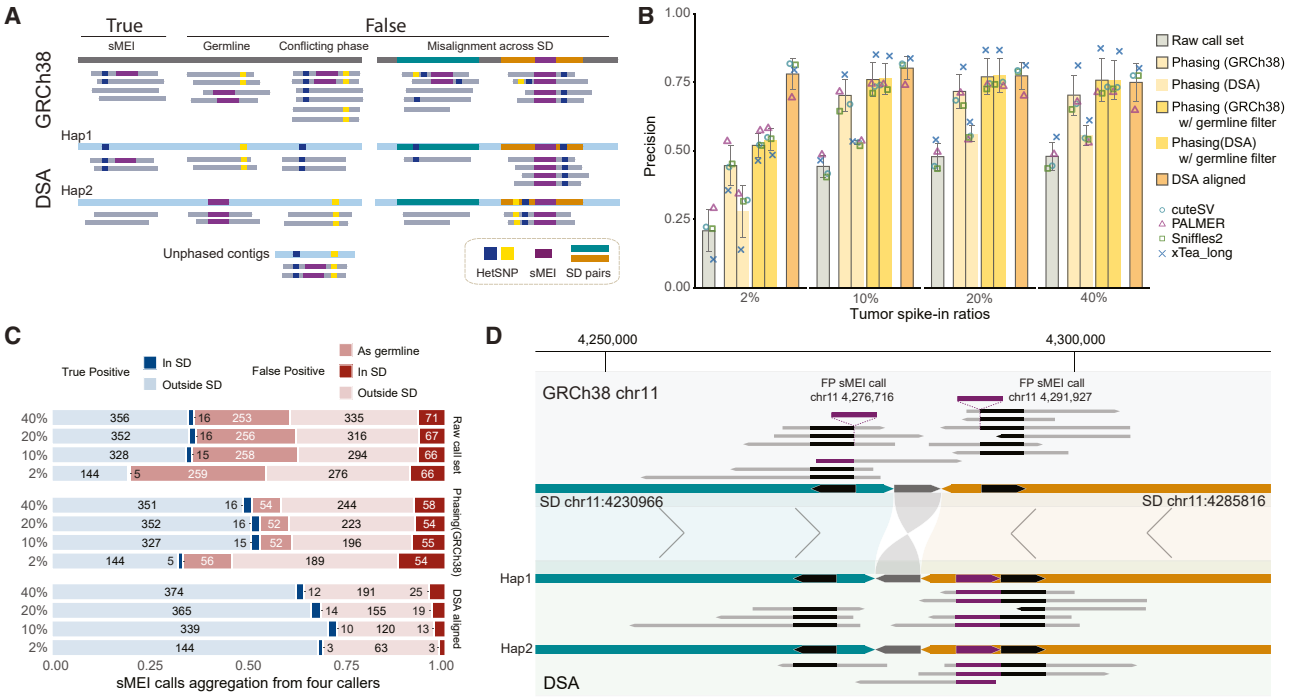


Figure 4. Refinement of sMEI detection from haplotype phasing and DSA analysis

(A) Schematic illustrating how phasing information and DSA are leveraged to mitigate errors caused by germline MEIs, conflicting signals, or segmental duplications (SDs).

(B) Precision of sMEI calls across four tumor spike-in mixtures (2%, 10%, 20%, and 40%). Bars indicate precision across four detection methods (cuteSV, PALMER, Sniffles2, and xTea_long) using GRCh38-based initial alignment, DSA-based alignment, and various refinement methods (phasing on GRCh38, phasing on DSA, GRCh38 phasing with germline site filter, and DSA phasing with germline site filter); colored points show per-caller values.

(C) Composition and numbers of deduplicated sMEI call observations classified as true positives (TPs) and false positives (FPs) inside vs. outside SDs, and FPs overlapping population-polymorphic germline MEI sites, aggregated across four callers. Results are shown for GRCh38-based alignment, phasing on GRCh38, and DSA-based alignment, based on unique observations.

(D) Mis-mapping in an SD pair generates false sMEI calls on GRCh38 but not on the DSA. Upper: GRCh38 shows two 5.0 kb reference blocks with L1 signal (black) within the inverted SD pair (chr11:4,230,966, green, and chr11:4,285,816, orange). Reads carrying an additional nearly identical 5.0 kb L1 sequence (purple) align arbitrarily into the two blocks, yielding two apparent sMEIs at chr11:4,276,716 and chr11:4,291,927. Lower: in the DSA, a germline tandem duplicate (purple) was constructed in one SD; reads align consistently to both SD regions with no somatic signals.

Figure S7). To simulate varying sMEI frequencies, we generated four WGS datasets by spiking tumor sequencing reads (PacBio 60×; N50 > 1.6 kb) into the normal dataset at rates of 40%, 20%, 10%, and 2% (Figure 1B; Table S5). To evaluate the performance of sMEI discovery in these mixtures, we followed a similar pipeline as for the HapMap mixture described earlier to generate a cancer benchmarking set of sMEIs in the tumor H2009 cell line (L1 only, n = 315, 376, and 379 in tiers 1–3, respectively), excluding calls present in normal cell line data (STAR Methods and Tables S13 and S14). The sMEIs in the cancer benchmarking set of these mixtures followed the expected distribution of VAFs, ranging from 1% to 20% (Figure S8A). We next evaluated the performance improvement from the phasing and direct sequence assembly in three ways: (1) phasing information extracted from the sequencing reads aligned to the GRCh38 reference genome; (2) phasing information derived from the assembly contigs aligned to GRCh38; and (3) calling efficacy directly from the sequencing reads aligned to the DSA (Figures 1B and 4A). We further applied a population-level germline MEI filter to these sets by excluding polymorphic insertions previously detected by large-scale sequencing projects (STAR Methods).

Using phasing information from long-read sequencing, we performed local, read-backed phasing (see STAR Methods) around each candidate insertion as the first refinement method. Relative to the pre-refined calls, precision increased by 22.3%–70.3% in the 40% tumor spike-in sample (Figure 4B; Table S15), removing an average of 43.4% of FPs across callers. For assembly-derived phasing, we utilized the empirical decision boundary (Figure S8 and STAR Methods) established in a previous study[43] to distinguish between FPs and potential TPs, achieving a 98.4% recall rate within the cancer benchmarking set and yielding a 7.7% gain to an average of 55.7% precision in the 40% tumor spike-in sample. Applying the germline MEI filter further increased precision in the 40% mixture for both read-backed and assembly-derived phasing by 5.5% and 20.1%, respectively, reaching mean precisions of 75.7% and 75.8%, while maintaining recalls up to 91.3% and 86.4%. We observed a similar pattern at other mixture levels. Notably, improvements extended into the most challenging, lowest mixture-level

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

sample: in the 2% spike-in, precision rose from 21.2% to 51.9% and 54.0% across callers for read-backed phasing and assembly-derived phasing, respectively, highlighting the value of phasing information for ultra-low-frequency sMEI detection (Figures 4B and S9).

We then asked whether aligning reads directly to a donor-matched DSA would further suppress reference-mismatch artifacts, particularly under low-tumor-fraction conditions where weak somatic signals are most vulnerable to mis-mapping and germline-confounding effects. Because a DSA intrinsically encodes the donor's germline alleles and architecture, direct DSA alignment accounts for donor-specific germline divergence and reduces mis-mapping in SDs and other complex genomic regions, thereby facilitating the identification of *bona fide* somatic insertion signals absent from the DSA. This benefit was most pronounced in the 2% mixture, where DSA alignment increased mean precision 3.7-fold (from 21.2% to 78.0%) relative to the raw callsets. DSA-aligned callsets also remained among the highest precision strategies at higher tumor spike-in ratios, indicating that direct alignment to a donor-matched assembly provides robust refinement across mixture levels while offering the greatest relative advantage under ultra-low-frequency conditions. Together, these results highlight the particular value of DSA alignment for ultra-low-frequency sMEI detection, a key regime for somatic mosaic variant discovery, where low allelic support makes reference-mismatch artifacts difficult to distinguish from true somatic insertions.

We further systematically assessed TP and FP sMEIs in the context of SDs and known germline sites across the three refinement methods. Read-derived phasing methods substantially refined sMEI signals (Figure 4C). For example, in the 2% mixture, the aggregated number of unique FP calls across callers in SD decreased from 66 in the pre-refined sets to 54, while FPs in germline sites decreased from 259 to 56. Notably, calls from DSA-based alignment yielded sparse FPs in the SD regions, ranging from 25 (40% mixture) to 3 (2% mixture). Additionally, DSA-aligned calls showed no overlap with predefined germline MEI sites (STAR Methods), consistent with effective suppression of germline-associated FPs. We observed that DSA substantially reduced the proportion of SD-localized FPs (adjusted odds ratio [OR] = 0.61, 95% confidence interval [CI] 0.50–0.73; Wald $z = -0.73$, $p = 9.95 \times 10^{-8}$; STAR Methods), leading to a clear shift in call compositions across refinement methods (Figure 4C) while preserving TPs.

As a case study, we examined two potential FP somatic L1 insertions that fell within SD regions as aligned to GRCh38: a 5.0 kb insertion (A) at chr11:4,276,716 and another (B) of the same length at chr11:4,291,927, both concordantly detected by cuteSV and Sniffles2 in the 40% spike-in sample, with insertion (A) additionally detected by xTea_long (Figures 4D and S10). One copy of the SD pair at chr11:4,230,966 and its inverted homologous copy at chr11:4,285,816 both contain a 5.0 kb reference segment with L1 signal. This segment was confirmed to be 99% identical to the insertion sequences, supported by 11 of 46 reads in call A and 9 of 74 reads in call B, yielding two putative somatic L1 calls. However, in the haplotype-resolved DSA, we identified a germline homozygous signal in one SD pair (chr11:4,285,816), representing a tandem duplication of the reference segment. When realigning the same molecules to the DSA, all supporting reads for both A and B were redistributed to the reconstructed SD (chr11:4,285,816 in GRCh38), which carries this germline tandem duplicate with L1 signal. In contrast, the other DSA SD copy (chr11:4,230,966 in GRCh38) showed no evidence of L1 insertion. This provides strong evidence that the DSA successfully reconstructed the exact insertion segment in both haplotypes, while in GRCh38 the reads were mis-mapped and arbitrarily assigned between these duplicated SD blocks (Figure 4D). This showcases how two FP L1 insertions can be introduced by a homozygous MEI germline event on the same SD copy and illustrates how homologous SD pairs in the traditional human reference introduce ambiguous alignments, leading the supporting reads for a germline insertion in one SD to be misdistributed and misinterpreted as sMEI at one or both loci. By reconstructing the germline insertion segment, the DSA facilitates accurate mapping and eliminates these false somatic signals.

Internal L1 sequence variation enables allele-specific source tracing

The human genome contains ~100 full-length L1HS (FL-L1HS) loci that are retrotransposition competent.[56] However, which FL-L1HS loci contribute to somatic retrotransposition in human tissues is still an open question. Traditionally, tracing somatic L1 insertions back to their source loci requires 3′ transduction sequences, leaving a substantial fraction of insertions untraceable.[44] Only 8.8%[57] to 15%[29] of germline L1 insertions contain transduction sequences, while for somatic L1 insertions, this number can range from 11% in normal colon tissues[24] to 35% in gastrointestinal tumors.[44]

Long-read sequencing enables the profiling of complete L1 insertion sequences, giving an opportunity to use the internal L1 sequence for source tracing. To assess the feasibility of this approach, we first collected reference and non-reference FL-L1HS sequences from the BL2009 cell line, referred to as the FL-L1HS catalog, totaling 627 unique sequences (566 reference and 61 non-reference; STAR Methods and Table S16). Next, using a transduction-based tracing approach, we established an evaluation set of 49 insertions confidently linked to 17 source loci (Table S17). These source sequences showed considerable sequence diversity, suggesting the possibility of source tracing using the sequence variation (Figure S11A). We developed a source-tracing pipeline that leverages L1 internal sequence variation compared against the FL-L1HS catalog (Figure 5A and STAR Methods). Applied to the transduction-based evaluation set, our pipeline successfully mapped 20 of 49 insertions (40.8% recall) to their sources with 95.2% precision (Figure 5A; Table S17).

We then applied the pipeline to 376 somatic L1 insertions identified from the H2009 cancer WGS and successfully traced 143 (38.0%) insertions to 29 unique source loci (Figure 5B; Table S17). Compared to the transduction-based method alone, our approach traced 2.9-fold more insertions to 1.7-fold more source L1s (Figure S11B). However, 29 (59.2%) of 49 insertions with 3′ transduction could only be traced by their transduction sequences, indicating that the two strategies are complementary and should be combined for comprehensive source tracing. Using both approaches, we traced sources for 171 (45.5%) insertions in

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

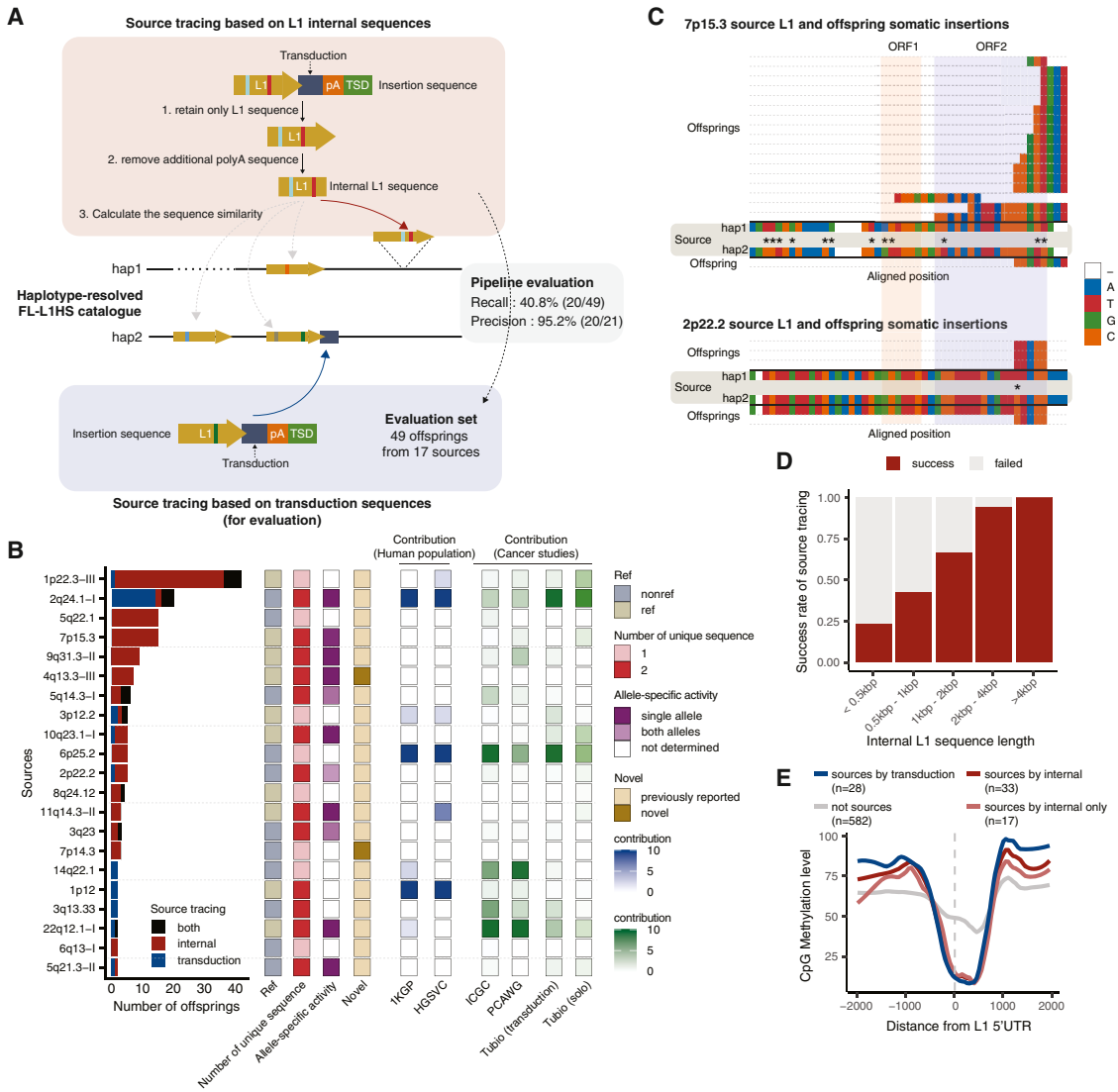


Figure 5. L1 source tracing using internal sequence variations in tumor-normal mixture

(A) Schematic overview of the transduction-based and internal sequence-based source-tracing pipelines. Recall and precision are shown for the internal sequence-based source-tracing method.

(B) Comprehensive characterization of L1 source loci with at least two offspring. The left bar chart shows the number of offspring from each source, with color indicating the prediction method (transduction, internal sequence, or both). The adjacent columns provide detailed annotations for each source locus, including reference (ref) or non-reference (nonref) status, unique allelic sequences, allele-specific activity, and source activity reported in human population studies (1KGP[58] and HGSVC[30]) and cancer (ICGC,[33] PCAWG,[6] and Tubio[44]) studies. For allele-specific activity, sources detected only by the transduction method and containing only one unique allelic sequence are shown in white, as allele-specific activity cannot be calculated for them. For other sources inferred from L1 internal sequence variation, if two unique allelic sequences are present at the source locus, allele-specific activity was calculated and was colored according to the major haplotype.

(C) Examples of two sources (7p15.3 and 2p22.2) and their offspring. Multiple sequence alignment of offspring sequences and source sequences was performed using Clustal Omega.[59] Bases differing among the four source sequences were extracted, with asterisks indicating allele-specific variants.

(D) Success rate of source tracing using L1HS internal sequence variations depending on the internal L1 sequence size.

(E) DNA methylation level at the 5′ UTR (±2 kb) of FL-L1HS for four source categories of source loci in the H2009 tumor cell line.

total (Figure 5C). As expected, sources identified by transduction were often previously reported, whereas several identified by L1 internal sequence variation were novel (Figure 5B).

Our pipeline can reveal allele-specific source activities in homozygous FL-L1HS loci. We calculated the allele-specific contributions of source loci with two unique sequences using the internal sequence variations. Notably, source L1 transcripts are often derived from one allele, suggesting that both allele-specific sequence variation and heterogeneous epigenetic status may influence the allelic activity of the source L1 element (Figure 5B). For example, a source L1 element at the 7p15.3 locus produced the highest number of somatic insertions but with a strong allelic

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

bias (Figure 5C). By contrast, a source L1 at the 2p22.2 locus contributed insertions equally from both alleles (Figure 5C). The two haplotypes differed by 12 base pairs at 7p15.3 and by only one base pair at 2p22.2. Since our pipeline relies on internal sequence variation, its performance improves with increasing offspring insertion length. For somatic insertions longer than 2 kb, we successfully traced 90% of events back to their source loci (Figures 5D and S11C).

We further examined allele-specific CpG methylation at the promoters (5′ UTR) of the identified source L1s. In the H2009 tumor cell line, sources identified by our pipeline exhibited significant hypomethylation. Notably, sources identified exclusively by internal sequence variation showed similar hypomethylation to transduction-based sources, reflecting their active epigenetic state (Figure 5E). We found that the promoters of these same sources were hypermethylated in the normal cell line (BL2009), confirming their cancer-specific active state (Figure S11D). Overall, our source-tracing pipeline leveraging L1 sequence diversity significantly expands the number of identifiable source loci.

A multi-platform integrative framework reveals structural and source diversity of low-frequency sMEIs in normal human tissues

Finally, we assessed the applicability of our integrative framework (haplotype phasing, multi-platform integration, and source L1 tracing) for sMEI analysis by analyzing six tissue samples—liver, lung, colon, and brain—from four *postmortem* human donors from the SMaHT Network (ST001–ST004; Figure 1C).[49] Callsets were generated using reads aligned to GRCh38, with sequencing depths of up to 450× Illumina, 100× PacBio, and 100× ONT from each GCC (Genome Characterization Center) (Figure 1C; Table S5). Each filtering step in our framework is illustrated using WGS data of the ST003 brain sample, which has sequencing depths similar to those used as the representative dataset in the HapMap mixture (Figure 6A; Table S18). We identified a total of 21 somatic L1 insertions with variable frequency across samples (Figure 6B; Table S19). Somatic L1 insertions have previously been reported in normal liver,[24] colon,[24] and brain[19,22,23,25] tissues. We also identified somatic L1 insertions in normal lung tissue. Each sample had at least one high-confidence insertion supported by both platforms—called by long reads and rescued by short reads (S3′, Figure S12). Twelve additional L1 insertion candidates were supported exclusively by long reads only (S5′) and passed manual inspection. Given their estimated low VAFs (only one supporting read detected; Table S19), this result coincided with the HapMap mixture benchmark, which showed that low VAF insertions were primarily supported by long reads (Figure 3C).

Re-sampling and validating somatic insertions from bulk tissue are challenging due to their low VAFs and likely focal distribution. Nonetheless, we performed the nested PCR for eight selected candidates by amplifying the 3′ L1-genome junctions (Table S20), using either the same gDNA extracted from the tissue homogenate that was used to generate the corresponding WGS data (ST001 and ST002) or the gDNA extracted from a different core but in the same tissue (all four donors). Three candidates with estimated VAF larger than 0.5% (ST001 liver, ST002 colon, and ST003 brain) were validated in both cases (Figures 6B and S13; Table S19). By contrast, the two insertions in ST002 lung with lower estimated VAF (0.20%–0.25%) were only validated in the same gDNA extracted from the tissue homogenate, and two other insertions in ST002 lung with similar estimated VAF (0.25%) or much lower (below 0.20%) were not validated even in the same gDNA extracted from the tissue homogenate (Table S19). Among the five insertions validated by 3′ nested PCR, three insertions in ST001 liver, ST002 colon, and ST003 brain samples were also detected by TEnCATS (Figure S14; Table S19), and two out of these three insertions (ST002 colon and ST003 brain samples) were also detected by eHAT-seq (Table S19), all performed on different cores from the homogenate used for WGS.

All identified insertions were 5′ truncated (71.4%, 15/21 shorter than 1 kb), with no full-length insertions observed (Figure 6C). Intra-inversion, indicative of twin-priming, was observed in 33.3% (7/21) of the insertions (Figure 6D). It is lower than that reported in pan-cancer analyses (40%),[15] consistent with observations from normal colorectal epithelium (29.5%),[24] but relatively higher than the rate (21%–25%) reported in the cell culture or population studies.[29,57,60] This suggests that twin-priming rates may vary across tissue types and biological contexts. Interestingly, one somatic L1 insertion in a brain sample contained a 3′-truncated U5 snRNA (*RNU5A-1*) fused in an antisense orientation to a 5′-truncated L1HS element, a structural pattern consistent with very rare polymorphic events previously reported[61–64] (Figure 6E). The micro-homology found at the U5-L1 and insertion site junctions is consistent with a twin-priming model of formation.[63] We resolved its sequence and structure using nested full-length PCR followed by Sanger sequencing (Figures 6E and S13). This somatic snRNA/L1 chimeric insertion in human tissue suggests that long-read-based analysis will enable more comprehensive investigation of these non-canonical forms of sMEI in the somatic state.

Our source-tracing pipeline identified eight somatic insertions originating from seven source L1s (Figure 6F). Three insertions (14.3%) with 3′ transductions originated from two reference and one non-reference full-length source L1s. The other five insertions were traced to their source loci based on internal sequence variation. The identified source L1s belonged to various L1HS subfamilies: preTa, Ta0, Ta1nd, Ta1d-CCA, Tad1d-CAT, and Ta1d-TCA (Figure 6F). Notably, two somatic L1 insertions found in the lung and the colon samples of the same donor (ST002) were derived from the same source of Ta1d-TCA subfamily, known active L1HS subfamily[65] (Figure 6F). Eleven (52.4%) somatic insertions were located within introns of 11 different protein-coding genes, including genes associated with tissue-specific diseases. For example, one insertion in ST001 liver occurred in intron 14 of *PTPRM*, a gene involved in cell adhesion,[66] liver fibrosis,[67] and hepatocellular carcinoma.[68] Another insertion in ST004 brain occurred in intron 2 of *CSMD1*, implicated in neurodevelopmental or psychiatric disorders such as schizophrenia[69,70] and intellectual disability.[71] Most somatic insertions detected in normal tissues are likely neutral, but further investigation of their clonal structure across tissues and their functional

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

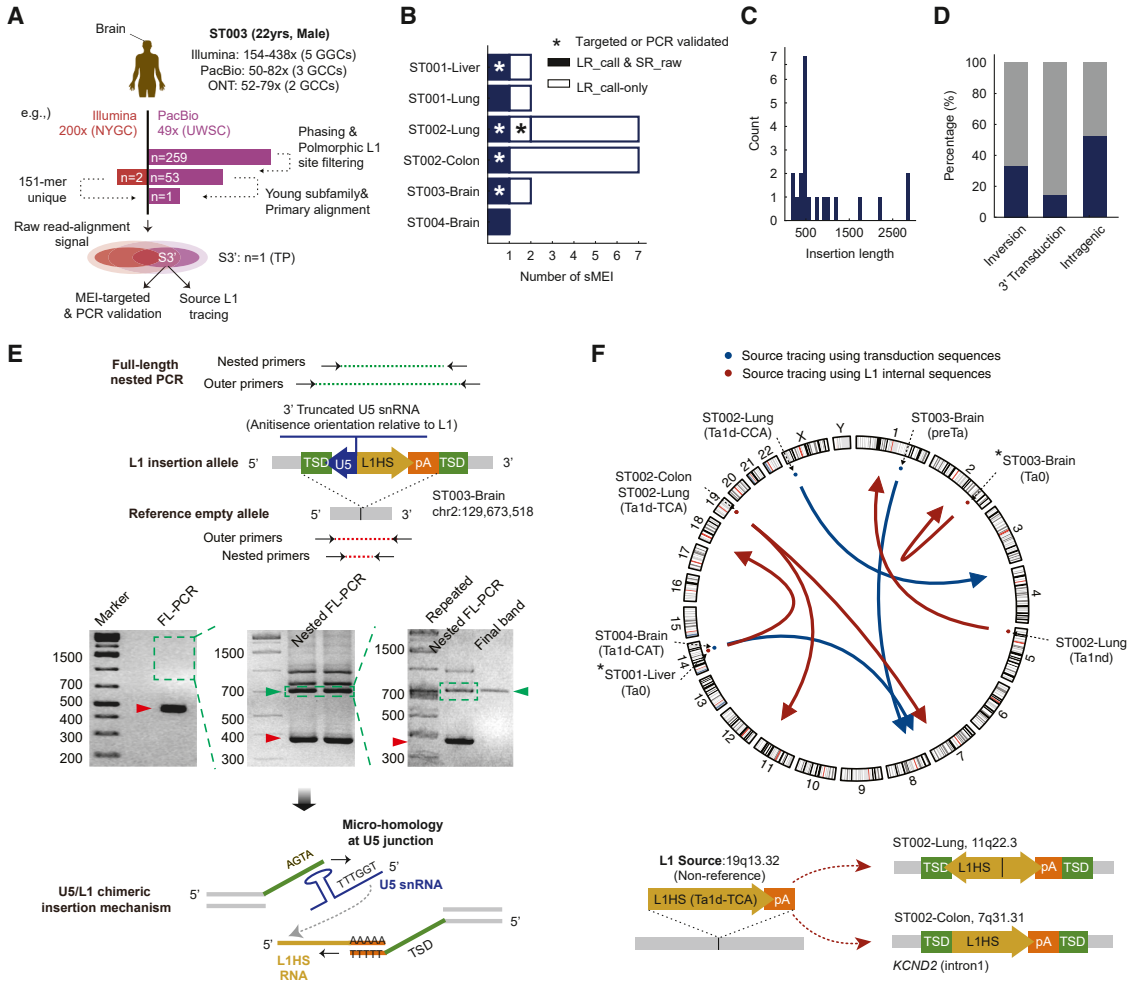


Figure 6. Characterization of sMEIs and source L1s in normal human tissue homogenates

(A) Overview of the somatic L1 detection pipeline applied to ST003 brain yielding one high-confidence insertion supported by long-read call and short-read alignment signal. This process was repeated for all samples from each of 5 GCCs and for other donors.

(B) The bar plot shows the number of somatic L1 insertions per sample.

(C) Length distribution of somatic L1 insertions across the tissue samples.

(D) Somatic L1 categorization by 5′ inversion, 3′ transduction, and intragenic vs. intergenic genomic regions.

(E) Schematic diagram of the U5/L1 chimeric insertion and representative agarose gels (bottom) for the insertion in ST003 brain sample. Due to low VAF, the L1 insertion band was not visible after the initial PCR (FL-PCR). A region of the gel at the expected size was excised (dashed green box) for subsequent DNA extraction and re-amplification. This repeated nested PCR approach successfully isolated a clean band corresponding to the L1 insertion allele (green arrowheads). Red arrowheads indicate the bands from the reference empty allele. The image provides a concise overview of the experimental validation. See Figure S13 for the additional experimental details.

(F) Source L1s identified by transduction and internal L1 sequences and schematic diagram of the L1 insertions in ST002 lung and colon samples from the same L1 source traced by internal L1 sequence variation.

impact may provide insights into their regulation and potential roles in altering the epigenetic and transcriptomic landscape in the human body.

DISCUSSION

sMEIs are increasingly recognized as contributors to genomic diversity in normal tissues, yet their reliable detection has remained challenging due to their low frequency, repetitive nature, and the lack of standardized benchmarks. To overcome these challenges, we established an integrative framework for sMEI analysis and demonstrated its applicability in normal human tissues. As a foundational component of this framework, we generated a high-confidence sMEI benchmarking set stratified by TPRT features using the HapMap cell-line mixtures. This benchmarking set provides a quantitative reference for method development and evaluation and will be valuable to the community for advancing computational and MEI-targeted sequencing approaches as interest in the roles of sMEIs in disease and normal aging continues to grow.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

Using this benchmarking set, we systematically evaluated the strengths and limitations of the nine sMEI detection methods across sequencing platforms, sequencing depth, genomic stratification, VAF, and insertion length. Building on this in-depth analysis, we derive practical recommendations for sMEI detection, particularly tailored to data availability and expected VAF ranges (Figure 7). When a single platform is available, and the expected VAF is greater than 3%, sequencing depths of 200× for short-read WGS and 60× for long-read WGS provide an optimal balance between cost and performance. In this regime, short-read performance reaches saturation at 200×, whereas long-read WGS at 60× achieves comparable performance. By contrast, when the expected VAF falls below 3%, long-read WGS becomes the preferred strategy, owing to its superior sensitivity at low VAFs and in challenging genomic regions, as well as its capacity to support downstream epigenomic analyses such as CpG DNA methylation profiling. In this low-VAF range, MEI-targeted sequencing represents a cost-effective alternative to both short-read and long-read WGS.

When both short-read and long-read WGS are available, we recommend integrating results at the raw read-alignment level with platform-specific FP filtering strategies. Identifying orthogonal raw-alignment signals at shared breakpoints enhanced the high-confidence sMEI detections supported by both platforms while maintaining high precision, compared with simply intersecting callsets. For short-read WGS, a substantial fraction of FPs can be removed by restricting calls to 151-mer uniquely mappable regions. For long-read WGS, somatic MEI calls can be refined by filtering FPs through haplotype phasing and by retaining only young subfamily calls based on complete insertion sequences.

We further recommend aligning sequence reads directly to DSAs when available. Using a tumor-normal mixture, we confirmed that DSA alignment improves precision by reducing FPs, particularly within SD regions. As with any assembly-enabled approach, the extent of improvement depends on assembly quality and varies by use case. For example, the levels of consensus accuracy, contiguity, and haplotype consistency needed for alignment-based sMEI refinement may differ from those required for long-range, genome-wide reconstruction tasks, such as resolution of megabase-scale copy-number changes or chromosome-scale phasing.

An assembly suitable for the intended application is shaped by several interacting factors, including sequencing depth, read-length distribution, sequencing modality, base accuracy and error profile, representation of repetitive or complex genomic regions, and DNA quality. In particular, sequencing strategies often involve a trade-off between depth and molecule length, which affects the traversability of complex regions and the achievable phase-block length. In this manuscript, we benchmark approaches for capturing sMEIs, including using DSA-based alignment as a support for haplotype-aware sMEI calling. Although it is difficult to define a strict minimum read-depth requirement for DSA construction, the current 60× long-read depth, together with long-range linkage data for phasing and scaffolding, produced DSAs that were suitable as mapping references for the analyses presented here. However, these coverage metrics were chosen specifically for the use case of this study; in some datasets with abundant ultra-long molecules and strong linkage information, lower total depth may still produce assemblies adequate for mapping-based analyses.

Leveraging long-read technology, we demonstrated that full-length L1 sequences harbor sufficient sequence variation to serve as reliable source markers and developed a new source-tracing pipeline based on the internal L1 sequence variation. Our pipeline identified three times more sources compared to the traditional transduction-based approach, uncovering previously unrecognized hot source elements and revealing allele-specific source activity at homozygous source loci. It is particularly effective in normal tissues, where transductions are expected to be less common.

Our integrative framework enabled reliable detection of low-frequency mosaic L1 insertions in normal human tissues, revealing structural and source diversity. We identified somatic L1 insertions in all six tissue samples examined, representing four tissue types, including lung, as well as a somatic U5 snRNA-L1 chimeric insertion in a brain sample. Furthermore, we observed multiple somatic insertions originating from the same source element across different tissues within an individual, providing insights into somatic retrotransposition dynamics. While our study primarily focused on high-depth WGS, several alternative strategies have recently emerged for sMEI detections, including approaches based on single-cell whole-genome amplification (WGA),[25] laser capture microdissection (LCM),[72] and single-cell-expanded clones.[24] These methods offer important complementary strengths, particularly by reducing cellular heterogeneity and enabling clonal or near single-cell resolution analyses. At the same time, each approach has inherent limitations. For instance, WGA-based methods must account for amplification artifacts and allelic dropout, whereas LCM-based strategies depend on the presence of clonally expanded cell populations within tissues. Furthermore, WGS approaches based on single-cell-derived clonal expansion require cell proliferation, limiting their applicability to post-mitotic cell types such as neurons and cardiomyocytes. Also, future studies that integrate single-cell and spatial sequencing as well as functional assays will be essential to determine how sMEIs influence cellular function. Analyzing larger and more diverse donor cohorts will clarify the prevalence, tissue distribution, and inter-individual variability of sMEIs and their active source loci. Additional questions also remain: whether the sMEI landscape in normal tissues mirrors the high variability observed across cancer types, whether specific genetic factors influence sMEI frequencies, and whether distinct sets of source loci or hotspot loci are active across different tissues during development and aging. Our integrative framework lays the foundation for answering these questions and advancing our understanding of the role of MEIs in human health.

Limitations of the study

Several limitations should be considered when interpreting the results in this study. First, the somatic insertions in HapMap mixture were inferred from germline insertions present in

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

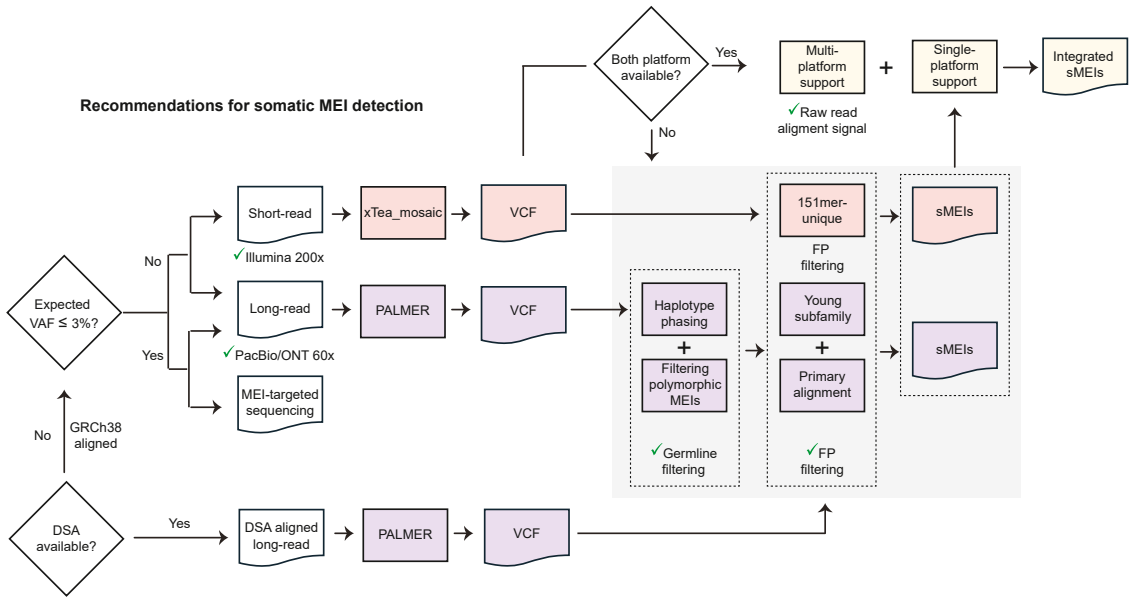


Figure 7. Recommendations for somatic MEI detection and source element tracing

Best-practice approaches for sMEI detection and source element tracing are summarized according to expected variant allele frequency (VAF) and data availability, including sequencing platforms and donor-specific assemblies (DSAs).

constituent cell lines. Although this design enables benchmarking across controlled allele fractions, these germline-derived insertions may not fully recapitulate the structural characteristics of somatic insertions. Somatic retrotransposition events can differ from germline insertions in features such as 5′ truncation or inversion, 3′ transduction, or breakpoint architectures, which could influence their detectability by different sequencing platforms and computational methods. Thus, the benchmark primarily evaluates the detection of low-frequency MEIs under controlled conditions and may not capture the full structural diversity of somatic retrotransposition *in vivo*. Additionally, our benchmarking set was generated using stringent criteria to ensure high confidence; therefore, it is conservative and potentially omitting some *bona fide* MEIs. Nevertheless, the detailed annotation of individual events in the benchmarking set enables users to perform stratified evaluations tailored to specific insertion characteristics or benchmarking purposes.

We also acknowledge practical limitations of the refinement and source-tracing approaches presented here. Generating high-quality DSAs remain technically demanding and resource intensive, which may not yet be uniformly accessible across research settings. The SMaHT network has been generating DSAs from fibroblast cultures for its production donors, and as these assemblies become available, the haplotype-aware DSA-based approaches presented here can be directly applied to further refine sMEI detection across the SMaHT production cohort. Lastly, although our source-tracing method substantially expands source identification, the accuracy of our approach depends heavily on the quality and completeness of the full-length germline L1HS insertion catalog for each sample. In addition, its performance with ONT WGS data has not yet been fully evaluated. Future work will aim to address these limitations and further refine the pipeline into a more comprehensive and robust framework for L1 source tracing.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Eunjung Alice Lee (ealice.lee@childrens.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study was conducted as part of the NIH Common Fund consortium initiative, SMaHT. The benchmark datasets (dbGaP: phs004193) described in this study are available through dbGaP (http://www.ncbi.nlm.nih.gov/gap). The data used in this work were provided by the SMaHT Data Analysis Center (DAC) on behalf of the SMaHT network. More information about the SMaHT network is available online at the SMaHT Data Portal (https://data.smaht.org/). The sequencing samples of H2009 and BL2009 can be found at CASTLE project (https://github.com/CASTLE-Panel/castle, SRA: PRJNA1086849). The open access data (callsets for HapMap mixture and tumor-cancer mixture, and haplotype-resolved assembly for BL2009) are available at https://doi.org/10.5281/zenodo.17254344. We used GRCh38 (GenBank: GCA_000001405.15) as the primary reference in this project. The reference can be downloaded at https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/. The scripts, command lines, and related pipeline links used in this project can be found at https://github.com/ealee-lab/SMaHT-MEI-Benchmarking and https://doi.org/10.5281/zenodo.23104249.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

ACKNOWLEDGMENTS

We are grateful to the SMaHT donors and donor families who have generously provided such precious gifts to support this important work. We thank Jayoung Ku for critical feedback on cancer benchmarking set generation. This research is supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under awards 1UM1DA058230, UM1DA058219, UG3NS132127, UG3NS132084, and UG3NS132146.

AUTHOR CONTRIBUTIONS

E.A.L., W.Z., R.E.M., and A.P.B. supervised the project. M.B., X. Zeng, and Z.W. generated the benchmarking set. S.W., C.E.S., B.M., S.D., Z.C., J.P., C.C., T.W., A.E.U., X. Zhu, and W.Z. generated callsets for WGS-based computational methods. B.Z., K.N., S.M., J.A.S., S.J.L., T.L.M., C.M., R.K.G., and W.Z. generated the datasets and callsets for MEI-targeted sequencing. S.W. performed benchmarking and integrative strategy. J.W., S.W., M.B., H.L., and W.Z. designed and performed haplotype phasing and DSA analysis. M.B. developed an internal sequence variation-based source-tracing pipeline. B.Z., K.N., and A.M.T. performed nested PCR validation. A.B., H.S., and K.H.B. prepared and managed samples for MEI-targeted sequencing and PCR validation. W.C.F. and H.-J.E.C. managed sequencing data. S.W., M.B., J.W., B.Z., R.E.M., W.Z., and E.A.L. drafted the manuscript. S.W., M.B., J.W., B.Z., A.P.B., R.E.M., W.Z., and E.A.L. revised the manuscript with input from all authors. All authors read and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Experimental design
- METHOD DETAILS
  - Library preparation and sequencing in MEI targeted sequencing
  - Donor-specific assembly generation
  - Benchmarking set generation
  - sMEI detection
  - Performance evaluation
  - Integrative strategy combined with call set and raw read alignment signals
  - Refinement of sMEIs using phasing information
  - Refinement of sMEIs using phasing information from DSA
  - DSA-based alignment analysis
  - Constructing the germline full-Length L1HS sequence catalog in haplotype resolution
  - Source identification based on transduction sequences
  - Source inference based on internal L1HS sequences
  - Discovery of sMEIs in human tissue homogenates
  - Experimental PCR validation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2026.101418.

Received: October 3, 2025
Revised: December 22, 2025
Accepted: October 12, 2026

REFERENCES

1. Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell *87*, 905–916. https://doi.org/10.1016/s0092-8674(00)81997-2.
2. Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. EMBO J. *21*, 5899–5910. https://doi.org/10.1093/emboj/cdf592.
3. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703. https://doi.org/10.1038/nrg2640.
4. Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327–338. https://doi.org/10.1016/s0092-8674(02)00839-5.
5. Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. Cell *110*, 315–325. https://doi.org/10.1016/s0092-8674(02)00828-0.
6. Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al. (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. *52*, 306–319. https://doi.org/10.1038/s41588-019-0562-0.
7. Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. *52*, 643–645.
8. Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. *26*, 745–755. https://doi.org/10.1101/gr.201814.115.
9. Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285. https://doi.org/10.1038/nrg2072.
10. De Cecco, M., Ito, T., Petrashen, A.P., Elias, A.E., Skvir, N.J., Criscione, S.W., Caligiana, A., Brocculi, G., Adney, E.M., Boeke, J.D., et al. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature *566*, 73–78. https://doi.org/10.1038/s41586-018-0784-9.
11. Crow, M.K. (2010). Long interspersed nuclear elements (LINE-1): potential triggers of systemic autoimmune disease. Autoimmunity *43*, 7–16. https://doi.org/10.3109/08916930903374865.
12. Gorbunova, V., Seluanov, A., Mita, P., McKerrow, W., Fenyö, D., Boeke, J.D., Linker, S.B., Gage, F.H., Kreiling, J.A., Petrashen, A.P., et al. (2021). The role of retrotransposable elements in ageing and age-associated diseases. Nature *596*, 43–53. https://doi.org/10.1038/s41586-021-03542-y.
13. Sanchez-Luque, F.J., Kempen, M.-J.H.C., Gerdes, P., Vargas-Landin, D.B., Richardson, S.R., Troskie, R.-L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., Salvador-Palomeque, C., et al. (2019). LINE-1 Evasion of Epigenetic Repression in Humans. Mol. Cell *75*, 590–604.e12. https://doi.org/10.1016/j.molcel.2019.05.024.
14. Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., et al. (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967–971. https://doi.org/10.1126/science.1222077.
15. Solovyov, A., Behr, J.M., Hoyos, D., Banks, E., Drong, A.W., Thornlow, B., Zhong, J.Z., Garcia-Rivera, E., McKerrow, W., Chu, C., et al. (2025). Pan-cancer multi-omic model of LINE-1 activity reveals locus heterogeneity of retrotransposition efficiency. Nat. Commun. *16*, 2049. https://doi.org/10.1038/s41467-025-57271-1.
16. Shukla, R., Upton, K.R., Muñoz-Lopez, M., Gerhardt, D.J., Fisher, M.E., Nguyen, T., Brennan, P.M., Baillie, J.K., Collino, A., Ghisletti, S., et al. (2013). Endogenous retrotransposition activates oncogenic pathways

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

in hepatocellular carcinoma. Cell *153*, 101–111. https://doi.org/10.1016/j.cell.2013.02.032.

17. Burns, K.H. (2017). Transposable elements in cancer. Nat. Rev. Cancer *17*, 415–424. https://doi.org/10.1038/nrc.2017.35.
18. Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., et al. (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. Neuron *81*, 306–313. https://doi.org/10.1016/j.neuron.2013.10.053.
19. Zhao, B., Wu, Q., Ye, A.Y., Guo, J., Zheng, X., Yang, X., Yan, L., Liu, Q.-R., Hyde, T.M., Wei, L., and Huang, A.Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. PLoS Genet. *15*, e1008043. https://doi.org/10.1371/journal.pgen.1008043.
20. Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C.N., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc. Natl. Acad. Sci. USA *108*, 20382–20387. https://doi.org/10.1073/pnas.1100273108.
21. Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., and Walsh, C.A. (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49–59. https://doi.org/10.1016/j.neuron.2014.12.028.
22. Erwin, J.A., Paquola, A.C.M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I.A., Butcher, C.R., Herdy, J.R., et al. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. *19*, 1583–1591. https://doi.org/10.1038/nn.4388.
23. Zhu, X., Zhou, B., Pattni, R., Gleason, K., Tan, C., Kalinowski, A., Sloan, S., Fiston-Lavier, A.-S., Mariani, J., Petrov, D., et al. (2021). Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. Nat. Neurosci. *24*, 186–196. https://doi.org/10.1038/s41593-020-00767-4.
24. Nam, C.H., Youk, J., Kim, J.Y., Lim, J., Park, J.W., Oh, S.A., Lee, H.J., Park, J.W., Won, H., Lee, Y., et al. (2023). Widespread somatic L1 retrotransposition in normal colorectal epithelium. Nature *617*, 540–547. https://doi.org/10.1038/s41586-023-06046-z.
25. Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., et al. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483–496. https://doi.org/10.1016/j.cell.2012.09.035.
26. Erwin, J.A., Marchetto, M.C., and Gage, F.H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. Nat. Rev. Neurosci. *15*, 497–506. https://doi.org/10.1038/nrn3730.
27. Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nat. Rev. Genet. *24*, 464–483. https://doi.org/10.1038/s41576-023-00590-0.
28. Prodanov, T., and Bansal, V. (2020). Sensitive alignment using paralogous sequence variants improves long-read mapping and variant calling in segmental duplications. Nucleic Acids Res. *48*, e114. https://doi.org/10.1093/nar/gkaa829.
29. Zhou, W., Emery, S.B., Flasch, D.A., Wang, Y., Kwan, K.Y., Kidd, J.M., Moran, J.V., and Mills, R.E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. *48*, 1146–1163. https://doi.org/10.1093/nar/gkz1173.
30. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117. https://doi.org/10.1126/science.abf7117.
31. Chu, C., Borges-Monroy, R., Viswanadham, V.V., Lee, S., Li, H., Lee, E.A., and Park, P.J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat. Commun. *12*, 3836. https://doi.org/10.1038/s41467-021-24041-8.
32. Chu, C., Zhao, B., Park, P.J., and Lee, E.A. (2020). Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. Curr. Protoc. Hum. Genet. *107*, e102. https://doi.org/10.1002/cphg.102.
33. Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K., et al. (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343. https://doi.org/10.1126/science.1251343.
34. Faulkner, G.J., and Garcia-Perez, J.L. (2017). L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet. *33*, 802–816. https://doi.org/10.1016/j.tig.2017.07.004.
35. Faulkner, G.J., and Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. Mob. DNA *9*, 22. https://doi.org/10.1186/s13100-018-0128-1.
36. Chen, J., Cheng, J., Chen, X., Inoue, M., Liu, Y., and Song, C.-X. (2022). Whole-genome long-read TAPS deciphers DNA methylation patterns at base resolution using PacBio SMRT sequencing technology. Nucleic Acids Res. *50*, e104. https://doi.org/10.1093/nar/gkac612.
37. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods *14*, 407–410. https://doi.org/10.1038/nmeth.4184.
38. Chen, X., Baker, D., Dolzhenko, E., Devaney, J.M., Noya, J., Berlyoung, A.S., Brandon, R., Hruska, K.S., Lochovsky, L., Kruszka, P., et al. (2025). Genome-wide profiling of highly similar paralogous genes using HiFi sequencing. Nat. Commun. *16*, 2340. https://doi.org/10.1038/s41467-025-57505-2.
39. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.
40. Logsdon, G.A., Ebert, P., Audano, P.A., Loftus, M., Porubsky, D., Ebler, J., Yilmaz, F., Hallast, P., Prodanov, T., Yoo, D., et al. (2025). Complex genetic variation in nearly complete human genomes. Nature *644*, 430–441. https://doi.org/10.1038/s41586-025-09140-6.
41. Ewing, A.D., Smits, N., Sanchez-Luque, F.J., Faivre, J., Brennan, P.M., Richardson, S.R., Cheetham, S.W., and Faulkner, G.J. (2020). Nanopore sequencing enables comprehensive transposable element epigenomic profiling. Mol. Cell *80*, 915–928.e5. https://doi.org/10.1016/j.molcel.2020.10.024.
42. Xiao, C., Chen, Z., Chen, W., Padilla, C., Colgan, M., Wu, W., Fang, L.-T., Liu, T., Yang, Y., Schneider, V., et al. (2022). Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. Genome Biol. *23*, 237. https://doi.org/10.1186/s13059-022-02803-x.
43. Zhou, W., Mumm, C., Gan, Y., Switzenberg, J.A., Wang, J., De Oliveira, P., Kathuria, K., Losh, S.J., Bessell, B., McDonald, T.L., et al. (2025). A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex. Preprint at bioRxiv. https://doi.org/10.1101/2024.12.18.629274.
44. Zumalave, S., Santamarina, M., Espasandín, N.P., Zamora, J., Garcia-Souto, D., Temes, J., Baker, T.M., Rodríguez-Castro, J., Otero, P., Pequeño-Valtierra, A., et al. (2026). Concurrent L1 retrotransposition events promote reciprocal translocations in human tumorigenesis. Science *392*, eaee4513. https://doi.org/10.1126/science.aee4513.
45. Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534–537. https://doi.org/10.1038/nature10531.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

46. McKerrow, W., Tang, Z., Steranka, J.P., Payer, L.M., Boeke, J.D., Keefe, D., Fenyö, D., Burns, K.H., and Liu, C. (2020). Human transposon insertion profiling by sequencing (TIPseq) to map LINE-1 insertions in single cells. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190335. https://doi.org/10.1098/rstb.2019.0335.

47. McDonald, T.L., Zhou, W., Castro, C.P., Mumm, C., Switzenberg, J.A., Mills, R.E., and Boyle, A.P. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nat. Commun. *12*, 3586. https://doi.org/10.1038/s41467-021-23918-y.

48. Coorens, T.H.H., Oh, J.W., Choi, Y.A., Lim, N.S., Zhao, B., Voshall, A., Abyzov, A., Antonacci-Fulton, L., Aparicio, S., Ardlie, K.G., et al. (2025). The Somatic Mosaicism across Human Tissues Network. Nature *643*, 47–59. https://doi.org/10.1038/s41586-025-09096-7.

49. (2026). Somatic Mosaicism across Human Tissues Network. Comprehensive benchmarking of somatic mutation detection by the SMaHT Network. Cell *189* (*21*), 6594–6612. https://doi.org/10.1016/j.cell.2026.09.037.

50. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.

51. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat. Biotechnol. *42*, 1571–1580. https://doi.org/10.1038/s41587-023-02024-y.

52. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker Open-4.0. https://www.repeatmasker.org.

53. (2025). Li H. Finding easy regions for short-read variant calling from pangenome data. Gigascience *14*, giaf103. https://doi.org/10.1093/gigascience/giaf103.

54. Keskus, A.G., Bryant, A., Ahmad, T., Yoo, B., Aganezov, S., Goretsky, A., Donmez, A., Lansdon, L.A., Rodriguez, I., Park, J., et al. (2026). Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. Nat. Biotechnol. *44*, 247–257. https://doi.org/10.1038/s41587-025-02618-8.

55. Park, J., Cook, D.E., Chang, P.-C., Kolesnikov, A., Brambrink, L., Mier, J.C., Gardner, J., McNulty, B., Sacco, S., Keskus, A.G., et al. (2026). Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. Nat. Biotechnol. *44*, 1569–1578. https://doi.org/10.1038/s41587-025-02839-x.

56. Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA *100*, 5280–5285. https://doi.org/10.1073/pnas.0831042100.

57. Schloissnig, S., Pani, S., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., Asparuhova, M., et al. (2025). Structural variation in 1,019 diverse humans based on long-read sequencing. Nature *644*, 442–452. https://doi.org/10.1038/s41586-025-09290-7.

58. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and 1000 Genomes Project Consortium; and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929. https://doi.org/10.1101/gr.218032.116.

59. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539. https://doi.org/10.1038/msb.2011.75.

60. Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. *11*, 2059–2065. https://doi.org/10.1101/gr.205701.

61. Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res. *31*, 4385–4390. https://doi.org/10.1093/nar/gkg496.

62. Garcia-Perez, J.L., Doucet, A.J., Bucheton, A., Moran, J.V., and Gilbert, N. (2007). Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. Genome Res. *17*, 602–611. https://doi.org/10.1101/gr.5870107.

63. Doucet, A.J., Droc, G., Siol, O., Audoux, J., and Gilbert, N. (2015). U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. Mol. Biol. Evol. *32*, 1815–1832. https://doi.org/10.1093/molbev/msv062.

64. Law, C.-T., and Burns, K.H. (2025). Comparative Genomics Reveals LINE-1 Recombination with Diverse RNAs.. Preprint at bioRxiv. https://doi.org/10.1101/2025.02.02.635956.

65. Chuang, N.T., Gardner, E.J., Terry, D.M., Crabtree, J., Mahurkar, A.A., Rivell, G.L., Hong, C.C., Perry, J.A., and Devine, S.E. (2021). Mutagenesis of human genomes by endogenous mobile elements on a population scale. Genome Res. *31*, 2225–2235. https://doi.org/10.1101/gr.275323.121.

66. Aricescu, A.R., Siebold, C., Choudhuri, K., Chang, V.T., Lu, W., Davis, S.J., van der Merwe, P.A., and Jones, E.Y. (2007). Structure of a tyrosine phosphatase adhesive interaction reveals a spacer-clamp mechanism. Science *317*, 1217–1220. https://doi.org/10.1126/science.1144646.

67. Liu, W., Liu, Y., Zhang, L., Li, L., Yang, W., Li, J., and He, W. (2025). Nucleic acid spheres for treating capillarisation of liver sinusoidal endothelial cells in liver fibrosis. Nat. Commun. *16*, 4517. https://doi.org/10.1038/s41467-025-59885-x.

68. Luo, Z., Mao, X., and Cui, W. (2019). Circular RNA expression and circPTPRM promotes proliferation and migration in hepatocellular carcinoma. Med. Oncol. *36*, 86. https://doi.org/10.1007/s12032-019-1311-z.

69. Koiliari, E., Roussos, P., Pasparakis, E., Lencz, T., Malhotra, A., Siever, L.J., Giakoumaki, S.G., and Bitsios, P. (2014). The CSMD1 genome-wide associated schizophrenia risk variant rs10503253 affects general cognitive ability and executive function in healthy males. Schizophr. Res. *154*, 42–47. https://doi.org/10.1016/j.schres.2014.02.017.

70. Byrne, R.A.J., Nimmo, J., Torvell, M., Carpanini, S.M., Daskoulidou, N., Hughes, T.R., Noble, L.V., Veteleanu, A., Watkins, L.M., Zelek, W.M., et al. (2025). The schizophrenia-associated gene CSMD1 encodes a complement classical pathway inhibitor predominantly expressed by astrocytes and at synapses in mice and humans. Brain Behav. Immun. *127*, 287–302. https://doi.org/10.1016/j.bbi.2025.03.026.

71. Athanasiu, L., Giddaluru, S., Fernandes, C., Christoforou, A., Reinvang, I., Lundervold, A.J., Nilsson, L.-G., Kauppi, K., Adolfsson, R., Eriksson, E., et al. (2017). A genetic association study of CSMD1 and CSMD2 with cognitive function. Brain Behav. Immun. *61*, 209–216. https://doi.org/10.1016/j.bbi.2016.11.026.

72. Olafsson, S., McIntyre, R.E., Coorens, T., Butler, T., Jung, H., Robinson, P.S., Lee-Six, H., Sanders, M.A., Arestang, K., Dawson, C., et al. (2020). Somatic Evolution in Non-neoplastic IBD-Affected Colon. Cell *182*, 672–684.e11. https://doi.org/10.1016/j.cell.2020.06.036.

73. Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., and MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat. Methods *15*, 595–597. https://doi.org/10.1038/s41592-018-0054-7.

74. Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome Biol. *21*, 265. https://doi.org/10.1186/s13059-020-02168-z.

75. Shiraishi, Y., Koya, J., Chiba, K., Okada, A., Arai, Y., Saito, Y., Shibata, T., and Kataoka, K. (2023). Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. Nucleic Acids Res. *51*, e74. https://doi.org/10.1093/nar/gkad526.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

76. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

77. Fernandes, J.D., Zamudio-Hurtado, A., Clawson, H., Kent, W.J., Haussler, D., Salama, S.R., and Haeussler, M. (2020). The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. Mob. DNA *11*, 13. https://doi.org/10.1186/s13100-020-00208-w.

78. Antipov, D., Rautiainen, M., Nurk, S., Walenz, B.P., Solar, S.J., Phillippy, A.M., and Koren, S. (2025). Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. Genome Res. *35*, 1583–1594. https://doi.org/10.1101/gr.280383.124.

79. Astashyn, A., Tvedte, E.S., Sweeney, D., Sapojnikov, V., Bouk, N., Joukov, V., Mozes, E., Strope, P.K., Sylla, P.M., Wagner, L., et al. (2024). Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biol. *25*, 60. https://doi.org/10.1186/s13059-024-03198-7.

80. Stanojević, D., Lin, D., Nurk, S., de Sessions, P.F., and Šikić, M. (2026). Telomere-to-telomere assembly using HERRO-corrected Simplex nanopore reads. Nature, 1–3. https://doi.org/10.1038/s41586-026-10563-y.

81. Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J.M., Miga, K.H., et al. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat. Methods *18*, 1322–1332. https://doi.org/10.1038/s41592-021-01299-w.

82. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. *23*, 258. https://doi.org/10.1186/s13059-022-02823-7.

83. Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., and Zhang, Y. (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience *9*, giaa094. https://doi.org/10.1093/gigascience/giaa094.

84. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

85. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics *34*, i142–i150. https://doi.org/10.1093/bioinformatics/bty266.

86. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. *21*, 245. https://doi.org/10.1186/s13059-020-02134-9.

87. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

88. Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

89. Van Der Auwera, G.O., and Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media).

90. Daniels, C.A., Abdulkadir, A.A., Cleveland, M.H., McDaniel, J.H., Jáspez, D., Rubio-Rodríguez, L.A., Muñoz-Barrera, A., Lorenzo-Salazar, J.M., Flores, C., Yoo, B., et al. (2026). Characterization of subclonal variants in HG002 Genome in a Bottle reference material as a resource for benchmarking variant callers. Cell Genom. *6*, 101104. https://doi.org/10.1016/j.xgen.2025.101104.

91. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., and Smit, A.F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob. DNA *12*, 2. https://doi.org/10.1186/s13100-020-00230-y.

92. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11. https://doi.org/10.1186/s13100-015-0041-9.

93. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

94. University of Michigan, Zhou, W., Gan, Y., Boyle, A., and Mills, R. (2025). A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex. Zenodo. https://zenodo.org/records/15733390.

95. R Core Team (2024). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

96. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

97. Niu, Y., Teng, X., Zhou, H., Shi, Y., Li, Y., Tang, Y., Zhang, P., Luo, H., Kang, Q., Xu, T., and He, S. (2022). Characterizing mobile element insertions in 5675 genomes. Nucleic Acids Res. *50*, 2493–2508. https://doi.org/10.1093/nar/gkac128.

98. Kojima, S., Koyama, S., Ka, M., Saito, Y., Parrish, E.H., Endo, M., Takata, S., Mizukoshi, M., Hikino, K., Takeda, A., et al. (2023). Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. Nat. Genet. *55*, 939–951. https://doi.org/10.1038/s41588-023-01390-2.

99. Borges-Monroy, R., Chu, C., Dias, C., Choi, J., Lee, S., Gao, Y., Shin, T., Park, P.J., Walsh, C.A., and Lee, E.A. (2021). Whole-genome analysis reveals the contribution of non-coding de novo transposon insertions to autism spectrum disorder. Mob. DNA *12*, 28. https://doi.org/10.1186/s13100-021-00256-w.

100. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Cell lines		
HG005 lymphoblastoid cell line	Coriell Institute	HG005 Cat# GM24631, RRID:CVCL_1C97
HG02622 lymphoblastoid cell line	Coriell Institute	HG02622 Cat# HG02622, RRID:CVCL_GV49
HG002 lymphoblastoid cell line	Coriell Institute	HG002 Cat# GM24385, RRID:CVCL_1C78
HG02257 lymphoblastoid cell line	Coriell Institute	HG02257 Cat# HG02257, RRID:CVCL_Q703
HG02486 lymphoblastoid cell line	Coriell Institute	HG02486 Cat# HG02486
HG00438 lymphoblastoid cell line	Coriell Institute	HG00438 Cat# HG00438, RRID:CVCL_P889
HapMap lymphoblastoid cell line mixture	WashU-VAI	HapMap6
Biological samples		
Liver from donor ST001	TPC/NDRI	ST001-1A
Lung from donor ST001	TPC/NDRI	ST001-1D
Colon from donor ST002	TPC/NDRI	ST002-1G
Lung from donor ST002	TPC/NDRI	ST002-1D
Brain from donor ST003	TPC/NDRI	ST003-1Q
Brain from donor ST004	TPC/NDRI	ST004-1Q
Deposited data		
Data for HG005, HG02622, HG02486, HG02257, HG002, HG00438	Human Pangenome Reference Consortium	humanpangenome.org
Data for HapMap6	dbGaP	phs004193
Data for HapMap6	SMaHT DAC	https://data.smaht.org
Data for ST001, ST002, ST003, ST004	dbGaP	phs004193
Data for BL2009, H2009	SRA	PRJNA1086849
Software and algorithms		
PALMER (v2.0.1)	Zhou et al.[29]	https://github.com/WeichenZhou/PALMER
xTea_long (v0.1.0)	Chu et al.[31]	https://github.com/parklab/xTea
cuteSV (v2.1.2)	Jiang et al.[50]	https://github.com/tjiangHIT/cuteSV
Sniffles2 (v2.2, v.2.6.3)	Smolka et al.[51]	https://github.com/fritzsedlazeck/Sniffles
xTea_mosaic (v0.1.9)	Chu et al.[31]	https://github.com/parklab/xTea
MELT (v2.2.2)	Gardner et al.[58]	https://melt.igs.umaryland.edu/index.php
RetroSom (v2)	Zhu et al.[23]	https://github.com/Czhuofu/RetroSomV2
eHAT-seq pipeline (v0.1.0)	Zhao et al.[19]	https://github.com/ealee-lab/HATseq-pipeline
TEnCATS pipeline	McDonald et al.[47]	https://doi.org/10.17504/protocols.io.kqdg3q66ev25/v2
Minimera	Losh et al.	https://github.com/Boyle-Lab/minimera/
Dipcall (v0.2)	Heng et al.[73]	https://github.com/lh3/dipcall
Minigraph (v0.34)	Heng et al.[74]	https://github.com/lh3/minigraph
Nanomonsv (v0.7.0)	Shiraishi et al.[75]	https://github.com/friend1ws/nanomonsv
Severus (v1.2)	Keskus et al.[54]	https://github.com/KolmogorovLab/Severus
BLAST	Camacho et al.[76]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
RepeatMasker	Smit et al.[52]	http://www.repeatmasker.org
RepeatBrowser	Fernandes et al.[77]	https://repeatbrowser.ucsc.edu/
pbmm2	PacBio Inc.	https://github.com/PacificBiosciences/pbmm2
Verkko (v2.2.1)	Rautiainen et al.[78]	https://github.com/marbl/verkko
FCS (v0.5.5)	Astashyn et al.[79]	https://github.com/ncbi/fcs
HERRO (v0.1.1)	Stanojević et al.[80]	https://github.com/lbcb-sci/herro

(*Continued on next page*)

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PEPPER-Margin-DeepVariant (r0.8)	Shafin et al.[81]	https://github.com/kishwarshafin/pepper
RagTag (v2.1.0)	Alonge et al.[82]	https://github.com/malonge/RagTag
TGS-Gapcloser (v1.2.1)	Xu et al.[83]	https://github.com/BGI-Qingdao/TGS-GapCloser
BUSCO (v.5.7.1)	Manni et al.[84]	https://github.com/metashot/busco
QUAST (v5.2.0)	Mikheenko et al.[85]	https://github.com/ablab/quast
Merqury (v1.3)	Rhie et al.[86]	https://github.com/marbl/merqury
Scipy (v1.10.1)	Virtanen et al.[87]	https://github.com/scipy/scipy
BLAT	Kent et al.[88]	https://genome.ucsc.edu/cgi-bin/hgBlat
GATK	Auwera & O'Connor[89]	https://gatk.broadinstitute.org/
Code for performance evaluation and multi-platform integration (v1.0.0)	This paper	https://doi.org/10.5281/zenodo.23104249 (https://github.com/ealee-lab/SMaHT-MEI-Benchmarking)
Others		
Sequence data for in-silico BL2009-H2009 mixture	This paper	https://doi.org/10.5281/zenodo.17254344
Haplotype-resolved assembly for BL2009	This paper	https://doi.org/10.5281/zenodo.17254344

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Experimental design

HapMap mixture

Briefly, the SMaHT network designed HapMap mixtures by combining six cell lines from International HapMap Projects at predefined ratios to simulate somatic variants: HG005 (83.5%; Male; Chinese), HG02622 (10%; Female; Gambian in Western Division, Mandinka), HG02486 (2%; Male; African Caribbean in Barbados), HG02257 (2%; Female; African Caribbean in Barbados), HG002 (2%; Male; Ashkenazim Jewish), and HG00438 (0.5%; Female; Chinese Han in the South). These six cell lines were selected because they have been extensively sequenced and comprehensively characterized, with high-quality genome assemblies available through the Genome in a Bottle (GIAB)[90] and Human Pangenome projects.[39] Together, they represent diverse ancestral and sex backgrounds, thereby increasing genetic diversity and improving the robustness of benchmarking analyses. Each of the five SMaHT GCCs generated high-coverage WGS data for the mixtures, including short-read (Illumina) and long-read (PacBio and ONT) datasets. We obtained the aligned sequencing data (BAM) of the mixtures from the SMaHT data portal (https://data.smaht.org/) for benchmarking WGS-based computational sMEI detection methods.

In silico *BL2009-H2009 mixtures*

PacBio-HIFI sequencing data for BL2009 (normal) and H2009 (lung adenocarcinoma) cell lines were obtained from the Sequence Read Archive (SRA, PRJNA1086849) and the CASTLE project (https://github.com/CASTLE-Panel/castle).[54,55] We generated four *in silico* BL2009-H2009 mixtures at 60x coverage with tumor read spike-in ratios of 2%, 10%, 20%, and 40%, respectively. Ideally, the heterozygous sMEIs in the H2009 tumor cell line will be downsampled to the predefined VAFs (1%, 5%, 10%, and 20%). The mixed reads were aligned to the GRCh38 human reference genome using pbmm2 (https://github.com/PacificBiosciences/pbmm2).

Human tissue homogenate

The SMaHT network collected six tissue samples from four *postmortem* donors: ST001 (liver and lung; Male; 22 years old), ST002 (lung and colon; Male; 74 years old; male), ST003 (brain; Male; 22 years old), and ST004 (brain; Male; 73 years old). To ensure cellular homogeneity, the tissue samples were homogenized prior to sequencing. WGS was performed by five SMaHT GCCs using Illumina, PacBio, and ONT platforms. Aligned sequencing data (BAM) of each sample were obtained from the SMaHT data portal (https://data.smaht.org/).

Detailed sample preparation and processing protocols for the samples that we obtained from the SMaHT network are described in the SMaHT flagship publication,[49] including "Preparing HapMap Samples" and "Human Tissue Preparation and Distribution" for sample preparation, and "Sequencing at BCM/Broad/NYGC/UW/WashU-VAI" for sequencing procedures. For the in silico tumor–normal mixtures, sample preparation and sequencing procedures are described in the original CASTLE projects.[54,55] Data summary of WGS samples are provided in Table S5.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

METHOD DETAILS

Library preparation and sequencing in MEI targeted sequencing

TEnCATS

TEnCATS library preparation was performed following McDonald et al.[47] with changes described here for SQK-LSK114 and R10.4.1 flow cells (data and code availability). 30 μL of gDNA (1900ng/μL) from the SMaHT HapMap mixture or 10 μg of gDNA from the benchmarking donor tissue samples were dephosphorylated in a 40μL reaction with 6μL Quick CIP (M0525S, NEB) and 4μL 10X rCutSmart buffer (B7204S, NEB). This reaction was inverted and gently tapped to mix, and then incubated at 37°C for 30 min followed by a 2 min heat inactivation at 80°C. The Cas9 ribonucleoprotein (RNP) was formed by combining 850ng of *in vitro* transcribed guide RNA, 1 μL of a 1:5 dilution of Alt-R S.p. HiFi Cas9 Nuclease V3 (1081060, IDT), and 1X rCutSmart buffer (B7204S, NEB) in a total of 30 μL. This reaction was incubated at RT for 20 min. Next, both the prepped gDNA and RNP were placed on ice and the RNP was added to the dephosphorylated gDNA. 1μL 10mM dATP and 1.5μL Taq DNA Polymerase (M0273S, NEB) were added to the gDNA:RNP reaction, then inverted and gently tapped to mix. This reaction was incubated at 37°C for 30 min for Cas9 cutting and brought to 75°C for a-tailing for 10 min. For adapter ligation, the cut reaction was transferred to a 1.5mL microcentrifuge tube. We then added 5μL T4 DNA ligase (M0202M, NEB) and 5μL Ligation Adapter (LA; SQK-LSK114, ONT). This reaction was inverted to mix and incubated at RT for 20 min with rotation. Following ligation, we added 1 volume of 1X Tris EDTA (TE) and inverted to mix. Next 0.4X Ampure beads (SQK-LSK114, ONT) are added and incubated for 5 min with rotation followed by 5 min at RT without rotation. The beads were then washed twice with 150μL Long Fragment Buffer (LFB; SQK-LSK114, ONT) followed by incubation with 50μL Elution Buffer (EB; SQK-LSK114, ONT) at 37°C for 30 min. Finally, we loaded the R10.4.1 MinION flow cell following the ONT protocol using 10μL of the library and sequenced for 72 h on a MinION. For sequencing of the libraries on PromethION flow cells, 12μL of the same library sequenced on a MinION was used and brought to a total volume of 32μL with EB then sequenced for 72 h on a PromethION2 Solo.

eHAT-seq

We prepared enhanced Human Active Transposon-Sequencing (eHAT-seq) libraries by adapting the protocol from previous work.[19] Starting with 500 ng of genomic DNA extracted from the SMaHT HapMap cell line mixture and SMaHT tissue samples, we first fragmented the DNA using a Covaris R230 focused-ultrasonicator (Covaris, LLC). These fragments were subsequently processed for end repair, A-tailing, and adapter ligation with the KAPA HyperPrep Kit (Kapa Biosystems, Inc). Using the adapter-ligated DNA as a template, we generated four technical replicate libraries for the HapMap mixture and two replicates for each of SMaHT tissue samples. The L1 enrichment and indexing PCR steps included two key modifications to the previous protocol: 1) the P7_Ns_L1Hs primer contained a mix of 2,4,6,8 staggers to enhance the sequence diversity of the semi-amplicon library. 2) unique dual indices were used for both P5 and P7 adapters to ensure precise library demultiplexing. Library fragments ranging from 340 to 450 bp were enriched using a dual size selection with AMPure beads from 0.575x to 0.75x beads ratio (Beckman Coulter). We quantified libraries with the KAPA Library Quantification Kit (KAPA Biosystems) and then combined them into an equimolar pool. The final library pool was sequenced with a 5% PhiX spike-in on an Illumina NovaSeq sequencing system, generating 2 × 150 bp paired-end reads. A list of adapter and primer sequences used for eHAT-seq library construction is available in Table S21.

Donor-specific assembly generation

For cancer cell line mixture analysis, we built donor-specific assemblies (DSA) using PacBio-HIFI, Nanopore ONT, UL-ONT, and Pore-C sequencing data from the normal cell line (BL2009) in the CASTLE panel project (SRA PRJNA1086849).[54,55] These data were collected to generate a DSA without downsampling for this normal-tissue pair using Verkko[78] v2.2.1, the resulting DSA then went through a decontamination process through FCS[79] v0.5.5 (Figure S7).

For benchmarking donors, due to the lack of phasing information such as Hi-C or trio data, we defined the *target* as the draft donor-specific assembly and introduced the concept of *bridges*—additional sequencing datasets that share overlap with the target to extend phase blocks within the DSA (Figure S7). The initial draft DSA was assembled from PacBio HiFi, Oxford Nanopore (ONT), and HERRO-corrected ONT reads. HiFi reads were randomly downsampled to 100× coverage, and ONT reads were corrected with HERRO[80] using the *model_R10_v0.1.pt* model, retaining only reads longer than 10 kb. The draft DSA was partitioned into two haplotypes using a custom graph-based approach: contig overlaps were used to construct a weighted graph, and a greedy bi-partitioning algorithm assigned contigs to haplotypes so as to maximize phase block continuity within each group. Phased heterozygous SNVs were then identified using dipcall[73] v0.3 on the two resulting haplotypes. For bridges, original ONT reads were used (rather than HERRO-corrected reads), as HERRO correction discards base quality scores. Phased heterozygous SNVs were called using PEPPER-Margin-DeepVariant[81] r0.8.

Draft DSA contigs (“targets”) were connected via ONT reads (“bridges”) using overlapping phased heterozygous SNPs as anchors. For each merged interval, phased genotypes at each shared SNP between target and bridge blocks were compared, and the most common phase switch decision (match or flip) was determined for each target phase block. A concordance matrix was constructed, and target blocks connected through bridges with valid phased SNPs overlap were grouped using a disjoint-set union (DSU) algorithm. Within each group, genomic coordinates were merged to define extended phase blocks. The flip status for each phase block was propagated from an anchor block, ensuring consistent haplotype assignment across the group. Whenever a flip was detected, contigs previously assigned to haplotype 1 were switched to haplotype 2 and vice versa, maintaining phase consistency until another flip was encountered. Final phase switch corrected phase blocks were exported in BED format for downstream

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

analysis. Phase switch-corrected contigs were further scaffolded using RagTag[82] v2.1.0 and gap-filled with ONT reads using TGS-Gapcloser[83] v1.2.1. For each haplotype, only reads assigned to the corresponding haplotype were used in the scaffolding and gap-filling steps. The quality assessment of all diploid assemblies was performed thoroughly by three tools: BUSCO[84] v5.7.1, QUAST[85] v5.2.0, and Merqury[86] v1.3 (Table S22).

Benchmarking set generation

HapMap mixtures

To simulate somatic insertions in HapMap mixtures, we first identified high-confidence non-reference MEIs in six HapMap cell lines (Figure S1A). We obtained publicly available PacBio long-read whole-genome sequencing data (BAM files) and haplotype-resolved assemblies for these cell lines from the Human Pangenome Reference Consortium (humanpangenome.org).[39] We called germline insertions using a combination of assembly-based methods (DipCall v0.20,[73] Minigraph v0.34[74]) and an alignment-based method (Sniffles v2.2[51]).

After calling the insertions using three different callers, we extracted insertions annotated to any repeat type using RepeatMasker,[52] which database was merged with both Dfam[91] and RepBase[92] v.20181026. This process yielded a set of candidate mobile element insertions. For more detailed repeat annotation, we used BLAST[76] against consensus sequences from RepeatBrowser,[77] To focus on potentially active elements, we restricted our analysis to insertions from young subfamilies known to be active in the human genome: L1HS and L1PA2 for L1 family; *Alu*Y, *Alu*Yb8, and *Alu*Ya5 for *Alu* family; and SVA_E and SVA_F for SVA family.

We then developed a custom pipeline to identify the canonical signatures of target-primed reverse transcription: target site duplications (TSD) and polyA/T tails. To identify TSD sequences, we generated synthetic paired-end reads for each insertion. For each event, we extracted 500bp of the upstream and downstream reference sequence to serve as the left and right pseudo-read, respectively. Then, the first 100bp of the insertion sequences was appended to the left pseudo-read, and the last 100bp of the insertion sequence was prepended to the right pseudo-reads. These synthetic reads were then aligned to the reference genome using bwa-mem with default options, and the resulting CIGAR strings were inspected. A TSD was identified if the alignment match ("M" in the CIGAR string) extended beyond the original 500bp reference sequence length. This method allowed us to annotate the TSD sequence and its size, defining a canonical TSD as being between 4 and 30bp in length. polyA/T tails were identified by searching for sequences of more than 10 A at the 3′ end or 10T at the 5′ end of the insertion, allowing for up to two mismatches after excluding the TSD sequence.

Based on the subfamily, TSD, polyA/T features, we stratified the filtered insertions into three confidence tiers. Tier 1 included insertions from young subfamilies with canonical TPRT signatures: both a TSD and a polyA/T. Tier2 included insertions with only a polyA tail including events with target site deletions, and Tier 3 comprised all remaining young subfamily insertions, regardless of TPRT evidence. To annotate zygosity, we calculated the VAF for each insertion and applied a Gaussian Mixture Model to the resulting distribution to distinguish between heterozygous and homozygous events. Finally, we manually inspected all candidate insertions by comparing them with another manually curated set of mobile element insertions in the HG002 cell line from a previous study,[31] and then applied the same pipeline to the other five cell lines.

After establishing a high-confidence set of MEIs for each of the six HapMap cell lines, we combined all identified insertions to simulate the somatic insertions in the HapMap mixture. For this simulation, any insertion present in the HG005 cell line was assigned as germline, while insertions found only in the other cell lines were assigned as somatic. The expected VAF in the cell line mixture was then calculated using the following equation:

$$VAF_{Expected} = \sum_{c \in C} R_c * Z_c$$

where C is the set of cell lines {HG005, HG02622, HG02486, HG02257, HG00, HG00438}, R_c is the mixture ratio for cell line c, and Z_c is the zygosity of the insertion in cell line c (0.5 for heterozygous, 1.0 for homozygous) (Figure S1B). For tiering of HapMap mixture insertion, we assigned the most frequently occurring tier for the final benchmarking set (Tables S1, S2, and S3).

In silico BL2009/H2009 mixtures

The H2009 lung cancer cell line exhibits extensive copy number variation, which prevents the generation of a diploid assembly from long-read sequencing data. Therefore, instead of using assembly-based methods, we applied multiple read-based structural variation callers (nanomonsv v0.7.0[75] and Severus v1.2[54]) in a paired tumor-normal mode with the matched normal BL2009 cell line, modifying only a single parameter by setting the minimum number of supporting reads to 1. The resulting insertion calls were processed using the same annotation pipeline as the HapMap cell lines, including subfamily classification, TSD identification, and polyA/T tail analysis. As the majority of detected insertions belonged to the L1HS subfamily, we constructed a benchmarking set focused exclusively on L1 elements. All insertions, particularly those reported by only a single caller, were manually inspected, and confidence tiers were assigned using the same criteria as the HapMap mixture benchmarking set (Table S14). In addition, because cancer cell lines can accumulate subclonal mutations during cell culture, we examined the VAF distribution of the insertions to distinguish the intended clonal insertions from culture-derived subclonal events. The distribution exhibited a bimodal pattern (Figure S2E), with a distinct low-VAF peak centered around 1.5%, suggesting that these insertions likely represented subclonal events arising during cell culture. Because the *in silico* normal–tumor mixtures were designed based on the VAF levels of clonal

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

insertions (around 50%), only insertions with VAFs above the 15th percentile of the clonal VAF distribution (VAF = 15%) were included in the benchmarking set (Figure S2E). Insertions below this threshold were not removed but instead designated as a masking set and excluded from performance evaluation to avoid underestimating caller accuracy due to the presence of unintended subclonal events (Figure S2F). Bimodal distribution analysis was performed using Scipy[87] (v1.10.1).

sMEI detection

The WGS-based methods were included if they are capable of detecting sMEIs in single-sample mode given normal human tissues where matched case–control samples are not available and meet one of the following criteria: (1) specifically designed for sMEI detection (xTea_mosaic[31] and RetroSom[23]) used for sMEI detection in previous studies[24,43] (MELT[58] and PALMER[29]), or (3) originally developed for germline MEIs but adaptable for sMEI detection with simple parameter tuning—lowering supporting read counts (xTea_long,[31] cuteSV,[50] and Snilffles2[51]) (data and code availability). Since cuteSV and Sniffles2 are general structural variant callers, we applied RepeatMasker[52] to annotate sMEI candidates for their insertion calls.

xTea_mosaic

We ran a mosaic mode (-M) of xTea[31] (v0.1.9) with the modified parameters related to the number of supporting clipped and discordant reads for sMEI detection at the lower frequency. We lowered the cutoffs for the minimum number of clipped reads (–nclip 2, default 3), clipped reads whose mates map in the repetitive region (–cr 0, default 1), clipped reads in the filtering step (–nflip 1, default 3), and discordant pair in the filtering step (–nfdisc 1, default 5).

MELT

Mate coordinates and insert size field tags were added to BAM files using samtools[93] (fixmate) to improve the runtime of MELT. MELT[58](v2.2.2) was implemented with default parameters in 'Single' genome analysis mode.

RetroSom

RetroSom[23] (v2) recommends to run control and case samples independently (step 1–6) and combine the results for somatic insertion detection (step 7–8). However, since our study aims to detect sMEI in the normal tissues without matched control-case samples, we ran RetroSoms as single sample mode by running step 1–6 using only HapMap mixture samples. The maximum number of supporting reads to be considered as a putative somatic insertion (-n) was set as 30% of sequencing depth of the sample. We omitted the polymorphic filtering step since the sMEI in the HapMap mixture originated from the germline insertions.

PALMER

PALMER[29,30,43,47] (v2.0.1) was implemented with the default parameter setting. The calls with at least one and two high-confidence supporting reads were included in the callsets of PacBio-HIFI and ONT sequences, respectively. PALMER then implements a masking process to exclude the calls in low-confidence regions for the final callsets. The mask file can be found here https://doi.org/10.5281/zenodo.15733390.[94]

xTea_long

Since xTea_long[31] (v0.1.0) was originally developed for germline insertion detection, we modified the parameters related to the number of supporting reads to detect sMEIs (–user -c 1) in the clip (-C) and assembly (-A) step.

cuteSV

cuteSV[50] (v2.1.2) was implemented with a modified parameter for the number of supporting reads for sMEI detection (–min_support 1, default 10). Since cuteSV was originally developed for SV detection, we applied RepeatMasker[52] and remained the calls pass following three criteria: (1) 'L1', '*Alu*', 'SVA' class, (2) less than <10% divergence (combined % of substitution, deletion, and insertion bp) by comparing consensus ME sequence, (3) insertion sequence larger than 50bp.

Sniffles2

We ran two different modes of Sniffles2 (v.2.6.3.).[51] We first ran a germline mode of Sniffles2 with a modified parameter for the number of supporting reads for sMEI detection (–minsupport 1). Additionally, we ran a mosaic mode of Sniffles2 for sMEI detection (–mosaic). Since the minimum VAF levels of mosaic mode was set as 5%, we lowered the threshold of minimum VAF levels to also detect the sMEI lower than 5% (–mosaic-af-min 0, default 0.05). Then, we applied RepeatMasker[52] and the same filtering criteria with cuteSV. The performance from modifying the number of supporting reads was mainly used for the evaluation and that of mosaic mode was provided in Tables S7, S8, and S9.

TEnCATS

For both L1 and *Alu* datasets, on-target rate, and MEI calling were characterized using NanoPal (data and code availability), adapted from our prior study.[47] Briefly, reads were filtered using Minimera (data and code availability) to remove sequencing artifacts, aligned to the reference genome, and non-reference MEI insertions were detected with PALMER. Next, TEnCATS reads were classified as on-target using BLASTn, and reads that support the MEIs were clustered by location. Variant calls with fewer than two read support were filtered from the final results.

eHAT-seq

Bulk eHAT-seq data were analyzed using a computational pipeline for the SMaHT HapMap cell line mixture, adapted from our prior study.[19] Raw, de-multiplexed FASTQ reads were trimmed of adapters and low-quality bases. Read pairs containing L1Hs sequences in Read 2 were retained; the corresponding Read 1, capturing the 3′ genomic flank, was aligned to the human reference genome (build hg38/GRCh38, GCA_000001405.15, without alternate contigs). Putative L1 insertion sites were then identified by peak calling.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

Identified peaks were annotated by comparison to insertion databases and with their genomic context and quantitative features (i.e., reads per million [RPM], polyA/T tail, template count). Peaks were subsequently classified as known reference (KR), known non-reference (KNR), unknown (UNK)/somatic insertions, or artifacts. Peak reclassification was applied to the eHAT-seq libraries through an integrative analysis of replicates. For the artificial KNR L1 spike-ins designed to mimic somatic events in HapMap mixture, a rescue step was implemented: after initial classification as KNRs, they were rescued and re-routed to undergo the same downstream filtering to generate the final unknown (UNK)/somatic insertion callset.

Performance evaluation

To evaluate the performance of each sMEI detection method, we computed recall, precision and F1 score of each call set against the benchmarking set. Using the germline insertion information (present in the HG005) in the benchmarking set, germline calls were excluded from the callsets to ensure that evaluation was limited to somatic candidates. A call was classified as a true positive (TP) if its breakpoint was located within ±100 bp (for WGS-based methods) or ±200 bp (MEI-targeted sequencing) of an insertion in the benchmarking set; otherwise, it was considered a false positive (FP).

Recall was defined as the proportion of true sMEIs in the benchmarking set that were successfully detected by the method:

$$Recall = \frac{TP}{TP+FN}$$

Precision was defined as the proportion of detected insertions that were true sMEIs. To avoid penalizing calls that are absent in the current benchmarking set tier but present in a lower-confidence tier, we masked such calls during precision calculation. For example, when evaluating a call set against the tier 2 benchmarking set, we excluded calls that would be considered true positives in the tier 3 set from the false positive count. In *in silico* cancer-normal mixture, we additionally excluded putative subclonal insertion events using the masking set.

Precision was then calculated as:

$$Precision = \frac{TP}{TP+FP_{masked}}$$

F1 score, the harmonic mean of recall and precision, was calculated as:

$$F1 = 2 * \frac{(Recall * Precision)}{(Recall+Precision)}$$

All metrics were evaluated globally as well as in stratified analyses by genomic regions and VAF. Specifically, to evaluate the performance in each genomic region, we used two genomic stratifications: (1) 151-mer unique vs. non-unique region[53] and (2) ME-free vs. ME region. We obtained genomic stratification bed files from the SMaHT data portal (https://data.smaht.org/) and used the combination of 'easy' and 'difficult' region as 151-mer unique and 'extreme' region as non-unique region, respectively. We defined the ME region based on 'L1', 'Alu', and 'SVA' annotation in the reference genome by RepeatMasker.[52] The benchmarking sets and the callsets then were stratified by genomic regions based on the breakpoint positions. Recall and precision were then calculated independently within each region using the same equation as in the global analysis:

$$Recall_{region} = \frac{TP_{region}}{TP_{region}+FN_{region}}$$

$$Precision_{region} = \frac{TP_{region}}{TP_{region}+FP_{region}}$$

Similarly, to evaluate the performance in each VAF range, we used different VAF definitions in recall and precision calculation. In recall, the benchmarking set was stratified into four VAF bins ($0\%<VAF\leq1\%$, $1\%<VAF\leq3\%$, $3\%<VAF\leq5\%$, and $VAF>5\%$) and three insertion length groups ($length\leq1kb$, $1kb<length\leq5kb$, $length>5kb$), based on the expected VAF and the length that we defined in the benchmarking set. Each VAF or length bin was designed to include enough insertions to ensure robustness of the performance (Table S4). The entire call set was then compared to the stratified benchmarking set in each VAF or length bin to calculate recall:

$$Recall_{\ VAF\ or\ Length} = \frac{TP_{expected\ VAF\ or\ Length}}{TP_{expected\ \ VAF\ or\ Length} + FN_{expected\ \ VAF\ or\ Length}}$$

On the other hand, for precision, we used caller VAF or length that each method reports for each call, since the expected VAF and length for false positives cannot be defined in the benchmarking set. So, the call set was stratified into same VAF and length bins with the recall using the caller VAF and insertion length and the entire benchmarking set was compared to the stratified call set to calculate precision:

$$Precision_{\ VAF\ or\ Length} = \frac{TP_{caller\ \ VAF\ or\ Length}}{TP_{caller\ \ VAF\ or\ Length} + FP_{caller\ \ VAF\ or\ Length}}$$

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

Integrative strategy combined with call set and raw read alignment signals

We selected xTea_mosaic and PALMER for the integration—two methods that showed the best overall performance in benchmarking in each platform. The supporting type of each call in the integrated call set was classified into five types according to short-read call (SR_call), short-read raw signal (SR_raw), long-read call set (LR_call), and long-read raw signal (LR_raw). In the integrated call set, we first included calls supported by both methods (SR_call & LR_call, S1) with overlapping insertion breakpoints (within 200bp). Except for the calls belonging to S1, we searched the raw signals in orthogonal platforms. To this end, we extracted the 7bp from the breakpoint coordinates in the clipped reads and examined whether this sequence matched with the insertion sequence identified in the other platform. The calls that have raw signals from the other platforms were classified as (SR_call & LR_raw (S2) or LR_call & SR_raw (S3)). For the calls supported by a single platform (SR_call-only (S4) and LR_call-only (S5)), we applied method-specific FP filtering strategies. For the xTea_mosaic set, we excluded the calls located in the non 151-mer unique regions (S4′). For the PALMER call set, we ran BLAST[76] using the complete insertion sequence and only retained calls classified into young subfamilies (L1HS and L1PA2). We also excluded calls that were supported by only supplementary alignment reads (S5′). We also applied the same FP strategy to S3 (S3′). The final integrated callsets included the calls belonging to five supporting types: S1, S2, S3′, S4′ and S5'.

Refinement of sMEIs using phasing information

To determine the sMEI, we performed local phasing using long-read sequencing data. First, we called single nucleotide polymorphisms (SNPs) in a 40 kb window centered on the insertion locus using GATK4 HaplotypeCaller.[89] From the resulting variants, we selected up to five high-quality (QUAL >1000) heterozygous SNPs (hetSNPs) that were closest to the insertion site for the analysis. For phasing, we analyzed long reads that covered both a selected hetSNP locus and the corresponding insertion site. This phasing process was conducted for each of the selected hetSNPs (up to five per insertion), and the final classification was assigned based on the most frequent outcome. An insertion was categorized as 'No hetSNPs' if no suitable heterozygous SNPs were identified within the 40 kb flanking region, or as 'No covering reads' if no single long read spanned both the insertion site and a selected hetSNP locus. If the insertion was linked to alleles from both haplotypes, it was classified as 'Insertion in multiple haplotypes'. The insertion was classified as 'Germline' if it was confidently linked to a single haplotype while the alternate haplotype was well-covered by reads lacking the insertion. Finally, an insertion was labeled as 'Somatic MEI (sMEI)' if it was linked to a single haplotype but supported by a low number of reads, indicative of a potential somatic or mosaic event.

Refinement of sMEIs using phasing information from DSA

We employed two strategies to leverage phasing information for refining somatic MEIs using DSA: an GRCh38 reference-based approach and a DSA reference-based approach. For the GRCh38-based approach, tumor-normal mixture data were aligned to the GRCh38 reference genome. Reads were then phased using LRphasing (data and code availability) based on phased heterozygous SNVs identified from the DSA, and phased VAFs were then computed for each haplotype. The haplotype with the higher phased VAF was designated as hapA, provided the two haplotypes were supported by at least five reads. To further reduce false-positive somatic sMEI candidates, we applied an empirical decision boundary defined by $VAF_{HapB} = 0.2 \times VAF_{HapA}$. Variants lying outside this boundary were flagged as probable artifacts and excluded from downstream analyses.

DSA-based alignment analysis

Six samples were aligned to the DSA using pbmm2. To prevent alignment to unphased contigs, only contigs assigned to haplotype 1 and haplotype 2 were extracted and merged to create the reference genome. Callers (PALMER, cuteSV, Sniffles2 and xTea_long) were then run on each DSA-aligned sample with the DSA as the customized reference genome. For xTea_long, a DSA-specific repeat library was generated with the library preparation module of xTea (data and code availability).

We aligned DSA contigs to GRCh38 and performed coordinate projection between DSA and GRCh38 with a custom pipeline (data and code availability). A projection was declared successful only if the relevant contig-to-reference alignment had MAPQ $\geq$60 and fully covered the ±100 bp window around the queried coordinate. Sites lacking such coverage were excluded from rate denominators. The same criterion was applied symmetrically for DSA-to-GRCh38 projections. We performed these projections for MEI calls and the population-polymorphic germline sites. DSA-aligned calls were deduplicated within each caller and mixture level after coordinate projection to GRCh38.

To evaluate whether DSA preferentially reduced false positives (FPs) in segmental duplications (SD) more than outside SD. Let g = (caller, mixture, method), for each g, among FP calls, we counted.

- Y_g^{SD} = # FP in SD.
- Y_g^{nonSD} = # FP outside SD.
- $n_g = Y_g^{SD} + Y_g^{nonSD}$.

Polymorphic germline calls were excluded from the analysis.

We fit a binomial GLM with logit link (logistic regression) to FP composition:

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

CellPress
OPEN ACCESS

$$Y_g^{SD} \sim Binomial\left(n_g, \pi_g\right), logit\left(\pi_g\right) = \alpha + \beta I\{Method = DSA\ aligned\} + \gamma_{caller} + \delta_{mixture}$$

Caller and mixture level were included as fixed effects with 4 levels each. Coefficient β represents the adjusted log odds ratio of an FP occurring within an SD versus outside an SD for DSA-aligned calls relative to raw calls, controlling for caller and mixture level. The adjusted odds ratio (OR) was calculated as e^{β}.

We tested H_0:$\beta \geq 0$ vs. H_1:$\beta < 0$ using the Wald z statistic from GLM. We reported the adjusted OR with 95% confidence interval and the one-sided p-value. We also checked for overdispersion, where it indicated no overdispersion (residual deviance/df = 0.923, p = 0.456), supporting the binomial variance assumption. All analyses above were implemented in R (v4.4.1).[95]

Constructing the germline full-Length L1HS sequence catalog in haplotype resolution

Germline L1HS elements can exist in both reference and non-reference forms. To extract all full-length reference L1HS sequences, we first identified candidate loci from RepeatMasker[52] tracks and filtered out entries shorter than 4000 bp. Next, we extracted reads mapped to L1HS regions with mapping quality over 30, excluding supplementary alignments. These reads were converted into FASTA format and assembled using hifiasm[96] with default parameters except for the -f0 option. Unitig GFA files generated from the assembly were converted to FASTA, and RepeatMasker was used to annotate L1HS loci. The L1HS sequences were then extracted from these locally assembled FASTA files. If the L1HS sequences were identical between the two haplotypes, only one unitig was present in the FASTA file. Otherwise, two unitigs represented the distinct haplotypes.

Non-reference L1HS sequences were extracted using a similar approach, except for the identification of L1HS loci. When a non-reference L1HS insertion occurred at a locus overlapping a reference L1HS, RepeatMasker annotations alone were insufficient to identify the correct position. Therefore, we re-aligned the locally assembled FASTA files to the reference genome using pbmm2, and extracted the insertion sequences from the aligned assemblies.

After obtaining fully resolved L1HS sequences, we identified full-length elements via BLAST[76] alignment against the 5′UTR, ORF1, and ORF2 sequences. The reference sequences for 5′UTR, ORF1, and ORF2 are provided in the supplementary materials.

Source identification based on transduction sequences

To extract the transduction sequence from the insertion sequence, we first performed a BLAST[76] search using repeat consensus sequences. If either the left or right flanking sequence of an insertion was longer than 50 bp, we further aligned these flanking sequences to the human genome. If the flanking sequence mapped within 5000 bp of any full-length L1HS catalog, whether reference or non-reference, we defined that locus as the source.

Source inference based on internal L1HS sequences

During mobile element insertion, small structural variations such as insertions, deletions, or inversions can occur. Additional poly-A signals may also be appended during transcription. To avoid confounding elements, we extracted only the internal L1HS sequence based on BLAST[76] annotation of the insertion sequence and further removed the polyA/T tails. The trimmed sequence was then aligned on the haplotype-resolved FL-L1HS catalog using BLAST,[76] and the locus with the highest bitscore was identified as the putative source. If the bitscores of the top and second hits were identical, we annotated the insertion as ‘‘source untraceable’’.

Discovery of sMEIs in human tissue homogenates

We first ran xTea_mosaic (Illumina) and PALMER (PacBio and ONT) to generate callsets for each tissue WGS sample generated from each of five GCCs (Table S4). For the PALMER call set, MEI subfamily of each call was annotated using BLAST,[76] and the old subfamily calls and supplementary alignment calls were proactively filtered out. Then, we applied haplotype phasing to filter out germline calls. To exclude the calls that arose from misalignment or misclassification during haplotype phasing, we cross-checked the haplotype phasing results across the SMaHT tissue samples from the same donor and the calls classified as germline in at least one sample were also classified as germline in other samples. To further exclude germline calls, we curated the polymorphic MEI insertion positions from previous studies that used short-read[31,33,58,97–100] and long-read[29,30,40] sequencing and excluded overlapping sites (±100 bp).

Next, we applied the integrative strategy (‘Integrative strategy combined with call set and raw read-level signals’ in STAR Methods) to all combinations of short-read and long-read callsets. For example, ST003 sample has five short-read and five (three PacBio and two ONT) long-read WGS samples, and we performed the integrative strategies for 25 combinations. All candidates were manually inspected by a structured, multiple criteria review of each candidate insertion, performed by at least two independent reviewers. Specifically, manual inspection included the following criteria: (1) at least one target-primed reverse transcription (TPRT) feature using IGV – target site duplication (TSD; 2–30 bp) or poly(A) tail (≥10 bp); (2) correct alignment of both 5′ and 3′ flanking sequences to the breakpoints using BLAT[88] (https://genome.ucsc.edu/cgi-bin/hgBlat). Candidates that did not meet these criteria were classified as false positives. We described three classes of false positives observed during manual inspection (translocation-like events, other insertion/deletion-derived events, and ONT-specific sequencing errors) in Figure S15 and Supplemental Notes. The VAF of each insertion was estimated by average the VAF in each sample which was calculated by dividing the number of supporting reads by the local depth in the long-read WGS data. The L1HS subfamily of source L1s was identified by canonical positions defining L1HS subfamilies from previous work.[65]

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

Please cite this article in press as: Wang et al., Multi-platform framework for mapping somatic retrotransposition in human tissues, Cell Genomics (2026), https://doi.org/10.1016/j.xgen.2026.101418

Experimental PCR validation

Putative somatic L1 insertions were validated using two nested PCR strategies adapted from the previous work.[19] For 3′ junction validation, the first-round PCR (20 μL) contained 20 ng of gDNA, 10 μL of 2× KAPA2G Robust HotStart ReadyMix (KAPA Biosystems, KK5702), and 0.5 μL each of L1Hs-AC-28 and a site-specific 3′ primer (20 μM). Cycling conditions were: 95°C for 3 min; 20 cycles of 95°C for 15s, 60°C for 15s, and 72°C for 15s; and a final extension at 72°C for 1 min. For the second round, 1–2 μL of the first-round product was re-amplified with L1Hs-nest-28 and a nested 3′ primer for 20, 25, or 30 cycles to optimize target band visibility.

For full-length PCR validation, the first-round PCR contained 50 ng of gDNA with 1 μL of PrimeSTAR GXL DNA Polymerase (Takara Bio, R050A), 5 μL of PrimeSTAR GXL buffer (5×), 2 μL of dNTPs (2.5 mM each), and 1 μL each of site-specific 5′ and 3′ primers. Cycling conditions were: 98°C for 3 min; 30 cycles of 98°C for 20 s, 60°C for 20 s, and 68°C for 1 min 15 s; and a final extension at 68°C for 3 min. The resulting full-length amplicon was excised from the gel, purified, and used as a template for a second amplification round with nested primers.

All amplicons were visualized on 2% agarose gels. The 3′ junction PCR products were submitted for Amplicon-EZ sequencing (Azenta, Inc.). Full-length L1 amplicons were excised and purified using the PureLink Quick Gel Extraction Kit (Invitrogen, K210012) and PCR Purification Kit (Invitrogen, K310001). A portion of the purified product was submitted for Sanger sequencing, while the remainder was sonicated into fragments (Covaris R230) for deep sequencing via Amplicon-EZ. Primer sequences used for validation are listed in Table S20.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF MICHIGAN from ClinicalKey.com by Elsevier on November 09, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.