1146–1163 *Nucleic Acids Research, 2020, Vol. 48, No. 3* *Published online 19 December 2019*
doi: 10.1093/nar/gkz1173

Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology

Weichen Zhou [1], Sarah B. Emery[2], Diane A. Flasch[2], Yifan Wang[2], Kenneth Y. Kwan[2,3], Jeffrey M. Kidd [1,2], John V. Moran [2,4] and Ryan E. Mills [1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA, [2]Department of Human Genetics, University of Michigan Medical School, 1241 East Catherine Street, Ann Arbor, MI 48109, USA, [3]Molecular and Behavioral Neuroscience Institute, University of Michigan Medical School, 109 Zina Pitcher Place, Ann Arbor, MI 48109, USA and [4]Department of Internal Medicine, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, MI 48109, USA

Received August 07, 2019; Revised November 14, 2019; Editorial Decision December 04, 2019; Accepted December 05, 2019

ABSTRACT

Long Interspersed Element-1 (LINE-1) retrotransposition contributes to inter- and intra-individual genetic variation and occasionally can lead to human genetic disorders. Various strategies have been developed to identify human-specific LINE-1 (L1Hs) insertions from short-read whole genome sequencing (WGS) data; however, they have limitations in detecting insertions in complex repetitive genomic regions. Here, we developed a computational tool (PALMER) and used it to identify 203 non-reference L1Hs insertions in the NA12878 benchmark genome. Using PacBio long-read sequencing data, we identified L1Hs insertions that were absent in previous short-read studies (90/203). Approximately 81% (73/90) of the L1Hs insertions reside within endogenous LINE-1 sequences in the reference assembly and the analysis of unique breakpoint junction sequences revealed 63% (57/90) of these L1Hs insertions could be genotyped in 1000 Genomes Project sequences. Moreover, we observed that amplification biases encountered in single-cell WGS experiments led to a wide variation in L1Hs insertion detection rates between four individual NA12878 cells; under-amplification limited detection to 32% (65/203) of insertions, whereas over-amplification increased false positive calls. In sum, these data indicate that L1Hs insertions are often missed using standard short-read sequencing approaches and long-read sequencing approaches can significantly improve the detection of L1Hs insertions present in individual genomes.

INTRODUCTION

At least 45% of the human genome is composed of transposable element (TE)-derived sequences (1). TEs can be subdivided into four major categories: (i) DNA transposons, (ii) long terminal repeat (LTR) retrotransposons, (iii) long interspersed elements (LINEs), and (iv) short interspersed elements (SINEs). L1 represents a subclass of LINEs and L1-derived sequences comprise ∼17% of the human genome (1,2).

The overwhelming majority (>99.9%) of L1-derived sequences contain mutations (i.e. 5′ truncations, internal DNA inversion/deletion structures, and/or point mutations) and cannot move (i.e. retrotranspose) to new genomic locations (1,3–6). However, an average human genome contains ∼80–100 active full-length human-specific L1s (L1Hs) (7–9) and a small number of highly active, or 'hot,' L1Hs sequences are responsible for the bulk of human retrotransposition activity (7,8,10,11).

Retrotransposition-competent L1s (RC-L1s) are ∼6 kb in length (12,13) and contain a 5′UTR that harbors an internal RNA polymerase II promoter, two open reading frames (ORF1 and ORF2), and a 3′UTR that ends in a polyadenosine (poly(A)) tract (12,14–16). L1 retrotransposition occurs by target-site primed reverse transcription (TPRT) (17–19), which requires biochemical activities encoded by the L1-encoded proteins (ORF1p and ORF2p), full-length polyadenylated L1 RNA, and host-encoded proteins (15,18,20–22). TPRT leads to the insertion of an L1 at a new genomic location. The newly inserted L1 contains diagnostic structural hallmarks (21) and: (i) is often 5′ truncated and ends in a 3′ poly(A) tract, (ii) is flanked by variable length target site duplications (TSDs), (iii) inserts into an L1 ORF2p endonuclease (L1 EN) consensus cleavage site (5′-TTTTT/AA, and variants of that sequence) (18,23,24),

*To whom correspondence should be addressed. Tel: +1 734-647-9628; Email: remills@umich.edu

and (iv) sometimes contains additional genomic sequences at their 3′ end (known as L1-mediated 3′ transductions) (16,25–27). ORF1p and/or ORF2p also can act *in trans* to mobilize SINEs (e.g. *Alu* elements and SVAs), non-coding RNAs, and messenger RNAs (21,28). Together, these L1-mediated retrotransposition events comprise at least 11% of the human genome (1,21).

L1-mediated retrotransposition events can be mutagenic and germline retrotransposition events within the exons or introns of genes can result in null or hypomorphic expression alleles that lead to sporadic cases of human disease (29). Moreover, recent studies have revealed that somatic L1 retrotransposition events can act as driver mutations in certain cancers (30). Somatic L1 retrotransposition events can also occur in neuronal progenitor cells (31–34), leading to suggestions that they may play a role in the etiology of neuropsychiatric diseases (35–38). However, the various sequencing methodologies and bioinformatics pipelines used to detect somatic L1Hs insertions in human neurons have yielded conflicting data with regard to the rate of L1 retrotransposition and whether specific brain areas accommodate higher levels of L1 retrotransposition than others (33,34,39–43). Indeed, the difficulty in uniquely aligning short-read sequences to repetitive genomic regions likely leads to an under-representation of L1Hs insertion in complex and/or repetitive genomic regions.

De novo L1 retrotransposition events can insert within endogenous L1s or other repeated DNA sequences, making their detection by short-read sequencing difficult (24). The advent of third-generation DNA sequencing technologies provides a powerful way to characterize repeat-rich genomic regions (44,45). However, computational approaches to identify and characterize L1Hs insertions in repeat-rich genomic regions require refinement, as many approaches only label L1Hs events as 'generic' insertions and/or are not designed to consider the unique sequence hallmarks of new L1 insertions (46–48). Moreover, approaches using pairwise sequence alignments to annotate repeat sequences may have difficulty differentiating between new retrotransposition events and duplications of existing genomic sequences containing endogenous repeats (49).

Here, we describe a computational tool, PALMER, which pre-masks long reads aligning to endogenous repeats present in the human reference sequence. We then applied this approach to identify non-reference L1Hs insertions in recently generated Pacific Bioscience (PacBio) long-read sequencing data from the well-characterized NA12878 benchmark genome. We assessed our approach by comparing non-reference L1Hs insertions detected by PALMER (herein called PALMER L1Hs insertions) with call sets identified using standard Illumina WGS data, whole genome 3′ targeted L1 capture assays, as well as recent large-scale efforts, including Genome in a Bottle (50), and the Human Genome Structural Variation Consortium (48). Finally, we used our PALMER call set to assess the efficacy of single-cell whole genome DNA amplification (WGA) in detecting L1Hs insertions in NA12878. Together, these efforts revealed that L1Hs insertions are often missed due to their integration into complex and repetitive genomic regions and/or are often incompletely annotated because the candidate L1Hs insertions were not systematically assessed for the presence of L1 structural hallmarks. In sum, we demonstrate that a combination of long-read sequencing data with repeat pre-masking and systematic feature identification can identify many previously overlooked L1Hs insertions into complex genomic regions.

MATERIALS AND METHODS

Resolving germline non-reference L1Hs insertions from PacBio data

We developed an approach, named PALMER (PreMAsking Long reads for Mobile Element inseRtion), to detect L1Hs insertions in the NA12878 benchmark genome. PALMER first pre-masks aligned long-read sequences containing known reference L1 sequences obtained from Repbase (51) and then searches against a 'hot L1' sequence (L1.3; GenBank: L19088) (9) to detect non-reference L1Hs insertions within the remaining unmasked sequences in the genome (Figure 1). To facilitate this process, neighboring pre-masked repeats within 100 bp of each other in individual PacBio subreads were combined into larger segments.

The Genome Institute at Washington University School of Medicine generated 50× coverage NA12878 PacBio sequence data (NCBI Sequence Read Archive: PRJNA323611). Notably, this is the same data resource used in a recent publication (49. All subsequent analyses were carried out using the hs37d5 (GRCh37+decoy) reference genome (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/). Ambiguous aligned reads with a samtools (52) FLAG value larger than 255, which excludes supplementary, secondary, and duplicate alignments, as well as low mapping-quality reads (MAPQ <10) in the PacBio data were excluded from subsequent analyses.

We implemented several filtering strategies to reduce potential false positive calls. We first implemented a lower threshold for the number of individual PacBio subreads required to consider loci for putative L1Hs insertions. This threshold was derived based on the mean (μ) and standard deviation (σ) of the PacBio sequence coverage across all putative insertion sites (Supplementary Figure S1A). We excluded sites with fewer than ($\mu - \sigma$) supporting subreads, where $\mu - \sigma = 5$ subreads was used for this analysis. The number of subreads required is a heuristic parameter and can be adjusted by the user to increase sensitivity at the potential cost of specificity. We further required that the putative non-reference L1Hs insertion contain ≥25 bp of sequence in the PacBio subread that is identical to the L1.3 sequence. This heuristic cut-off is based on the established use of oligonucleotides that are 25 bp in length for microarray hybridization (53) and is an adjustable parameter. We also require the presence of a poly(A) sequence of ≥20 bp in length and that the putative insertion is bounded by identical sequences of ≥6 bp in length reflective of target site duplications (TSDs).

To determine how sequencing read coverage influences the performance of PALMER, we conducted a downsampling analysis. We randomly chose 80%, 60%, 40% and 20% of the whole-genome PacBio reads to generate 40×, 30×, 20× and 10× coverage genome sequences, respectively. We then calculated the number of non-reference

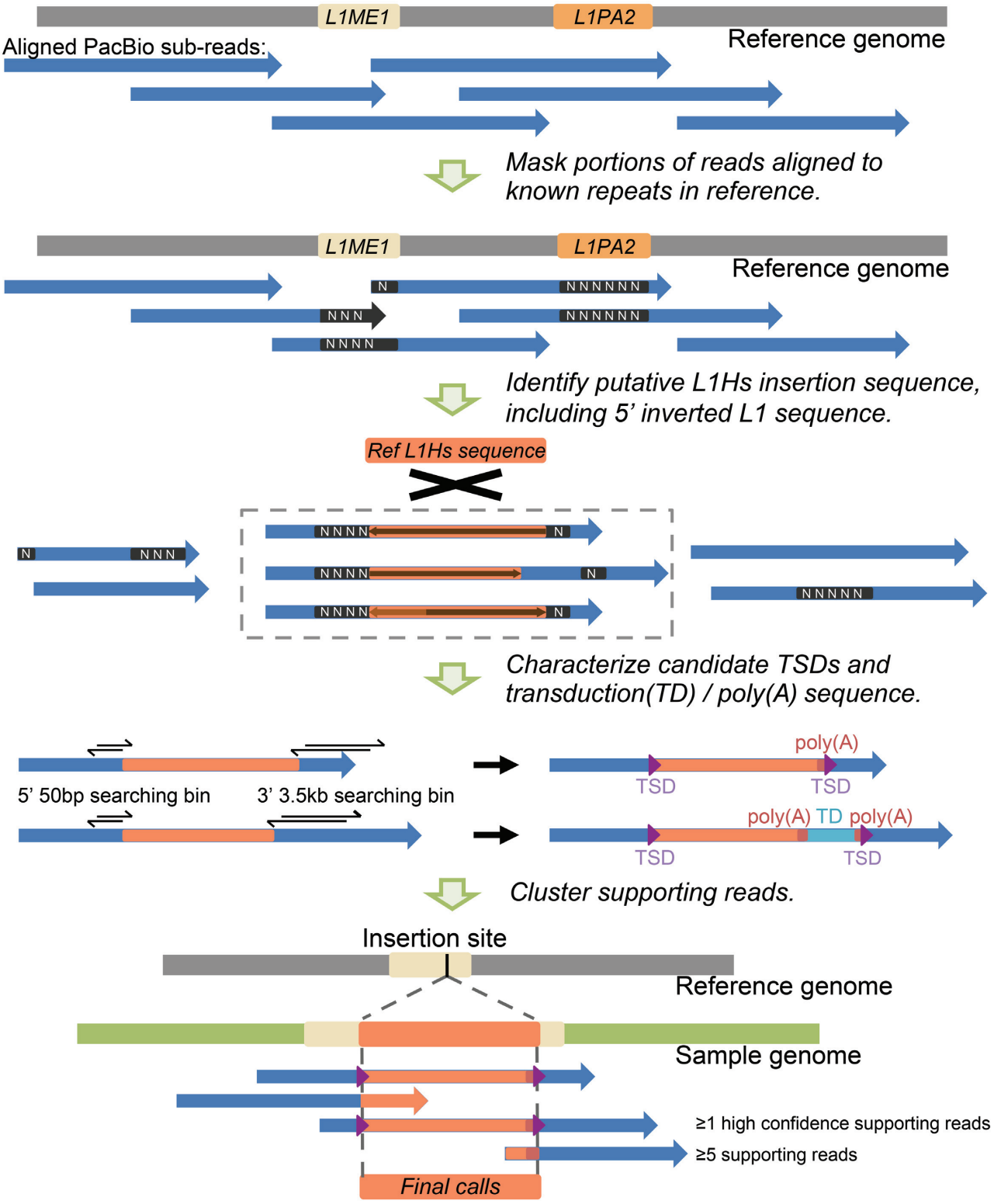

Figure 1. PALMER identifies non-reference L1Hs insertions from PacBio data. Reference-aligned BAM files from long-read technology are used as input. Known repeats (L1s, *Alu*s or SVAs in reference) are used to pre-mask the portions of individual reads that align to these repeats. After the pre-masking process, PALMER searches PacBio subreads against an insertion sequence, L1.3 (GenBank: L19088), and identifies reads with a putative insertion sequence (including 5′ inverted L1 sequence, if available) as candidate supporting reads. PALMER searches the bins in 50 bp 5′ upstream and 3.5 kb 3′ downstream of insertion sequence for each read and then identifies candidate TSD motifs, 5′ transduction and poly(A) sequence. All supporting reads are then clustered at each locus and those with a minimum number of supporting events are reported as putative insertions.

L1Hs insertions detected in our original PALMER call set at the various sequence coverage depths.

To improve the accuracy of non-reference L1Hs insertion sequences derived from individual PacBio subreads, which have lower per-read base pair accuracy, we used local sequence alignments and PacBio error correction strategies. Error correction was conducted by applying CANU (54) to PacBio subreads that contain the PALMER L1Hs insertion sequence, allowing the generation of error-corrected reads that served as inputs for local re-alignment using the long-read aligning software BLASR (55). A second-pass of the PALMER pipeline then was executed using these locally aligned error-corrected reads to generate a high-confidence call set of germline non-reference L1Hs insertions (summarized in Supplementary Table S1). All command lines running pipelines in this work can be found in the GitHub repository (https://github.com/mills-lab/PALMER_Pipelines). We also generated recurrence (dot) plots by comparing the reference sequence to the sequence of individual PacBio error-corrected reads using BLASTn (56,57), as part of our orthogonal validation efforts (Supplementary Table S2).

Validation from sequence data from BLAST database and clone data

We used BLASTn (56,57) to search for L1Hs sequences identified in other studies that are present in the NCBI nr/nt database. We required at least one matched alignment to be continuously extended through the insertion location with ≥99% identity to an error-corrected read containing an L1Hs insertion sequence.

To assess whether our predicted non-reference L1Hs insertions could be supported from end-sequence clone pair information generated from NA12878 fosmid libraries (58), we compared deviances of end-sequence alignment distances to the expected L1Hs insertion lengths on the same haplotype. Briefly, we obtained 839 373 clone end-sequence pairs with insert sizes between 10 kb and 100 kb and mapped them to hs37d5 using BWA-MEM (59). We calculated the insert size for each individual clone pair and intersected them with the coordinates of PALMER L1Hs insertions. We obtained the phased single nucleotide polymorphism (SNP) information from the GIAB Project (46), which is based on phasing information obtained from physical 10× Genomics linked-read sequencing and long-read PacBio sequencing. All PacBio error-corrected reads containing the L1Hs insertion sequence and the overlapped fosmid clones were assigned to individual haplotypes by interrogating the phased SNPs within each sequence (Supplementary Figure S2A). By using this strategy, 924 fosmid clone fragments overlapping 135 L1Hs insertions were assigned to specific haplotypes. The overall distribution of these fosmid clone insert sizes shows a mean value of ∼40 kb (Supplementary Figure S2B) and we derived a value (Δ) representing the deviation from the expected 40kb fosmid clone insert size. We then regressed the PALMER-predicted insertion sizes on the values of Δ in two categories: (i) those that overlapped PacBio error-corrected reads and fosmid clone pairs that were assigned to the same haplotype, and (ii) those that overlapped PacBio error-corrected reads and fosmid clone pairs assigned to different haplotypes. A similar strategy was used to leverage phased SNPs within each sequence to assign individual PacBio error-corrected reads to specific haplotypes around insertion sites predicted by non-PALMER approaches to verify the presence of both the post- and pre-integration insertion alleles.

Short-read WGS data of NA12878 and L1Hs 5′ genomic DNA/L1 junction sequence k-mer analysis

We used the following short-read WGS data in this study: (i) 50× coverage WGS data of three genomes (NA12878, NA12891 and NA12892) in the Centre d'Etude du Polymorphism Humain (CEPH) pedigree 1463 generated as part of the Illumina Platinum Genomes (the European Nucleotide Archive accession: PRJEB3381) project (60); and (ii) 30× coverage linked-read germline genome V2 data for NA12878 from 10× Genomics. We obtained the fastq file of the hs37d5 reference genome from the 1000 Genomes Project (61).

We constructed 26 bp L1Hs 5′ genomic DNA/L1 junction sequences (26mer) from the error-corrected PacBio reads containing each L1Hs insertion (Supplementary Table S1). This 26mer contains 13 bp of the L1Hs insertion sequence and 13 bp of the 5′ flanking genomic DNA sequences, respectively (Figure 2E). Concurrently, we artificially constructed 26mers to assess the specificity of L1Hs insertion predictions by randomly generating pseudo-L1Hs insertion locations in the reference consistent with the number of non-reference L1Hs insertions we identified (n = 203); we reiterated this process nine times. Hash tables of these genomes (WGS for NA12878, NA12891, NA12892, as well as 10x Genomics for NA12878 and the reference genome) for all 26mer (real or simulated) experiments were constructed using Jellyfish2.0 (62). We then counted the number of times that each 26mer appeared in each genome. An individual insertion (real or simulated) with a 26mer not present in the reference genome and exhibiting between 1 and 200 counts in the short-read data was considered as 'valid' (Supplementary Table S1). For the NA12878 10× Genomics data, we also calculated the counts in haplotype 1 (HP1), haplotype 2 (HP2) and non-haplotype information, separately, to determine whether identified 26mers were present on individual haplotypes consistent with expected Mendelian transmission.

We next applied valid L1Hs 5′ genomic DNA/L1 junction sequence k-mers (as described above) in our analysis of the 1000 Genomes phase 3 samples (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/). We downloaded aligned low-coverage sequencing bam files of all phase 3 samples (n = 2504) from the 1000 Genome FTP site. We calculated the counts of each 26mer from each insertion in every sample. The samples have a mean coverage of 4–6×; thus, we adjusted our parameters such that we considered 26mer counts that are larger than zero and less than 10 as 'valid' to support the appearance of the respective L1Hs insertion event in an individual sample. We omitted 26mers

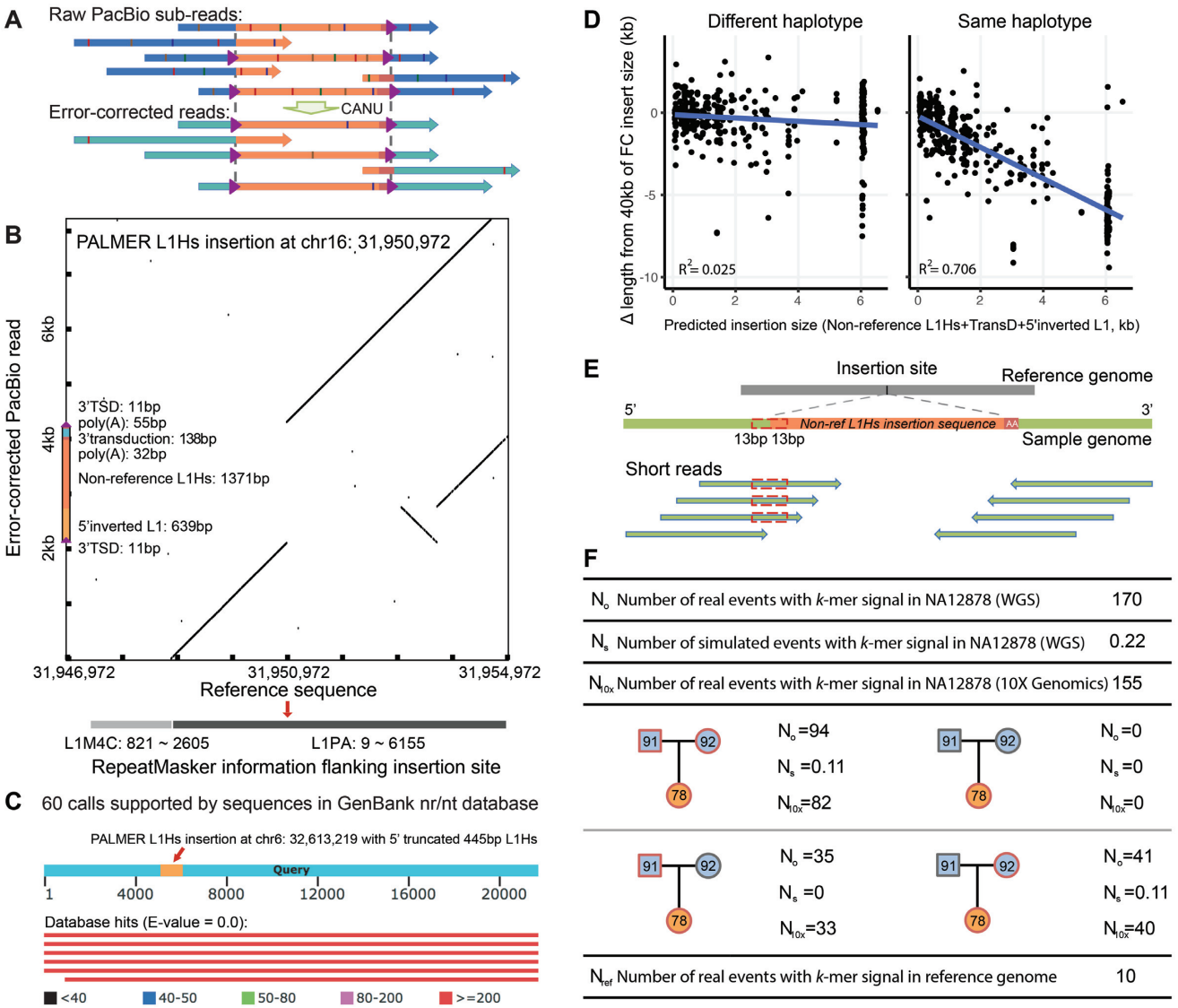

Figure 2. Validation of the PALMER L1Hs insertions using multiple strategies. (**A**) Error correction and local alignment for the supporting subreads were carried out to obtain high-quality sequence reads for each event. (**B**) A recurrence plot for a predicted insertion in chr16: 31 950 972. The structure of this event is shown on the Y-axis of the plot, including a 11 bp 5′TSD (purple arrow), a 639 5′ inverted L1 sequence (light orange bar), 1371 bp non-reference L1Hs sequence (dark orange bar), a 32 bp poly(A) tract (red bar), a 138 bp 3′ transduction (blue bar), a second 55 bp polyA tract (red bar), and a 11 bp 3′TSD (purple arrow). Y-axis is a 8 kb segment of error-corrected sequence, and X-axis is a 8 kb reference sequence at chr16 from 31 946 972 to 31 954 972. Information of RepeatMasker track was shown below in the same scale, demonstrating this event is inserted into a 6 kb reference L1PA region (red arrow shows the insertion site). (**C**) Example of supporting sequences from searching the BLAST GenBank nr/nt database using error-corrected reads containing putative insertion sequences. The lower panel shows the hits in the BLAST GenBank nr/nt database for one event (chr6: 32 613 219), whose sequence is 445 bp (orange) in an error-corrected read (green). The red bars underneath represent supporting results with *E*-value = 0 in the database. (**D**) Distributions between the predicted size of insertion sequence and the Δ length from 40 kb of the expected insert size of fosmid clone (FC) reads, categorized by fosmid clone read pairs assigned to the different haplotype of insertion sequence (left) and those assigned to the same haplotype of insertion sequence (right). (**E**) L1Hs 5′ genomic DNA/L1 junction sequence *k*-mer analysis for 203 germline non-reference L1Hs insertions of NA12878 in short-read data. The green bar represents the genome with inserted L1Hs sequence (orange); the green arrows are the short paired-end reads mapped to the genome. (**F**) L1Hs 5′ genomic DNA/L1 junction sequence *k*-mer analysis in five distinct sets: WGS data for NA12878, NA12891, NA12892 and 10× Genomics data for NA12878 and the reference genome (hs37d5). The red frame shows the events are supported in the specific genome. N_o: the number of the real event observed in WGS Illumina samples; N_s: the number of the simulated event observed in WGS Illumina samples; N_{10x}: the number of the real event observed in 10× Genomics data; N_{ref}: the number of the real event observed in the reference genome.

with ≥10 counts in >10 samples, as these likely represent over-amplified regions during sequence library construction and are likely not representative of true L1Hs insertion sequences. We then calculated the number of samples for each 'valid' 26mer and the sample frequency based on their presence in the following geographic population samples: super-population Africa (AFR, $n = 661$); East Asia (EAS, $n = 504$); Europe (EUR, $n = 503$); South Asia (SAS, $n = 489$); Americas (AMR, $n = 347$); and all phase 3 samples ($n = 2504$). We obtained individual L1Hs genotype information reported by the 1000 Genomes Project (63) for comparison.

Polymorphic L1Hs insertion datasets and genomic annotation information

We obtained other available datasets containing NA12878 sample for comparison, including dbRIP (64), a call set of structural variations (SVs, including L1Hs) from PacBio data (49), and two call sets for SVs (including generic insertions) from GIAB (i.e. calls from the Mt. Sinai School of Medicine for NA12878 PacBio data (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/) as well as calls from metaSV (65). We required that the sites overlapped and that ≥65% reciprocal overlap of length (if applicable) to be considered as intersections between two sets.

To assess whether we were able to distinguish between individual PALMER-specific L1Hs insertions inserted into specific copies of a segmental duplication (Supplementary Table S2), we used BLAT (66), as well as a map-free approach for assessing copy number based on *k*-mer counting, QuicK-mer (67). We first aligned error-corrected PacBio reads containing the non-reference L1Hs sequence to the reference using BLAT and required a full-length alignment of the entire read to have >90% identity to only one copy of an annotated segmental duplication (68). For PALMER-specific L1Hs insertions where we could not differentiate between near-identical segmental duplications, we applied QuicK-mer to the error-corrected read sequences. QuicK-mer uses a predefined set of informative 30mers to efficiently distinguish between duplicated regions of the genome. We assigned a non-reference L1Hs insertion to a specific segmental duplication copy if the error-corrected read in which it was encompassed exhibited at least one distinguishable 30mer at that position in the genome.

We obtained the RepeatMasker track (69) and gene annotation for hs37d5 from UCSC genome browser (http://genome.ucsc.edu/) (70). Accessible genome mask information was obtained from the 1000 Genomes FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/). We extended the sequences (±50 bp) surrounding each non-reference L1Hs insertion site to calculate the overlap with certain annotation information. We calculated the AT content flanking the L1Hs insertion site based on the reference sequence. The recombination rates were generated by deCODE (71) and downloaded from the UCSC genome browser. Any genome coordinates from another version of reference genome were transformed using the LiftOver tool (72).

NA12878 cell line preparation and whole genome 3′ targeted L1 capture technology for gDNA

We obtained the lymphoblastoid cell line of NA12878 from Coriell Cell Repositories (Camden, NJ) and cultured the cell line at 37°C under 5% carbon dioxide in RPMI 1640 with 2mM L-glutamine and 15% fetal bovine serum.

We performed a revised Iskow *et al.* (39) method for the preparation of the 3′ targeted L1 capture library. Genomic DNA (gDNA; 1–20 μg) was randomly sheared to 1 kb or 2 kb fragments with the Covaris M220 series. Following shearing, the DNA was end-repaired (NEBNext End Repair Module #E6050), column purified (QIAquick PCR purification Kit #28104), dA-tailed (NEBNext dA-tailing Module Protocol #E6053), and column purified once more (QIAquick PCR purification kit) so that our designed adapters with a T overhang could be ligated onto the sheared DNA. We used the same adapter sequences and followed the same annealing adapter and adapter ligation protocols as published previously (24). Ligation was followed by two amplification cycling conditions. The first PCR performed involves an L1Hs specific sequence containing the 'ACA' tri-nucleotide specific to L1Hs elements in the 3′UTR, as well as an 'outside' primer specific to the ligated adapter sequence (Supplementary Figure S3A). Amplification was performed with Invitrogen Platinum *Taq* DNA Polymerase (Invitrogen #10966018). Reactions contained 100 ng adapter-ligated gDNA, 1× manufactured PCR buffer including-Mg, 1.5 mM $MgCl_2$, 0.2 mM each dNTP, 0.4 μM each primer (L1Hs primer: 5′-ATACCTAATGCTAGATGACACA-3′ and 'outside' adapter primer: 5′-GCTTGACATTCTGGATCGATCG C-3′) and 2 U of Platinum *Taq* DNA Polymerase in a 50 μl total reaction volume. Reactions were incubated at 96°C for 2 min followed by 12 cycles of 96°C, 30 s; 60°C, 90 s; 72°C, 90 s; with a final 3-minute extension at 72°C. After the completion of the initial amplification, the first-round PCR products were used as a template (5 μl worth) in a subsequent PCR reaction. This second PCR amplification involves a downstream L1Hs primer sequence, residing upstream of the L1Hs poly(A) signal, which is tagged with an Illumina MiSeq priming sequence on the 5′ end of the primer sequence (5′-CAAGCAGAAGACGGCATACG AGATTCGAACCAGGGCACATGTATACATATGT AACTAACCTGCACAATGTG-3′), and an 'internal' adapter primer sequence tagged with Illumina MiSeq priming sequence on the 5′ end of the primer sequence (5′-AATGATACGGCGACCACCGAGATCTACACA TGTCACATGATCGATCGCTGCAGGGTATAGG-3′). Reactions contained the same components mentioned for the first PCR reactions, except the template consisted of 5 μl of the first PCR amplicon and the abovementioned primers. This second PCR reaction consisted of incubation at 96°C for 2 min followed by 20 cycles of 96°C, 30 s; 60°C, 30 s; 72°C, 90 s; with a final 5-minute extension at 72°C. Final PCR products were separated on a 1.2% UltraPure low melting point agarose gel (Invitrogen # 16520050). Products of ∼500 bp in size were gel extracted and purified with QIAquick Gel Extraction Kit (Qiagen #28704). Final DNA concentrations were determined using an Invitrogen Qubit Fluorometer. We then performed real-time PCR

to determine sample concentrations before sequencing the samples on an Illumina MiSeq using the protocols provided in the MiSeq Reagent Kit v3 (600-cycle) (Illumina MS102–3003). The following primers for sequencing: (i7 L1Hs primer: 5′-GGTACATGTGCACATTGTGC AGGTTAGTTACATATGTATACATGTGC-3′; Read 1 Sequencing Primer: 5′-ATCGATCGCTGCAGGGTA TAGGCGAGGACAACT-3′; Read 2 Sequencing Primer: 5′-GCACATGTATACATATGTAACTAACCTGCACA ATGTGCACATGTACCC-3′).

Resolving non-reference L1Hs insertions from whole genome 3′ targeted L1 capture data in bulk experiment

We customized a pipeline for analyzing the Illumina Miseq sequencing data (Supplementary Figure S3A). In this pipeline, we assigned paired-end alignments using BWA-MEM (59), removed the PCR duplicates using samtools (52), and trimmed the L1Hs sequence, as well as the poly(A) sequence, at both ends of 300 bp paired-end reads to facilitate alignment to the human genome reference sequence. These reads were then converted into fastq files and re-aligned to the reference genome. We discarded reads with low mapping quality (MAPQ < 10), clustered the reads as L1Hs regions, and counted the unique number of supporting shearing points (N_{sp}) for each clustered region. To obtain the values of N_{sp}, we used our high-confidence non-reference Palmer L1Hs insertion set ($n = 203$, described above) as a true positive training set to calculate Youden's J index (J = sensitivity + specificity – 1). This J index was used to set up the threshold of N_{sp} for calling an event in the 3′ targeted L1 capture technology (Supplementary Figure S3B).

Single-cell whole genome amplification cell line preparation

To obtain diploid GM12878 for single-cell studies, cultured GM12878 were washed twice in cold PBS, fixed by resuspending them in ice cold 50% ethanol, and incubated on ice for 20–30 min. Fixed cells were pelleted, resuspended at 1×10^7 cells/ml in cold PBS, and incubated on ice for 45 min with 1ug/ml RNase A (Qiagen) and 500 nM SYTO™ 13 Green Fluorescent Nucleic Acid Stain (Thermo Fisher Scientific, Waltham, MA, USA). Diploid cells were collected in the G1 phase of the cell cycle using a FACS Synergy flow cytometry (iCyt Mission Technology, Champaign, IL, USA). Single cells were isolated from the diploid population using a Cell Microsystems CellRaft and Cell Microsystems CellRaft apparatus according to manufacturer's instructions. WGA was done using an Illustra GenomiPhi V2 DNA Amplification Kit (GE Biosciences) with a modified protocol. Briefly, single cells on raft were incubated for 30 min on ice in 1 ul of 20 mM KOH and 50 mM DTT then frozen at –20°C or –80°C for a minimum of 30 min and maximum of 1 week. After freezing, the cell lysate was incubated at 65°C for 10 min and then on ice for 2 min. After cooling, 9ul of sample buffer was added to lysate and incubated for 10 min on ice. Then, a mix of 9 ul of reaction buffer and 1ul of enzyme was added and incubated for 2 h at 30°C and 10 min at 75°C. Amplified DNA was purified by ethanol precipitation.

WGS for single-cell whole genome amplification DNA

Purified WGA DNA was quantitated on a Qubit (Thermo Fisher Scientific) and 150 ng was used to make standard Illumina sequencing library using NEBNext kit (NEB) and Nextflex adapters (BIOO Scientific, Austin, TX, USA) according to manufacturer's protocols. Libraries were pooled and paired-end sequencing was done with 150 cycle MiSeq Reagent Kit v3 on MiSeq (Illumina, San Diego, CA). Reads were paired, mapped to the genome, and genome coverage of aligned reads was determined using Ginkgo (73) with variable length bins and equal numbers of uniquely mappable reads per bin. Samples with an index of dispersion <0.7 were used for higher depth WGS technology. Higher depth WGS was done at Novogene (Davis, CA) on HiSeq X Ten or the University of Michigan on HiSeq-4000.

We also obtained published single-cell sequencing WGS data from prior studies using multiple displacement amplification (MDA) (41,42) and multiple annealing and looping based amplification cycles (MALBAC) (74). We conducted sequencing quality, genome read alignment, and genome coverage analyses on these data as well as the WGS data from our four single cells. Lorenz curves (Supplementary Figure S4A) were constructed by plotting points with x-value equal to the fraction of the genome with $\leq r$ read depth and y-value equal to the fraction of reads that are in regions of the genome with $\leq r$ read depth. For read depth analysis in four single-cell WGS experiments, we calculated the raw read depth of each category in each single-cell experiment data and normalized them accordingly with the average read depth value of each experiment. The final curves were depicted based on the median of all values from four single-cell experiments data.

Resolving non-reference L1Hs insertions from WGS data in bulk and single-cell WGA experiments

We downloaded WGS data from the Illumina Platinum Genomes (60) to identify germline non-reference L1Hs insertions for NA12878 bulk experiment. We generated single-cell WGS data from WGA DNA as described above. Raw fastq files of NA12878 WGS data in bulk and single-cell WGA experiments were checked for quality using FastQC (75), resulting in the inclusion of four single cells for downstream analysis (scWGS2, scWGS5, scWGS9 and scWGS59). BWA-MEM (59) was used to align these files to reference genome and generate bam files. PCR duplicates were removed using samtools. We then utilized MELT (76) to identify non-reference L1Hs in the resulting BAM files. Since MELT is not specifically designed for single-cell WGS data, we plotted the ROC curves for four single-cell experiment call sets based on the number of split reads reported by MELT and determined a cutoff of the split-read number from MELT equal to six to support an event in our single-cell WGS data (Supplementary Figure S4C). The intersection between PALMER L1Hs insertions and MELT

call sets from four single-cell WGS data was drawn using an UpSet plot (77).

RESULTS

Identifying germline non-reference L1Hs insertions from PacBio data in NA12878

NA12878, a member of the CEPH pedigree number 1463 (78), is one of the most extensively investigated human genetic control samples. Several large-scale human genome projects, including HapMap (79), 1000 Genomes Project (63,80,81), the Human Genome Structural Variation Consortium (48,58,82), Genome In A Bottle (GIAB) (46) and reference genome improvement projects (49) have used a variety of approaches to generate NA12878 sequencing data. Subsequent analyses by these and other projects (45,46,60,83) have resulted in the generation of several SNP, insertion and/or deletion (INDEL), and SV call sets.

Non-reference germline L1Hs insertions have been identified using PCR capture-based approaches (25,39,43,84–86), the analysis of paired-end fosmid sequencing data (10,82), the analysis of population-scale short-read sequence datasets (81), and genome assembly comparisons (87). While powerful, each approach has limitations detecting L1Hs insertions in repetitive genomic sequences. Here, we leveraged NA12878 sequence data, including recently generated PacBio sequencing data from the Genome Institute at Washington University School of Medicine (see Materials and Methods), and developed a computational tool, PALMER, to discover non-reference L1Hs insertions from third-generation PacBio sequencing data (Figure 1).

The primary advancement of PALMER is the use of a pre-masking strategy that identifies and masks existing repetitive elements in the human genome reference sequence in individual sequencing reads, thereby enabling the detection of non-reference L1Hs sequences within the remaining unmasked portions of the genome. Thus, PALMER allows an increased resolution to identify and resolve nested 'repeat in repeat' insertions, which are often missed using short-read sequencing and can be difficult to detect using long-read sequencing approaches (88). PALMER further provides an automated way to detect L1 structural hallmarks, including target site duplications and poly(A) tracts, increasing the confidence that the detected events are *bona fide* L1Hs insertions (see below).

We applied PALMER to 50× coverage NA12878 PacBio subread sequence data and identified 203 candidate non-reference L1Hs insertions. Each putative insertion had, on average, 23 subreads supporting the L1Hs sequence, of which 19/23 (82%) were generated from different molecules (ZMWs). PALMER also allowed the annotation of additional L1Hs structural hallmarks (e.g. a poly(A) tract of ≥20 bp and TSDs of ≥6 bp), and an average of 5/23 supporting reads per insertion were defined as 'high-confidence' due to the presence of these features (Supplementary Table S1). These results remained robust in down-sampling experiments conducted with 30× PacBio sequencing coverage (see Methods, Supplementary Figure S1B).

The analysis of individual PacBio subreads resulted in efficient genome coverage. However, the accuracy of individual base pairs within a raw PacBio sequence read is only ∼80% (89), which led to variances in L1Hs sequence between individual reads. To overcome this sequence limitation, we performed read error correction using the CANU pipeline (54), which exploits multiple overlapping reads to correct individual read sequences. This process permits the generation of a high-quality L1Hs annotated sequence within each supporting subread (i.e. leading to <4.5% error rate). The use of error-corrected individual read sequences also enabled us to identify previously filtered subreads that exhibit L1Hs signals, leading to an overall increase in the average total number ($n = 23$) and high-confidence ($n = 12$) reads supporting each insertion at a given locus when compared to the raw PacBio subreads (Figure 2A). Thus, our final set of 203 candidate PALMER L1Hs insertion calls consist of error-corrected and accurately annotated sequence features. Overall, the non-reference L1Hs sequences detected using the error-corrected PacBio subreads were 98.190% ± 0.019 identical to the sequence of L1.3 (Supplementary Figure S1C), which is consistent with the percent sequence identity of L1.3 when compared to L1Hs sequences present in the reference genome (98.197% ± 0.024).

Additional validation for PacBio calls

To provide additional evidence that the 203 candidate PALMER L1Hs insertion calls represented authentic insertions, we first performed a recurrence plot analysis—a strategy that has long been used to visualize sequence differences (90) and has been more recently used to resolve complex structural rearrangements in the human genome (91). In a recurrence plot analysis, a region of one sequence (X-axis) is compared to another sequence (Y-axis) and small (i.e. 10 bp) segments that are identical between the two sequences are denoted with a plotted point. Thus, a continuous diagonal line comprising multiple points indicates portions of the compared sequences that are identical. By comparison, gaps and shifts from the diagonal denote an insertion or deletion in one sequence relative to the other. For example, Figure 2B depicts a PALMER L1Hs insertion at chr16: 31 950 972, whereas a respective vertical gap in the reference sequence indicates its absence. On the Y-axis, we also included a colored representation indicating the 2235 bp L1Hs insertion in a single PacBio error-corrected sequence (i.e. the sum of a 639 bp 5′ inverted L1 sequence, 1371 bp non-reference L1Hs sequence, 32 bp poly(A) tract, 138 bp 3′ transduction sequence and 55 bp poly(A) tract). Manual inspection of individual NA12878 PacBio supporting reads confirmed the size and breakpoints of the insertion (Supplementary Table S1).

As a control, we next applied PALMER to test whether we could detect polymorphic L1PA2 subfamily members. We reasoned that because the L1PA2 subfamily members are thought to have amplified prior to the divergence of humans and chimpanzees the overwhelming majority of them should not be polymorphic in human populations (92). L1PA2 is present at 4917 locations in the human reference sequence used by the 1000 Genomes Project (hs37d5), compared to 1544 locations for L1Hs, and provides a platform to test our L1Hs insertion predictions. PALMER did not detect any non-reference L1PA2 insertions in NA12878, supporting the assertion that PALMER is able to distin-

guish *bona fide* non-reference L1Hs insertions from other endogenous L1s present in the human reference genome and is consistent with the finding that L1PA2 subfamily members no longer contribute to active retrotransposition in the human genome (8,10)

We next used BLASTn to determine whether our characterized PALMER L1Hs insertions and their associated flanking genomic DNA sequences are present in the GenBank nr/nt database (Figure 2C). Eleven of 203 PALMER L1Hs insertions were previously identified in NA12878 using a fosmid-based approach (see below) (82). An additional 49/203 L1Hs insertions were supported by sequence data from other samples in the nr/nt database, which were primarily derived from bacterial artificial chromosome or fosmid-based DNA sequencing studies (1,58,82). Thus, 60/203 L1Hs insertions identified by PALMER have sequence-level support from previous studies (Figure 2C, Supplementary Tables S1 and S8).

We next examined whether the 203 candidate PALMER L1Hs insertions could be indirectly supported by previously generated fosmid clone data (10,58). Fosmid libraries are constructed to contain ∼40 kb of genomic DNA. The alignment of paired-end Sanger sequences derived from the 5′ and 3′ ends of the fosmid insert to the human genome allows a determination of whether the fosmid insert contains or lacks sequences in the reference sequence. For example, fosmid paired-end reads that map 40 kb apart suggest that the fosmid and reference sequences are generally co-linear. By comparison, discordant fosmid paired-end reads that map 34 kb apart suggest the presence of a ∼6 kb insertion in the fosmid insert relative to the reference sequence (10,58,82). Thus, identifying these apparently discordant fosmid read-pairs allows the identification of sequences present in individual genomes that are absent from the reference sequence. A limitation of fosmid clone datasets is that variation in the distribution of cloned fragment sizes can lead to difficulty in detecting insertions that are less than ∼4 kb in size (93).

We further leveraged the fact that the PacBio reads supporting non-reference L1Hs insertions were often long enough to detect flanking genomic SNPs, which should, in some cases, allow us to infer the haplotype on which individual L1Hs insertions arose. Using the high-quality phased SNP information from GIAB, we interrogated SNPs and assigned the supporting L1Hs reads and individual fosmid clone paired-end reads to specific haplotypes (see Materials and Methods). We then examined whether the decrease in the apparent span of paired-end reads in an individual fosmid (due to the polymorphic L1Hs insertion) correlated with the length of the L1Hs insertion predicted by PALMER and observed a strong linear relationship consistent with the presence of additional sequence in the PacBio reads ($R^2 = 0.706$ and $P < 0.0001$, Student's t-test, Figure 2D). As expected, we did not observe a correlation between individual fosmid clones that lacked an L1Hs insertion on the alternative haplotype with the length of the L1Hs insertion predicted by PALMER ($R^2 = 0.025$). Thus, these experiments allowed an additional level of support for 135/203 of our PALMER L1Hs insertion calls (Supplementary Figure S2C).

To explore whether PALMER L1Hs insertions possess specific sequence signals at their respective 5′ genomic DNA/L1 junction sequences, we conducted a k-mer analysis on short-read Illumina data (Figure 2E) generated from NA12878, the parental samples NA12891 and NA12892, as well as 10x Genomics linked-read data generated from NA12878. The k-mer protocol searches for 26mers that are present in the 5′ genomic DNA/L1Hs junction fragments, but are absent from the hs37d5 reference sequence. These analyses revealed that k-mers diagnostic for 83.7% (170/203) of the candidate PALMER L1Hs insertion 5′ genomic DNA/L1 junction sequences were present in NA12878 Illumina short-read datasets, which contrasts starkly with simulated data as a negative control, where only ∼0.1% (0.2/203) showed sequence level support for the insertion (Figure 2F).

Finally, we exploited CEPH pedigree sequencing data to examine whether the detection of L1Hs 5′ genomic DNA/L1 junction sequence k-mers were detected in a manner consistent with Mendelian inheritance. We utilized available 10x linked-read WGS data, which allows for the assignment of reads to individual haplotypes. In total, 76.4% (155/203) of the candidate PALMER L1Hs insertion 5′ genomic DNA/L1 junction sequences were detected in NA12878 10× linked-read genomic data. Of the 76 germline insertions detected in Illumina short-read data that were inherited from one parent, 73 were supported by 10× Genomics linked-read data with all reads derived from a single haplotype (Figure 2F, Supplementary Table S1). Thus, combined with the evidence above, each of the PALMER L1Hs insertions has support from at least one form of orthogonal evidence, allowing us to conclude that we have generated a high-confidence set of germline non-reference L1Hs insertions from the NA12878 sample.

Assessment of PALMER calls in existing PacBio sequenced resources

We next compared the 203 PALMER L1Hs insertions to a recent study that used PacBio sequencing to generate a large-scale SV call set from fifteen genomes, including NA12878 (49). This study used RepeatMasker (69) to annotate 118 non-reference L1Hs insertions in NA12878, but did not generate detailed annotations of the structural hallmarks associated with these L1Hs insertions. Ninety-four percent (111/118) of the insertions were present in the PALMER L1Hs call set (Figure 3A). Of the seven absent calls, five had fewer than five supporting reads (required by default in PALMER) and were excluded from further analyses (Supplementary Table S2). Manual inspection of the other two calls revealed the presence of 3′ truncated L1 sequences with no poly(A) tract or 3′UTR, suggesting they might not represent canonical retrotransposition events. Notably, 45.3% (92/203) of the non-reference PALMER L1Hs calls were not identified in NA12878 in this prior study (49).

We next extended our comparisons to insertions reported in Audano *et al.* ($n = 12797$) that were not designated as L1Hs insertions. We observed an additional 23 insertion calls in Audano *et al.* that overlapped with the PALMER L1Hs calls, including eleven calls categorized as 'L1P' subfamily members, three calls categorized as 'L1', seven calls categorized as 'Complex', one call categorized as a short

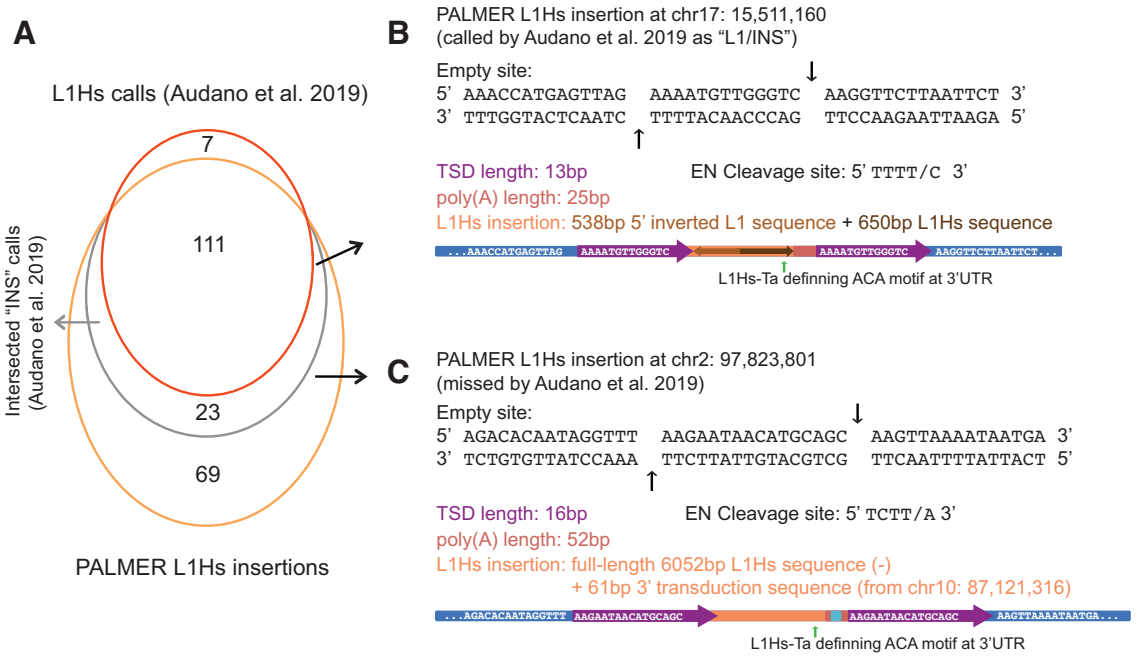

Figure 3. Comparison of PALMER L1Hs insertion calls with a high-quality structural variation set in NA12878. (**A**) Venn diagram of PALMER calls (orange) and L1Hs calls from Audano *et al.* 2019 (red) in NA12878. A subset of 'INS' calls (representing generic insertions not specifically annotated as L1Hs insertions) from Audano et al. 2019 that intersected with Palmer calls is also indicated (grey). (**B**) An example of a PALMER call that was reported by Audano *et al.* 2019 as a generic L1 insertion. We describe the EN Cleavage site sequence, the sequence at the empty site of insertion (bold font) and the TSD motif (purple font/arrow), poly(A) (red font/bar), and the detailed structure of L1Hs insertion (orange font/bar) with 5′ inverted L1 sequence (brown font/arrow) and non-inverted L1Hs insertion sequence (dark brown font/arrow). The green arrow shows that the L1Hs insertion sequence has the L1Hs-Ta defining 'ACA' motif at the 3′UTR region. (**C**) An example of a PALMER call that was missed by Audano et al., containing a 3′ transduction sequence (light blue bar) and other colors as described in (B).

tandem repeat (STR), and one call categorized as 'Not-Masked' (Supplementary Table S1).

For the 11 'L1P' calls, we examined the PacBio sequence data and observed 6/11 insertions with an 'ACG' in the L1 3′UTR, which is diagnostic for the L1Hs pre-Ta subfamily (10,86,94,95). Notably, some L1Hs pre-Ta subfamily members remain retrotransposition-competent in the human population (10,96). We further observed a 'GAG' motif in the 3′UTR region of 3/11 of the calls, which is indicative of L1PA2 or older L1 subfamilies and thus may not represent recent L1Hs insertions (86,94,95). The remaining two 'L1P' calls exhibit all the hallmarks of an L1Hs insertion, but either have an 'AGG' or were too short to have sequence at this diagnostic position, resulting in some ambiguity as to their origins. The three calls categorized as 'L1' have the diagnostic 'ACA' motif in the 3′UTR, but contain 5′ inverted L1 structures that likely led to an incorrect prediction of L1 insertion length, suggesting they were mis-annotated by RepeatMasker (Figure 3B). Moreover, of the eight calls annotated as 'Complex' or 'NotMasked', four contained the diagnostic 'ACA' motif in the L1 3′UTR, one contained an 'ACG' in the L1 3′UTR, two contained 'GAG' and 'GGG', and one was too short to exhibit the motif. Further, 7/8 contained L1-mediated 3′ transductions. In sum, by annotating the specific characteristics of L1Hs insertions, PALMER provides additional evidence that at least 14/23 Audano *et al.* insertions that intersected with the PALMER L1Hs insertion calls are likely true L1Hs insertion events, although there is some uncertainty about the remaining nine calls.

After considering the above comparisons, we still had 69/203 PALMER-specific L1Hs non-reference insertion calls in NA12878 (Supplementary Table S2). Thirty-four of 69 insertions were reported in other (non-NA12878) PacBio data, suggesting they are *bona fide* non-reference L1 insertions that were filtered from the NA12878 assembly (49). Relative to the other samples studied by Audano *et al.*, NA12878 had lower sequencing data quality, which may explain why they were not identified. The Audano et al. assembly approach is likely more sensitive to the length and quality of the input data than PALMER, which has been designed to specifically work with individual raw PacBio subreads. The other 35 PALMER specific L1Hs non-reference insertions were not identified in any of the PacBio samples reported in Audano et al. and include nine full-length L1Hs insertions (Figure 3C as an example, Supplementary Table S1).

Finally, comparisons of the PALMER L1Hs calls to a PacBio GIAB 'generic insertion' call set (46) revealed 107 overlapping calls (Supplementary Figure S5); however, 96 (47.3%) of 203 PALMER calls were absent from the >5000 reported GIAB insertions. Thus, PALMER can effectively re-annotate generic insertions sequences as *bona fide* retrotransposition events and allows the identification of >40% more L1Hs insertions than reported in other long-read NA12878 call sets.

Novel calls from PacBio data nested in existing reference LINE repeats

WGS is one of the most prevalent approaches used to discover inter-individual human genetic variation (97). We used the Mobile Element Locator Tool (MELT) (63,76) to identify L1Hs insertions in high coverage (50×) WGS data generated by the Illumina Platinum Genome Project (60). Of note, MELT was the primary tool to discover mobile genetic elements in the 1000 Genomes Project Phase 3 data set; however, NA12878 was not included in that analysis due to the inclusion of its parental samples and the focus of the project on unrelated individuals. MELT identified 164 L1Hs insertions (Figure 4, Supplementary Table S3) and 113 of these overlapped with the 203 PALMER L1Hs insertion calls. The remaining 44.3% (90/203) PALMER L1Hs insertions were absent from the MELT call set.

For the MELT-only calls, 8/33 exhibited some sequence support in the PacBio data and could represent *bona fide* L1Hs insertions; however, they were not called by PALMER because they were supported by fewer than five supporting subreads. An additional 3/33 MELT-only calls exhibited some concordance between the MELT-predicted size and the insert size of overlapping fosmid clone pairs, indicating that they also may be true L1Hs insertions. In the above 11/33 calls, two also were reported by Audano *et al.* (Supplementary Figure S6, Table S3).

We next examined the remaining 22/33 MELT-only calls and did not observe evidence of an L1Hs insertion in any overlapping PacBio reads or from Illumina short-reads in the respective regions (Supplementary Table S4). Manual inspection of these MELT-only calls revealed that they were located in A/T-rich or other repetitive regions and only exhibited non-L1Hs specific repetitive sequence content in the Illumina short-read data. Indeed, in more than half (13/22) of these predicted MELT-only insertions we observed a deletion or duplication CNV at the breakpoint, which likely led to the consideration of a distal L1 sequence in the reference genome as a putative variant. To further verify that we were capturing both haplotypes at MELT-only putative L1Hs locus, we conducted additional SNP-haplotype analyses on PacBio subreads in the regions of these 22 events (similar to our fosmid-based validation approach, see Methods, Supplementary Figure S2A). Using phased SNPs, we successfully assigned PacBio error-corrected reads to both haplotypes in 21/22 MELT-only putative L1Hs locations, but still did not observe L1Hs signals on either allele. Thus, these 22 MELT-only putative L1Hs call likely represent false positives that arise due to the limitations of using short-read technology to detect L1Hs insertions in copy number variable or repetitive regions (Supplementary Table S3).

We observed 39.4% (80/203) PALMER L1Hs insertions within pre-existing LINE sequences in the reference sequence (Figure 4C). This proportion increases to 81.1% (73/90) when considering L1Hs insertions only found by PALMER in PacBio data and decreases to 15.2% (5/33) when considering the WGS MELT-only calls (Figure 4C). We did not observe similar distributions in other categories of repeats, including SINEs, LTR retrotransposons, DNA transposons and tandem repeats. Moreover, the PacBio sequencing approach is more successful at identifying L1Hs insertions in regions of the genome that are not easily interrogated using short-read sequencing technologies (Figure 4D, Supplementary Table S1; 55 L1Hs in PacBio versus 35 in Illumina short-read data) (81). Finally, 74.4% (67/90) of PacBio-only calls are located in intergenic regions, whereas only 48.5% (16/33) of the L1Hs calls from standard Illumina WGS are in intergenic regions. As expected, none of the insertions are in coding exonic regions (Figure 4D, Supplementary Tables S1 and S3), suggesting that such insertions are likely subject to negative selection and are therefore underrepresented in humans (24,98,99). Thus, the above results highlight the importance of accurately assessing repetitive genomic regions with increased LINE occupancy for human-specific retrotransposon insertions (100) and demonstrate the limitations of standard WGS approaches in identifying L1Hs insertions within LINE-rich regions (48).

We further investigated whether the PALMER L1Hs insertions exhibited any enrichment or depletion around recombination hotspots. Although absent from the reference genome, many of the PALMER L1Hs insertions likely arose during the last ∼2 million years and thus were subject to forces of selection (86). Nonetheless, we examined our insertions for differences in recombination rate and did not observe any significant preference (female: $P = 0.293$ and male: $P = 0.055$, two-tailed Student's *t*-test, Supplementary Figure S7) consistent with cell-based studies of engineered retrotransposons that indicate that the presence of an endonuclease cleavage consensus and association with the DNA replication machinery are principle determinants that dictate the genomic distribution of new L1Hs insertions (24).

Characterizing germline non-reference L1Hs insertions from PacBio data in NA12878

We next sought to determine whether there were specific features of L1Hs insertions that could complicate their detection in PacBio versus short-read sequencing data. We did not observe significant differences between the 90 L1Hs PacBio-only calls with respect to 5′ truncation points or insertion lengths (Figure 4E), the distribution and prevalence of inverted L1 sequences at the 5′ end of L1Hs insertions (Figure 4F), the frequency or length of L1-mediated 3′ transduction sequences, or the sizes of the TSDs (Figure 4G, Supplementary Table S1; range 6–38 bp). Furthermore, 25.6% (52/203) of PALMER L1Hs insertions were full-length (Figure 4A) and 18/52 were only identified from PacBio sequences (Supplementary Table S1). Thus, these data suggest that the number of full-length, potentially active, L1Hs retrotransposons are underrepresented in existing human genomes.

L1Hs 5′ genomic DNA/L1 junction sequence k-mer analysis reveals PacBio L1Hs calls are polymorphic in populations of the 1000 Genomes Project

Phase 3 of the 1000 Genomes Project used Illumina 4–6× WGS of 2504 unrelated samples (63,81), which allowed us

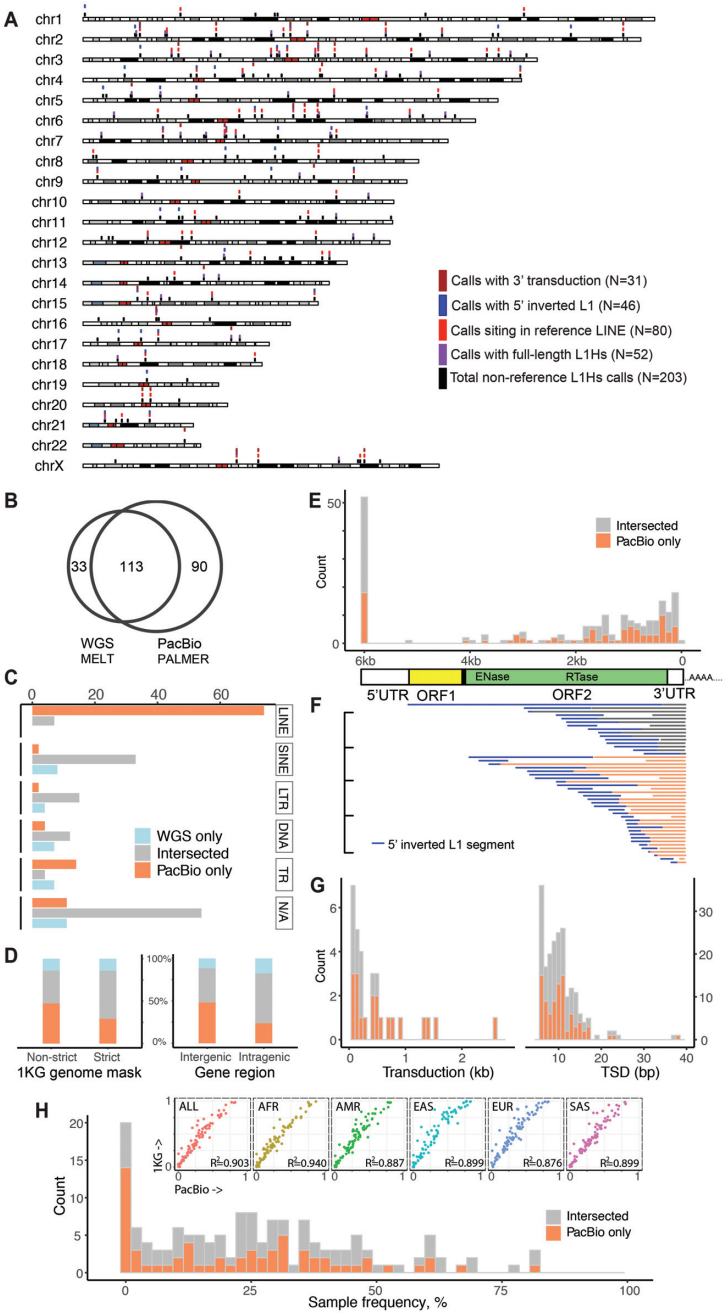

Figure 4. Characteristics of the 203 germline non-reference L1Hs insertions from NA12878 PacBio data. (**A**) Ideogram of PacBio call set. Four types of insertions are highlighted: insertions with 3′ transduction sequence (brown), insertions with 5′ inverted L1 sequence (dark blue), insertions located in reference LINE regions (red), and full-length events (purple). The black bar delineates all non-reference calls. (**B**) Venn diagram of non-reference L1Hs insertion sets of NA12878 from Illumina standard WGS by MELT and PacBio by PALMER. (**C**) Number of calls locating in different RepeatMasker categories based on (B). We show the calls in three categories: WGS-only calls (light blue), PacBio-only calls (orange) and calls intersecting in the two call sets (gray). We delineate reference repeat information into six categories: LINE (e.g. L1, L2), SINE (e.g. *Alu*, MIR), LTR (e.g. ERV, ERVK), DNA (DNA transposons), TR (tandem repeats, e.g. simple repeat, satellite, low complexity region), N/A (regions with no reference repeats annotated). (**D**) Number of calls located in genomic regions of different short-read accessibility (non-strict: less accessibility, and strict: more accessibility) on the left panel and different gene regions (intergenic and intragenic) on the right panel, on the scale of portion in overall two call sets. The figure legend is the same as in (C). (**E**) Distribution of truncated positions within the L1Hs sequence of PacBio calls with L1Hs structure annotated below. Bars filled with orange depicted PacBio-only calls, and gray bars depicted PacBio calls intersected with WGS call set. Lower panel demonstrated the detailed structure of a full-length L1Hs, including a 5′UTR, ORF1 (yellow), ORF2 containing endonuclease (EN) and reverse transcriptase (RT) domains (green), 3′UTR and a poly(A) tract. (**F**) Distribution of 5′ inverted L1 sequence possibly related to twin priming mechanism. The dark blue bar demonstrates the 5′ inverted L1 segments. (**G**) Length distributions of 3′ transduction sequence of PacBio calls (left) and TSD motif in the 5′ and 3′ flanking region of PacBio calls (right). (**H**) Histogram of sample frequency in all 1000 Genomes phase 3 samples of PacBio calls based on L1Hs 5′ genomic DNA/L1 junction sequence *k*-mer assessment. Upper panels show scatter plots of sample frequency by *k*-mer calculation (X-axis) versus sample frequency based on 1000 Genomes L1Hs call set (Y-axis), for calls intersected in PacBio call set and 1000 Genomes L1Hs call set across five super-populations: All (red), AFR (Africa, dark yellow), AMR (Americas, green), EAS (East Asia, sky blue), EUR (Europe, blue) and SAS (South Asia, pink). E, F, G and H share the same figure key.

to assess whether the PALMER L1Hs insertions identified in NA12878 were polymorphic with respect to presence or absence in humans. To query whether the PALMER L1Hs insertions were present in the 1000 Genomes Project Phase 3 samples (see Materials and Methods), we utilized the *k*-mers identified in our pedigree validation experiments (Figure 2E). We first compared the PALMER L1Hs insertions to the 1000 Genomes Project Phase 3 L1Hs call set (63). Approximately 42% (87/203) of the L1Hs insertions were present in both call sets and we observed a consistent correlation between the L1Hs frequencies reported in the 1000 Genomes Project Phase 3 call set and our *k*-mer analysis (Figure 4H; overall $R^2 = 0.903$) across all five super-populations (Africa [AFR], East Asia [EAS], Europe [EUR], South Asia [SAS] and the Americas [AMR]). An additional 57/90 PacBio-only L1Hs calls could be interrogated and 46 were identified in more than 5% of the phase 3 samples, indicating that they have been systematically missed in these population-scale data sets. Thus, these data demonstrate that we can assess whether the PALMER L1Hs insertions are present in existing short-read sequencing data derived from large cohorts of diverse individuals.

Whole genome 3′ targeted L1 capture technology in NA12878

Several studies have used targeted ligation-mediated PCR strategies to identify L1Hs insertions. Briefly, oligonucleotide primers that exploit the diagnostic 'ACA' trinucleotide sequence in the 3′UTR of L1Hs insertions and a ligated adapter sequence allow for the generation of complex PCR amplicons enriched for L1Hs/3′ flanking genomic DNA sequence junctions (39,86). Illumina short-read sequencing then is used to de-convolute the PCR amplicons and a variety of computational pipelines have been implemented to map individual L1Hs insertions to a reference sequence. Here, we sought to examine the effectiveness of ligation-mediated PCR targeted capture sequencing to identify L1Hs insertions identified in the NA12878 sample.

We used a previously reported approach to selectively amplify L1Hs/3′ flanking genomic DNA sequence junctions (39,86) from NA12878 genomic DNA and characterized them using overlapping 300 bp paired-end Illumina DNA sequencing. A customized computational pipeline (Supplementary Figure S3A) then was used to identify 128 non-reference germline L1Hs insertions; 85.9% (110/128) of these calls were also identified in the PALMER L1Hs insertion set (Figure 5A, Supplementary Table S5). We manually inspected the eighteen 3′ L1 capture calls, which were not called by PALMER, and 5/18 overlapped with potentially true-positive MELT calls (as described above). An additional L1 capture call (1/18) was also reported by Audano *et al.* and, after manual inspection, is likely to be a true positive. We applied the same SNP-haplotype analysis (see Methods, Supplementary Figure S2A) on PacBio error-corrected reads to the other 12/18 3′ targeted L1 capture technology calls and were able to assign reads to each haplotype. In each case, we did not detect evidence for an L1Hs insertion in these regions. In sum, comparisons between the PALMER L1Hs, Illumina WGS-MELT and the 3′ ligation-mediated PCR targeted capture call sets revealed that each technology identified a different constellation of L1Hs insertions (Supplementary Figure S6, Table S5); the PALMER L1Hs and 3′ targeted L1 capture approaches were more successful than Illumina WGS-MELT at identifying L1Hs elements in repetitive genomic regions (Figure 5B).

Amplification bias leads to variation of L1Hs calling in WGS among multiple single-cell experiments

We next explored the efficiency of detecting L1Hs insertions in single-cell WGA-derived DNA sequencing data. We reasoned that the high-confidence PALMER L1Hs insertions in NA12878 would serve as a valuable reference to assess both false positive and false negative L1Hs calls derived from single-cell WGA experiments. We isolated single diploid NA12878 cells, conducted MDA whole genome amplification, and generated standard Illumina libraries for subsequent paired-end DNA sequencing (see Methods). The resultant MDA-WGS exhibited a genome-wide average read depth of at least 20× and achieved an average coverage of 85 ± 2.8% of the genome at ≥1× read depth and 66 ± 8.6% at ≥5× read depth across all single cells, which is higher than the 2% MDA locus dropout rate reported in previous studies (41,42,101). We also analyzed our single-cell data for sequencing quality, genome read alignment, and genome coverage (Supplementary Table S6) and found the data to be comparable with those reported in previous MDA or MALBAC studies (41,42,74). However, we still observed genome coverage bias from both our single-cell WGA data as well as in data from these previous MDA and MALBAC studies (Supplementary Figure S4A and B), as reflected by the variance in each experiment in the cumulative fraction of the genome covered by specific read coverages. This result is consistent with previously reported amplification bias (102).

We next used MELT to detect non-reference L1Hs insertion calls in single-cell WGS data derived from four NA12878 single cells; however, we acknowledge the caveat that MELT was not specifically designed for single-cell analysis. The resulting number of L1Hs insertions identified in the four independent single-cell experiments ranged from 31 to 63 calls (Figure 5C, Supplementary Table S7). Almost all of the MELT L1Hs calls that were identified in multiple single-cell experiments (64/65) were present in the PALMER L1Hs insertion set; only one MELT L1Hs call identified in three separate single-cell experiments was absent in the PALMER L1Hs insertion set, suggesting that it could represent a false negative PALMER call. Moreover, MELT calls found only in individual single-cell experiments likely represent false positives (Figure 5C). Intriguingly, 51.1% (71/139) of calls only detected by PALMER (i.e. they are absent from all four single-cell experiments) are nested within endogenous LINEs within the human reference sequence, which is consistent with our inability to detect L1Hs calls within endogenous LINE sequences in bulk WGS data (Figures 4C and 5D). We compared the genotype information from MELT in both bulk and single-cell data. A portion of heterozygous events in bulk experiments were inconsistently reported as homozygotes in single-cell WGS experiments (11/56 in single single-cell experiment

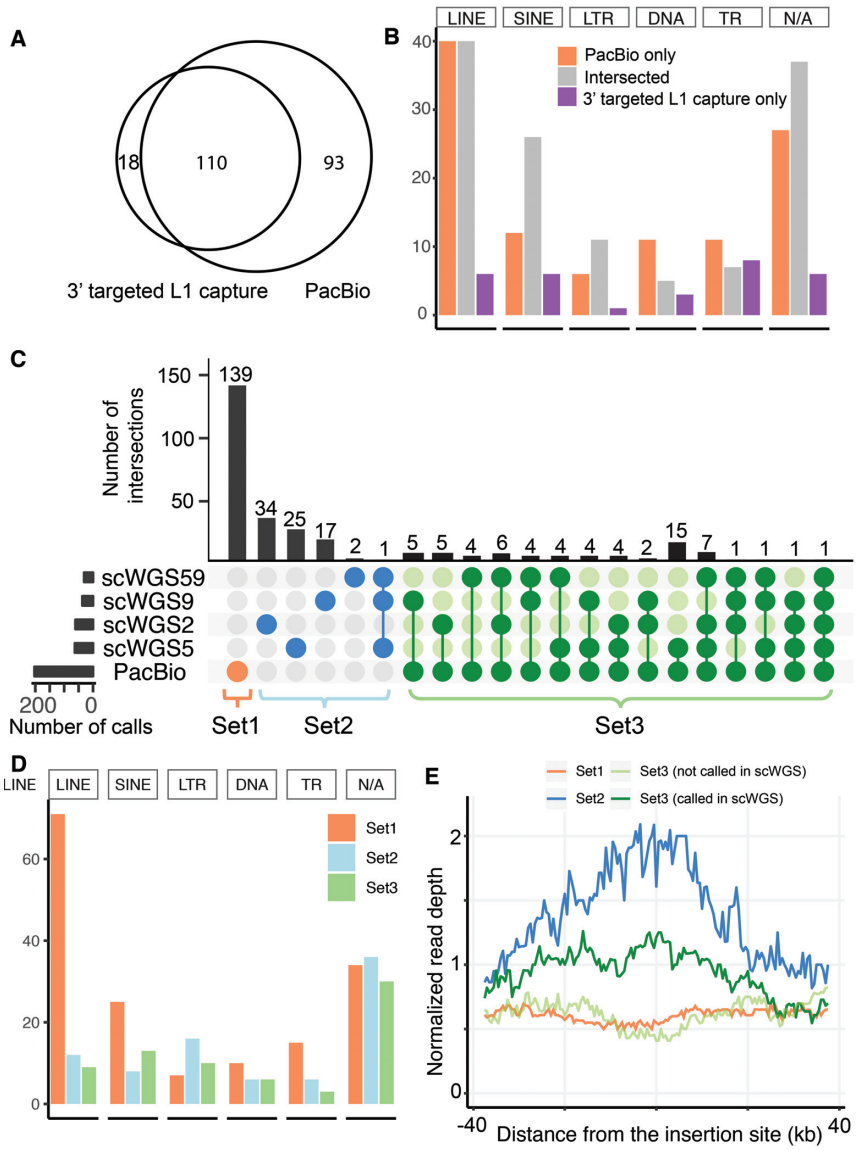

Figure 5. L1Hs insertion detection using 3′ targeted L1 capture in bulk experiments and WGS in single-cell experiments. (**A**) Venn diagram of non-reference L1Hs insertion sets of NA12878 from 3′ targeted L1 capture technology and PacBio by PALMER. (**B**) Number of calls located in different RepeatMasker categories based on (A) in three categories: 3′ targeted L1 capture technique-only calls (purple), PacBio-only calls (orange) and calls intersecting in two call sets (grey). (**C**) Upset plot of the intersection between PacBio call set and MELT call sets from four single-cell WGS data (batch id: scWGS59, scWGS9, scWGS2 and scWGS5). We delineate the calls into three sets: set1 (orange bracket and orange dots, the events in the PacBio call set but not called in any single-cell experiments), set2 (light blue bracket and blue dots, the events from the single-cell call sets but not in the PacBio call set), set3 (green bracket). In set3, we have two sub-sets: dark green dots show the intersection of the single-cell call sets and PacBio call set, and light green dots show the calls were absent in a certain single-cell experiment but called by the others and intersected with PacBio call set. (**D**) Number of calls located in different RepeatMasker categories based on sets defined in (C). We delineate the calls into three categories: set1 (orange), set2 (light blue) and set3 (green). (**E**) Read depth analysis for four single-cell WGS experiments. Categories of sets are based on (C). The curves of normalized read depth value in the ± 37.5 kb flanking regions of insertion sites are shown.

and 5/56 in multiple single-cell experiment, Supplementary Figure S4D), which may be indicative of allelic dropout in the single-cell experiments.

Finally, we investigated the read coverage of L1Hs calls in the four single-cell experiments. We normalized the read coverage values based on the average genome coverage, which should result in an average read depth of '1' at any given position in the genome. We observed that insertion calls that overlapped between the single-cell and PALMER L1Hs call sets had a value of '1', indicating that the DNA amplified as expected in these regions. However, for L1Hs insertions only detected in individual single cells, the sequences surrounding the putative L1Hs insertions exhibited higher read depth than the average (Figure 5E), indicating they are likely false positive calls induced by over-amplification. By comparison, the PALMER L1Hs insertions not called in single-cell data exhibited a lower read depth at the insertion loci relative to background signals, suggesting they were missed due to a lack of supporting reads. We further observed that these variances in coverage

tend to converge to the expected background levels when the flanking region extended to ± 20 kb or more, which suggests that they are in fact reflecting amplicon-level (∼10–50 kb for MDA) coverage biases (102) though they could also be due to the repetitive nature of the L1Hs sequence and/or the repetitive nature of the sequences at the L1Hs insertion site. Overall, these data support that the prevalence of over/under-amplification bias at a genome-wide level in single-cell experiments can affect the identification of endogenous L1Hs insertions.

DISCUSSION

We developed an approach to comprehensively detect a specific class of mobile genetic elements, L1Hs retrotransposons, in long-read DNA sequences. Notably, the proteins encoded by L1Hs retrotransposons continue to drive the mobilization of other retroelements, including *Alu* short interspersed elements (SINEs), SVAs (SINE/VNTR/*Alu*), and U6 snRNA (28,103–107). Collectively, these elements may contribute to the ongoing mutagenesis of the human genome.

Our approach makes use of PacBio long-read sequence data, which has allowed us to better characterize variation in repetitive regions of the genome that are often refractory to approaches using short-read sequencing technologies. Our implementation of a targeted pre-masking approach also provides advantages compared to assembly-based variant calling using long reads and is likely extendable to other emerging single-molecule technologies such as Oxford Nanopore Technologies. Together, these advancements have enabled us to identify previously overlooked L1Hs insertions, particularly those embedded in existing repetitive sequences. For example, we observed an L1Hs PacBio-only call (chr6: 32 613 219) (Figure 2C, Supplementary Figure S8) inserted into a reference L1MC1 segment within an intron of gene *HLA-DQA1*, which is located near the human major histocompatibility complex (MHC). Further investigation found that this insertion is actually present on an alternative haplotype of the human reference sequence; as such, it likely is a true polymorphic insertion missed by other studies that exclude alternative haplotypes from their analysis (108). Intriguingly, we observed a L1 deletion polymorphism just 3′ downstream of this insertion in other reference haplotypes, providing an example of how retroelements may potentially contribute to the diversity of the HLA complex (109).

We assessed the extent to which L1Hs insertions may be systematically missed in the myriad of genomes that are currently being sequenced with short-read technology. We compared our results to MELT, a short-read method that has seen broad use in the 1000 Genomes Project, as well as other large initiatives, and demonstrated that up to 45% of L1Hs insertions identified by PALMER were often missed in short-read data because they were embedded within complex regions containing pre-existing repetitive DNAs. This loss in L1Hs calling efficacy was consistent across our comparisons with a short-read based targeted approach, suggesting that limitations in L1Hs identification are currently driven more by limits in the sequencing technology (i.e. read lengths) rather than the underlying analytical strategies for identifying mobile elements from whole genome sequence data. Intriguingly, we identified 18 novel full-length L1Hs sequences in the NA12878 genome; functional assays (8,10,15) are needed to examine whether any of these elements are retrotransposition-competent and have the potential to contribute to inter- or intra-individual human genetic variation. The above being stated, it still is likely that PALMER is missing an unknown number of non-reference L1Hs insertions due to the conservative parameters that were used in our analysis (i.e. the number of supporting reads and identity between target site duplications).

Our focus has been on a single well-characterized sample, NA12878, as it has long been established as a gold standard for human genetic variant discovery and assessment. However, our analysis of the 2504 low-coverage, short-read 1000 Genomes Project data suggests that many of our newly identified L1Hs are, in fact, prevalent in human populations at a >1% allele frequency. Indeed, a recent deep human genome study (48) compared call sets of SVs from short and long-read technologies and showed a sensitivity of only ∼50% from Illumina short reads for detecting insertions >50bp. This lower sensitivity has impacted our ability to accurately assess the true extent of L1 activity in mammalian genomes. Indeed, the actual rate of L1 mobilization in the germline likely requires refinement. This observation is likewise true for insertions in somatic tissue where, for example, estimated rates of endogenous L1Hs mobilization in neuronal cells currently varies from 0.04 to as high as 13.7 per neuron (33,38–43). As long-read datasets continue to become available, we expect that PALMER will be a useful tool for identifying additional L1Hs elements that have been previously overlooked and could serve as source elements for potentially pathogenic insertions. Furthermore, we envision that future refinements of PALMER may be able to detect other classes of mobile genetic element insertions, including SINEs and/or SVA elements, in human, non-human primate, and other mammalian genomes.

DATA AVAILABILITY

The Illumina sequencing data generated in this study have been deposited into the NCBI Sequence Read Archive under accession PRJNA531679. The PALMER program is available at https://github.com/mills-lab/PALMER. All command lines in this project are available at https://github.com/mills-lab/PALMER_Pipelines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mike McConnell for advice on the isolation and amplification of single diploid cells and members of the NIMH Brain Somatic Mosaicism Network for critical feedback. The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM12878.
Author contributions: R.E.M. and W.Z. conceived the project. W.Z. developed the PALMER software. W.Z. and

Y.W. performed computational analysis. S.B.E. and D.A.F. constructed sequencing libraries and performed molecular analysis. All authors guided the data analysis strategy. R.E.M., W.Z., J.V.M. and J.M.K. wrote the manuscript. All authors edited the manuscript. All authors read and approved the final manuscript.

FUNDING

R.E.M., J.V.M., K.K. and J.M.K. were supported by a grant from the National Institutes of Health [MH106892]; D.A.F. was supported in part by National Institute of Health training grant T32 [HG000040]. Funding for open access charge: National Institutes of Health [MH106892].

Conflict of interest statement. J.V.M. is an inventor on patent US6150160, is a paid consultant for Gilead Sciences and a privately held company founded by Flagship Pioneering, and is on the American Society of Human Genetics Board of Directors. The other authors do not have competing interests.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Smit,A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, **9**, 657–663.
3. Grimaldi,G., Skowronski,J. and Singer,M.F. (1984) Defining the beginning and end of KpnI family segments. *EMBO J.*, **3**, 1753–1759.
4. Kazazian,H.H. Jr. and Moran,J.V. (1998) The impact of L1 retrotransposons on the human genome. *Nat. Genet.*, **19**, 19–24.
5. 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.
6. Larson,P.A., Moldovan,J.B., Jasti,N., Kidd,J.M., Beck,C.R. and Moran,J.V. (2018) Spliced integrated retrotransposed element (SpIRE) formation in the human genome. *PLoS Biol.*, **16**, e2003067.
7. Kazazian,H.H. Jr and Moran,J.V. (2017) Mobile DNA in health and disease. *N. Engl. J. Med.*, **377**, 361–370.
8. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 5280–5285.
9. Sassaman,D.M., Dombroski,B.A., Moran,J.V., Kimberland,M.L., Naas,T.P., DeBerardinis,R.J., Gabriel,A., Swergold,G.D. and Kazazian,H.H. Jr. (1997) Many human L1 elements are capable of retrotransposition. *Nat. Genet.*, **16**, 37–43.
10. Beck,C.R., Collier,P., Macfarlane,C., Malig,M., Kidd,J.M., Eichler,E.E., Badge,R.M. and Moran,J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, **141**, 1159–1170.
11. 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.
12. Scott,A.F., Schmeckpeper,B.J., Abdelrazik,M., Comey,C.T., O'Hara,B., Rossiter,J.P., Cooley,T., Heath,P., Smith,K.D. and Margolet,L. (1987) Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics*, **1**, 113–125.
13. Dombroski,B.A., Mathias,S.L., Nanthakumar,E., Scott,A.F. and Kazazian,H.H. Jr. (1991) Isolation of an active human transposable element. *Science*, **254**, 1805–1808.
14. Mills,R.E., Bennett,E.A., Iskow,R.C. and Devine,S.E. (2007) Which transposable elements are active in the human genome? *Trends Genet.*, **23**, 183–191.
15. Moran,J.V., Holmes,S.E., Naas,T.P., DeBerardinis,R.J., Boeke,J.D. and Kazazian,H.H. Jr. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917–927.
16. Moran,J.V., DeBerardinis,R.J. and Kazazian,H.H. Jr. (1999) Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530–1534.
17. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
18. 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.
19. 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.
20. Doucet,A.J., Wilusz,J.E., Miyoshi,T., Liu,Y. and Moran,J.V. (2015) A 3′ poly(A) tract is required for LINE-1 retrotransposition. *Mol. Cell*, **60**, 728–741.
21. Richardson,S.R., Doucet,A.J., Kopera,H.C., Moldovan,J.B., Garcia-Perez,J.L. and Moran,J.V. (2015) The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol. Spectr.*, **3**, doi:10.1128/microbiolspec.MDNA3-0061-2014.
22. Goodier,J.L. (2016) Restricting retrotransposons: a review. *Mob. DNA*, **7**, 16.
23. Cost,G.J. and Boeke,J.D. (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37**, 18081–18093.
24. Flasch,D.A., Macia,A., Sanchez,L., Ljungman,M., Heras,S.R., Garcia-Perez,J.L., Wilson,T.E. and Moran,J.V. (2019) Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. *Cell*, **177**, 837–851.
25. Beck,C.R., Garcia-Perez,J.L., Badge,R.M. and Moran,J.V. (2011) LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.*, **12**, 187–215.
26. 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.
27. Holmes,S.E., Dombroski,B.A., Krebs,C.M., Boehm,C.D. and Kazazian,H.H. Jr. (1994) A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.*, **7**, 143–148.
28. Moldovan,J.B., Wang,Y., Shuman,S., Mills,R.E. and Moran,J.V. (2019) RNA ligation precedes the retrotransposition of U6/LINE-1 chimeric RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 20612–20622.
29. Hancks,D.C. and Kazazian,H.H. Jr. (2016) Roles for retrotransposon insertions in human disease. *Mob DNA*, **7**, 9.
30. Scott,E.C. and Devine,S.E. (2017) The role of somatic L1 retrotransposition in human cancers. *Viruses*, **9**, E131.
31. Muotri,A.R., Chu,V.T., Marchetto,M.C., Deng,W., Moran,J.V. and Gage,F.H. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, **435**, 903–910.
32. Coufal,N.G., Garcia-Perez,J.L., Peng,G.E., Yeo,G.W., Mu,Y., Lovci,M.T., Morell,M., O'Shea,K.S., Moran,J.V. and Gage,F.H. (2009) L1 retrotransposition in human neural progenitor cells. *Nature*, **460**, 1127–1131.
33. Upton,K.R., Gerhardt,D.J., Jesuadian,J.S., Richardson,S.R., Sanchez-Luque,F.J., Bodea,G.O., Ewing,A.D., Salvador-Palomeque,C., van der Knaap,M.S., Brennan,P.M. *et al.* (2015) Ubiquitous L1 mosaicism in hippocampal neurons. *Cell*, **161**, 228–239.
34. Evrony,G.D., Lee,E., Park,P.J. and Walsh,C.A. (2016) Resolving rates of mutation in the brain using single-neuron genomics. *Elife*, **5**, e12966.
35. 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.
36. Harbom,L.J., Michel,N. and McConnell,M.J. (2018) Single-cell analysis of diversity in human stem cell-derived neurons. *Cell Tissue Res.*, **371**, 171–179.

37. 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.
38. McConnell,M.J., Moran,J.V., Abyzov,A., Akbarian,S., Bae,T., Cortes-Ciriano,I., Erwin,J.A., Fasching,L., Flasch,D.A., Freed,D. *et al.* (2017) Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science*, **356**, eaal1641.
39. Iskow,R.C., McCabe,M.T., Mills,R.E., Torene,S., Pittard,W.S., Neuwald,A.F., Van Meir,E.G., Vertino,P.M. and Devine,S.E. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, **141**, 1253–1261.
40. Erwin,J.A., Paquola,A.C., Singer,T., Gallina,I., Novotny,M., Quayle,C., Bedrosian,T.A., Alves,F.I., Butcher,C.R. and Herdy,J.R. (2016) L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.*, **19**, 1583–1591.
41. Evrony,G.D., Lee,E., Mehta,B.K., Benjamini,Y., Johnson,R.M., Cai,X., Yang,L., Haseley,P., Lehmann,H.S. and Park,P.J. (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, **85**, 49–59.
42. Evrony,G.D., Cai,X., Lee,E., Hills,L.B., Elhosary,P.C., Lehmann,H.S., Parker,J., Atabay,K.D., Gilmore,E.C. and Poduri,A. (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, **151**, 483–496.
43. Faulkner,G.J. and Garcia-Perez,J.L. (2017) L1 mosaicism in mammals: extent, effects, and evolution. *Trends Genet*, **33**, 802–816.
44. Chin,C.S., Alexander,D.H., Marks,P., Klammer,A.A., Drake,J., Heiner,C., Clum,A., Copeland,A., Huddleston,J., Eichler,E.E. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.
45. Jain,M., Koren,S., Miga,K.H., Quick,J., Rand,A.C., Sasani,T.A., Tyson,J.R., Beggs,A.D., Dilthey,A.T., Fiddes,I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.
46. Zook,J.M., Catoe,D., McDaniel,J., Vang,L., Spies,N., Sidow,A., Weng,Z., Liu,Y., Mason,C.E., Alexander,N. *et al.* (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.
47. Sedlazeck,F.J., Rescheneder,P., Smolka,M., Fang,H., Nattestad,M., von Haeseler,A. and Schatz,M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**, 461–468.
48. Chaisson,M.J., Sanders,A.D., Zhao,X., Malhotra,A., Porubsky,D., Rausch,T., Gardner,E.J., Rodriguez,O.L., Guo,L. and Collins,R.L. (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.
49. Audano,P.A., Sulovari,A., Graves-Lindsay,T.A., Cantsilieris,S., Sorensen,M., Welch,A.E., Dougherty,M.L., Nelson,B.J., Shah,A., Dutcher,S.K. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell*, **176**, 663–675.
50. Parikh,H., Mohiyuddin,M., Lam,H.Y., Iyer,H., Chen,D., Pratt,M., Bartha,G., Spies,N., Losert,W., Zook,J.M. *et al.* (2016) svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, **17**, 64.
51. Jurka,J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, **8**, 333–337.
52. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
53. Xu,Q., Schlabach,M.R., Hannon,G.J. and Elledge,S.J. (2009) Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 2289–2294.
54. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
55. Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.
56. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
57. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
58. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
59. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 16 March 2013,preprint: not peer reviewed.
60. Eberle,M.A., Fritzilas,E., Krusche,P., Kallberg,M., Moore,B.L., Bekritsky,M.A., Iqbal,Z., Chuang,H.Y., Humphray,S.J., Halpern,A.L. *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27**, 157–164.
61. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
62. Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
63. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Fritz,M.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
64. Wang,J., Song,L., Grover,D., Azrak,S., Batzer,M.A. and Liang,P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.*, **27**, 323–329.
65. Mohiyuddin,M., Mu,J.C., Li,J., Bani Asadi,N., Gerstein,M.B., Abyzov,A., Wong,W.H. and Lam,H.Y. (2015) MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, **31**, 2741–2744.
66. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
67. Pendleton,A.L., Shen,F., Taravella,A.M., Emery,S., Veeramah,K.R., Boyko,A.R. and Kidd,J.M. (2018) Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.*, **16**, 64.
68. Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W. and Eichler,E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
69. Smit,A., Hubley,R. and Green,P. (2013) RepeatMasker Open-4.0.
70. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
71. Kong,A., Thorleifsson,G., Gudbjartsson,D.F., Masson,G., Sigurdsson,A., Jonasdottir,A., Walters,G.B., Jonasdottir,A., Gylfason,A. and Kristinsson,K.T. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, **467**, 1099.
72. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A. and Hsu,F. (2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
73. Garvin,T., Aboukhalil,R., Kendall,J., Baslan,T., Atwal,G.S., Hicks,J., Wigler,M. and Schatz,M.C. (2015) Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods*, **12**, 1058–1060.
74. Zong,C., Lu,S., Chapman,A.R. and Xie,X.S. (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, **338**, 1622–1626.
75. Andrews,S. (2010) *Babraham Bioinformatics*. Babraham Institute, Cambridge.
76. Gardner,E.J., Lam,V.K., Harris,D.N., Chuang,N.T., Scott,E.C., Pittard,W.S., Mills,R.E., Genomes Project,C. and Devine,S.E. (2017) The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res*, **27**, 1916–1929.
77. Lex,A., Gehlenborg,N., Strobelt,H., Vuillemot,R. and Pfister,H. (2014) UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–1992.
78. Dausset,J., Cann,H., Cohen,D., Lathrop,M., Lalouel,J.M. and White,R. (1990) Centre d'etude du polymorphisme humain

(CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575–577.
79. International HapMap, C. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
80. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K. and Cheetham,R.K. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
81. Consortium,G.P., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
82. Kidd,J.M., Graves,T., Newman,T.L., Fulton,R., Hayden,H.S., Malig,M., Kallicki,J., Kaul,R., Wilson,R.K. and Eichler,E.E. (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, **143**, 837–847.
83. Zook,J., McDaniel,J., Parikh,H., Heaton,H., Irvine,S.A., Trigg,L., Truty,R., McLean,C.Y., De La Vega,F.M. and Salit,M. (2018) Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. bioRxiv doi: https://doi.org/10.1101/281006, 13 March 2018, preprint: not peer reviewed.
84. Ovchinnikov,I., Troxel,A.B. and Swergold,G.D. (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.*, **11**, 2050–2058.
85. Huang,C.R., Schneider,A.M., Lu,Y., Niranjan,T., Shen,P., Robinson,M.A., Steranka,J.P., Valle,D., Civin,C.I., Wang,T. *et al.* (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell*, **141**, 1171–1182.
86. Badge,R.M., Alisch,R.S. and Moran,J.V. (2003) ATLAS: a system to selectively identify human-specific L1 insertions. *Am. J. Hum. Genet.*, **72**, 823–838.
87. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
88. Churakov,G., Grundmann,N., Kuritzin,A., Brosius,J., Makalowski,W. and Schmitz,J. (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. *BMC Evol. Biol.*, **10**, 376.
89. Koren,S., Schatz,M.C., Walenz,B.P., Martin,J., Howard,J.T., Ganapathy,G., Wang,Z., Rasko,D.A., McCombie,W.R., Jarvis,E.D. *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.*, **30**, 693–700.
90. Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.*, **16**, 1–11.
91. Zhao,X., Weber,A.M. and Mills,R.E. (2017) A recurrence based approach for validating structural variation using long-read sequencing technology. *GigaScience*, **6**, 129.
92. Ovchinnikov,I., Rubin,A. and Swergold,G.D. (2002) Tracing the LINEs of human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 10522–10527.
93. Donahue,W.F. and Ebling,H.M. (2007) Fosmid libraries for genomic structural variation detection. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0520s54.
94. Boissinot,S., Chevret,P. and Furano,A.V. (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.*, **17**, 915–928.
95. Xing,J., Zhang,Y., Han,K., Salem,A.H., Sen,S.K., Huff,C.D., Zhou,Q., Kirkness,E.F., Levy,S. and Batzer,M.A. (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.*, **19**, 1516–1526.
96. Kazazian,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
97. Lappalainen,T., Scott,A.J., Brandt,M. and Hall,I.M. (2019) Genomic analysis in the age of human genome sequencing. *Cell*, **177**, 70–84.
98. Soares,M.L., Edwards,C.A., Dearden,F.L., Ferron,S.R., Curran,S., Corish,J., Rancourt,R., Allen,S., Charalambous,M., Ferguson-Smith,M.A. *et al.* (2018) Targeted deletion of a 170 kb cluster of LINE1 repeats: implications for regional control. *Genome Res.*, **28**, 345–356.
99. Sultana,T., van Essen,D., Siol,O., Bailly-Bechet,M., Philippe,C., El Aabidine,A.Z., Pioger,L., Nigumann,P., Saccani,S. and Andrau,J.-C. (2019) The landscape of L1 retrotransposons in the human genome is shaped by pre-insertion sequence biases and post-insertion selection. *Mol. Cell*, **74**, 555570.
100. Slotkin,R.K. (2018) The case for not masking away repetitive DNA. *Mobile DNA*, **9**, 15.
101. Hou,Y., Song,L., Zhu,P., Zhang,B., Tao,Y., Xu,X., Li,F., Wu,K., Liang,J., Shao,D. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, **148**, 873–885.
102. Zhang,C.Z., Adalsteinsson,V.A., Francis,J., Cornils,H., Jung,J., Maire,C., Ligon,K.L., Meyerson,M. and Love,J.C. (2015) Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.*, **6**, 6822.
103. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.*, **35**, 41–48.
104. Raiz,J., Damert,A., Chira,S., Held,U., Klawitter,S., Hamdorf,M., Lower,J., Stratling,W.H., Lower,R. and Schumann,G.G. (2012) The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.*, **40**, 1666–1683.
105. Hancks,D.C., Goodier,J.L., Mandal,P.K., Cheung,L.E. and Kazazian,H.H. Jr. (2011) Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.*, **20**, 3386–3400.
106. 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.
107. 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.
108. Horton,R., Gibson,R., Coggill,P., Miretti,M., Allcock,R.J., Almeida,J., Forbes,S., Gilbert,J.G., Halls,K., Harrow,J.L. *et al.* (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics*, **60**, 1–18.
109. Andersson,G., Svensson,A.C., Setterblad,N. and Rask,L. (1998) Retroelements in the human MHC class II region. *Trends Genet.*, **14**, 109–114.