PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: McKerrow W, Tang Z, Steranka JP, Payer LM, Boeke JD, Keefe D, Fenyö D, Burns KH, Liu C. 2019 Human transposon insertion profiling by sequencing (TIPseq) to map LINE-1 insertions in single cells. *Phil. Trans. R. Soc. B* **375**: 20190335. http://dx.doi.org/10.1098/rstb.2019.0335

Accepted: 14 October 2019

One contribution of 15 to a discussion meeting issue 'Crossroads between transposons and gene regulation'.

Subject Areas:
genetics, genomics, molecular biology, biotechnology

Keywords:
mobile genetic element, retrotransposon, TIPseq, whole-genome amplification, somatic mosaicism, tumour heterogeneity

Authors for correspondence:
David Fenyö
e-mail: david@fenyolab.org
Kathleen H. Burns
e-mail: kburns@jhmi.edu
Chunhong Liu
e-mail: cliu87@jhmi.edu

†Present address: Memorial Sloan Kettering Cancer Center, New York, NY, USA.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4796277.

THE ROYAL SOCIETY PUBLISHING

Human transposon insertion profiling by sequencing (TIPseq) to map LINE-1 insertions in single cells

Wilson McKerrow[1], Zuojian Tang[1,†], Jared P. Steranka[2], Lindsay M. Payer[2], Jef D. Boeke[1], David Keefe[3,4], David Fenyö[1], Kathleen H. Burns[2,5,6,7] and Chunhong Liu[2]

[1]Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, USA
[2]Department of Pathology, Johns Hopkins University School of Medicine, 733N Broadway, Baltimore, MD 21205, USA
[3]Department of Obstetrics and Gynecology, and [4]Department of Cell Biology, New York University Langone School of Medicine, 462 First Avenue, New York, NY 10016, USA
[5]McKusick-Nathans Institute of Genetic Medicine, and [6]High Throughput (HiT) Biology Center, Johns Hopkins University School of Medicine, 733N Broadway, Baltimore, MD 21205, USA
[7]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, 401N Broadway, Baltimore, MD 21231, USA

KHB, 0000-0003-1620-3761; CL, 0000-0002-0134-2745

Long interspersed element-1 (LINE-1, L1) sequences, which comprise about 17% of human genome, are the product of one of the most active types of mobile DNAs in modern humans. LINE-1 insertion alleles can cause inherited and de novo genetic diseases, and LINE-1-encoded proteins are highly expressed in some cancers. Genome-wide LINE-1 mapping in single cells could be useful for defining somatic and germline retrotransposition rates, and for enabling studies to characterize tumour heterogeneity, relate insertions to transcriptional and epigenetic effects at the cellular level, or describe cellular phylogenies in development. Our laboratories have reported a genome-wide LINE-1 insertion site mapping method for bulk DNA, named transposon insertion profiling by sequencing (TIPseq). There have been significant barriers applying LINE-1 mapping to single cells, owing to the chimeric artefacts and features of repetitive sequences. Here, we optimize a modified TIPseq protocol and show its utility for LINE-1 mapping in single lymphoblastoid cells. Results from single-cell TIPseq experiments compare well to known LINE-1 insertions found by whole-genome sequencing and TIPseq on bulk DNA. Among the several approaches we tested, whole-genome amplification by multiple displacement amplification followed by restriction enzyme digestion, vectorette ligation and LINE-1-targeted PCR had the best assay performance.

This article is part of a discussion meeting issue 'Crossroads between transposons and gene regulation'.

1. Introduction

A large proportion of the human genome is composed of interspersed repeat sequences, and a small subset of these are actively propagating as mobile genetic elements [1,2]. Long interspersed element-1 (LINE-1, L1) is one of the most active and abundant mobile DNAs in the human genome, and LINE-1 sequences comprise about 17% of the genome [1]. Most LINE-1 are old, fixed elements (i.e. homozygous insertion alleles in any individual human genome). However, a small subset of full-length LINE-1 insertions, members of the Ta subfamily of *Homo sapiens*-specific LINE-1 (L1Hs), are the evolutionarily youngest elements and have significant potential to retrotranspose through target primed reverse transcription (TPRT) [3–11].

These active LINE-1 are not only responsible for their retrotransposition, but also encode proteins that retrotranspose other repeat sequences in *trans*, namely, short interspersed elements (SINEs) and SVAs (SINE/VNTR/*Alu*) [12,13]. L1Hs elements, *Alu*Y SINE sequences and SVA insertions propagated by LINE-1 machinery together represent a significant source of structural variation in human populations [14–21].

There are at least 124 reported disease alleles caused by LINE-1-mediated retrotransposition events in the germline or early development [22–24]. Emerging data show that LINE-1 proteins are highly expressed in cancers and that somatic LINE-1 retrotransposition is commonplace in many cancers, indicating that LINE-1 expression and retrotransposition contribute to the genome instability in these malignancies [25–33]. LINE-1 insertions are frequent structural variants segregating in human populations, and many are not incorporated in the human reference genome assembly [34]. To understand genetic variation caused by these sequences and to find de novo insertions that distinguish cancer genomes from normal, many efforts have been made to profile LINE-1 insertions genome-wide using targeted or whole-genome sequencing [11,16,31,35–47]. Our laboratories contributed an approach we termed transposon insertion profiling by sequencing (TIPseq) [31,41,42,47]. This method is based on an insertion site-specific amplification, and covers the 3′ end of the L1Hs and downstream (3′), adjacent unique genomic DNA.

Genome-wide LINE-1 profiling in single cells has applications in many areas of research. As a marker of cellular lineage, it could be used to understand patterns of growth during development, somatic mosaicism in various tissues and clonal evolution in cancers. Several specific features of mobile element insertions make them useful as phylogenetic makers. First, they are directional, meaning that there is no ambiguity in distinguishing the pre-existing allele and the derivative allele. The 'empty' or pre-insertion allele is the antecedent allele, and the LINE-1 insertion allele arises later. Second, they are 'homoplasy-free', meaning that LINE-1 insertions are each unique [48–50]. In addition to its exact location, a LINE-1 insertion can be distinguished from another allele by its length, structure and target site duplication. Thus, finding the same insertion in two cells is strong evidence of a common origin or identity by division. We also have a lot to learn about retrotransposition, including tissue tropisms for the activity of specific source elements, whether LINE-1 activity is continual or episodic, and contributions of genotype and environment to retrotransposition activity. It seems that the activity of 'hot' LINE-1 loci is not constant throughout oncogenesis, but rather, apparently time-limited activities of different LINE-1 elements can cause new insertions in distinct tumour subclones [35,41]. Because of this, LINE-1 somatic insertions could be useful lineage markers in cancer heterogeneity and evolution. Somatically acquired LINE-1 insertions have been observed in neuronal cells, though other tissues and many non-malignant diseases have been less well studied [51–58].

There have been significant barriers to the development of LINE-1 mapping for single cells, owing to the prevalence of chimeric artefacts and the highly repetitive nature of LINE-1 sequences [56]. Whereas we previously have described TIPseq protocols requiring micrograms of genome DNA [47], here we demonstrate that TIPseq can be applied to whole-genome amplified DNA from little starting material—the genomic DNA content of a single cell. With data analysis using a modified version of TIPseqHunter2 software [31], the approach provides investigators with an economical and rigorous method for LINE-1 insertion site mapping in single cells.

2. Methods

(a) Cell line

The GM12878 lymphoblastoid cell line, which is one of the European HapMap cell lines [59], was obtained from Coriell Institute for Medical Research. GM12878 cells were cultured in RPMI 1640 medium supplemented with 2 mM L-glutamine (Quality Biological, cat no. 112-025-101), 15% FBS (Corning, cat no. 35-010-CV), 100 units ml^{-1} of penicillin and 100 µg ml^{-1} of streptomycin (Thermo Fisher, cat no. 15140122).

(b) Single-cell sorting

Single-cell suspensions of the cultured cells were washed with PBS, resuspended in the buffer (PBS + 1% BSA) and passed through 70 µm cell strainers (BD Pharmingen, cat no. 352235). Then the cells were stained with 1 mg ml^{-1} propidium iodide (PI) solution. Live single cells were sorted into PCR tubes. Doublet discrimination gates, including SSC-Height versus SSC-Width gate, FSC-Height versus FSC-Width gate, and PI gate, were used to ensure only one live cell was sorted per well (electronic supplementary material, figure S1). Cell sorting was performed in the Flow Cytometry and Cell Sorting Core Facility at Johns Hopkins Bloomberg School of Public Health.

(c) Whole-genome amplification

Single-cell whole-genome amplification (WGA) was performed using multiple displacement amplification (MDA) or multiple annealing and looping-based amplification cycles (MALBAC) methods. MDA was performed using REPLI-g Single Cell Kit (QIAGEN, cat no. 150343). MALBAC was performed using MALBAC® Single Cell WGA Kit (Yikon Genomics, cat no. YK001B). The sequences of L1 primer and MALBAC-L1 primer that were added during WGA are 5′-AGA TAT ACC TAA TGC TAG ATG ACA CA-3′ and 5′-GTG AGT GAT GGT TGA GGT CTT GTG GAG AGA TAT ACC TAA TGC TAG ATG ACA CA-3′, respectively.

(d) Quality control

The quality of the WGA for the samples was evaluated using qPCR. Twelve pairs of primers were designed for qPCR to amplify regions downstream of the 3′ end of fixed L1Hs insertions on different chromosomes (electronic supplementary material, table S1). Primers were synthesized by Integrated DNA Technologies (IDT). A sample containing 100 sorted cells was used as a positive control. qPCR was performed using SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad, cat no. 1725271) and run on Bio-Rad MyIQ™ Single-Color Real-Time PCR Detection System. Fold change was calculated based on the Ct value and normalized with the positive control samples.

(e) Vectorette PCR

Whole-genome amplified DNA samples were digested with *Ase*I, *Bsp*HI, *Bst*YI, *Hind*III, *Nco*I and *Pst*I (New England Biolabs). Alternatively, the whole-genome amplified DNA was end-repaired by 5′ phosphorylated and 3′ dA tailing using the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs, cat no. E7546S). A pair of vectorette oligonucleotides (synthesized by IDT) corresponding to each restriction enzyme or T tail were annealed to form vectorette adaptors with the sticky end created. See sequences reported in [47]. Then the digested or repaired

amplified genomic DNA were ligated with the vectorette adaptors using T4 DNA ligase (New England Biolabs, M0202S) at 4°C overnight. After ligation, PCR was performed with the L1 primer (5′- AGA TAT ACC TAA TGC TAG ATG ACA CA -3′) and the Vectorette Primer (5′- CTC TCC CTT CTC GGA TCT TAA -3′) using ExTaq (Takara, cat no. RR006A) with a touchdown programme (95°C 5 min; 95°C 1 min, 72°C 1 min, 72°C 5 min, 5 cycles; 95°C 1 min, 68°C 1 min, 72°C 5 min, 5 cycles; 95°C 45 s, 64°C 1 min, 72°C 5 min, 15 cycles; 95°C 45 s, 60°C 1 min, 72°C 5 min, 15 cycles; 72°C 15 min; 4°C hold).

(f) Next-generation sequencing

About 2 µg of the vectorette PCR products were sheared to fragments of around 300 bp. Sequencing libraries were prepared using the KAPA HTP DNA Library preparation Kit (Roche, cat no. KK8234). Libraries were sequenced on an Illumina HiSeq 4000 with paired-end 150 bp reads. Sequencing was performed in NYU Langone's Genome Technology Center.

(g) Data analysis using TIPseqHunter2 pipeline

Reference and non-reference L1Hs insertions were identified using a modified version of the TIPseqHunter2 pipeline [31]. Reads are trimmed using Trimmomatic [60] and then aligned to both to the hg38 reference human genome and to the consensus L1Hs sequence using bowtie2 [61]. Regions of hg38 that are continuously covered by aligned reads are identified as potential L1 primer amplification sites. Regions that do not have any reads that align to both the L1Hs consensus and to hg38 are excluded. TIPseqHunter2 then uses five features to separate true L1Hs insertions from noise:

1. length of the amplified region (from putative L1Hs insertion to vectorette ligation site),
2. mean coverage across the amplified region,
3. mean number of alignment mismatches per read,
4. presence of an intact polyA tail,
5. number of split reads that align partly to the amplified region and partly to the L1Hs consensus sequence.

A support vector machine (SVM) model with radial basis kernel is used to separate true insertions from false positive insertions. Two hundred fixed L1Hs insertions make up the positive training set [31]. Potential amplification regions must have a read for which one end aligns discordantly to hg38 and the other end aligns to the L1Hs consensus sequence at the 5′ end of the L1 primer binding site to be considered a candidate insertion. Regions that do not have such a read make up the negative training set. The SVM model then returns a probability that a candidate insertion is a true insertion. Insertions in pericentric heterochromatin are filtered out. TIPseqHunter2 accuracy is measured by sensitivity (i.e. true-positive rate or recall), defined to be the fraction of true insertions that are given a probability over some threshold (in this paper, we use 0.9) and positive predictive value (PPV; i.e. precision), defined to be the fraction of potential insertions with probability exceeding the threshold that are true insertions. If TP is the number of true-positive calls, FP is the number of false-positive calls and FN is the number of false-negative calls then

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

(h) Validation of unknown insertions

Validation of the unknown insertions was done by spanning PCR and 3′ junction PCR. Spanning PCR is designed to amplify the entirety of an insertion, with primers flanking the insertion site. 3′ junction PCR is designed to amplify the 3′ of LINE-1 insertion and the downstream flanking sequence using L1 primer in the 3′ of LINE-1 and the other primer in unique flanking sequence.

3. Results

(a) Single-cell sorting and whole-genome amplification

The single-cell TIPseq procedure consists of five steps: cell sorting, WGA, a quality control check, vectorette PCR for L1Hs insertion site amplification and next-generation sequencing (figure 1).

Live GM12878 lymphoblastoid cells were sorted into PCR tubes or 96-well plates with one single cell per well. Multiple gates, including FSC-Height versus FSC-Width, SSC-Height versus SSC-Width, and PI, were used to make sure that only single live cells were sorted, and the dead cells and doublets were excluded (electronic supplementary material, figure S1).

Then single-cell genomic DNA was amplified by WGA. We tested two methods for WGA: MDA [62] and MALBAC [63]. For each method, we also tested whether amplification was improved by the addition of L1Hs-specific primers, which is called 'L1 primer'. This L1 primer, ending with base pairs 'ACA', is designed to perfectly bind elements in the Ta subfamily, which is the youngest and most active subfamily of L1Hs, and causes most de novo retrotransposition in humans [8–11].

In the MDA WGA method, we tried the standard procedure with random hexamer primers in the reaction (MDA-R, 'R' to connote random primers); we also modified the procedure by adding additional L1 primer in the reaction (MDA-RL, 'RL' to connote random primers plus L1 primer). In both methods, the resulting amplicons ranged in size from 3 to 50 kb, and the yield was about 40 µg for MDA samples.

For the MALBAC WGA method, we tested the standard procedure (MALBAC-R), as well as the addition of the MALBAC-L1 primer in the pre-amplification step and L1 primer in the amplification step (MALBAC-RL). The size of amplicons ranged from 200 bp to 3 kb for MALBAC-R samples, and 100 bp to 3 kb for MALBAC-RL samples. The yield was 0.5–1 µg for MALBAC-R samples and 0.1–0.2 µg for MALBAC-RL samples.

(b) Quality control of whole-genome amplification

Quality control of the single-cell WGA was performed by qPCR. We designed 12 pairs of qPCR primers in unique DNA sequence located at the regions 3′ of homozygous (fixed present) L1Hs insertions on 12 different chromosomes (electronic supplementary material, table S1). qPCR was performed using these 12 pairs of primers to evaluate the performance of the WGA in these regions. Although results were quantitative, amplification of these regions was essentially binary, with either low Ct values and robust amplification or negligible amplification, and so data are shown as numbers of loci amplified here.

A sample containing 100 sorted cells was used as a positive control, which showed 12/12 loci amplified. In a representative experiment (figure 2), 2 of 5 MDA-R and 3 of 5 MDA-RL samples showed effective amplification of all

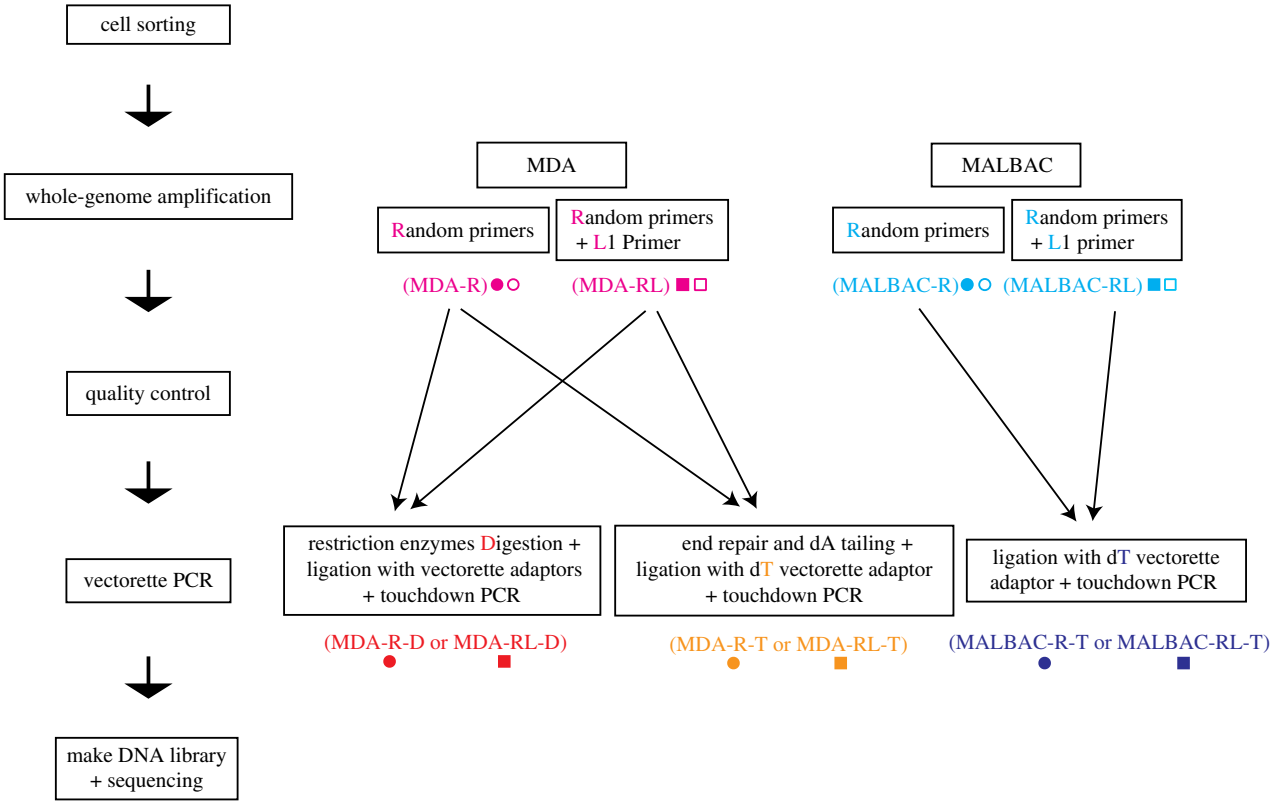

Figure 1. Overview of single-cell TIPseq workflows. The single-cell TIPseq procedure consists of five steps: cell sorting, WGA, a quality control check, vectorette PCR for L1Hs insertion site amplification and next-generation sequencing. Pink, MDA WGA with or without L1 primer (MDA); light blue, MALBAC WGA with or without L1 primer (MALBAC); red, MDA WGA followed by restriction enzyme digestion and ligation with vectorette adaptors (MDA-D); orange, MDA WGA followed by end repair, dA tailing and ligation with dT vectorette adaptor (MDA-T); dark blue, MALBAC WGA followed by ligation with dT vectorette adaptor (MALBAC-T). Circles represent WGA using random hexamers only (R); squares represent WGA using random hexamers and the L1 primer (RL).

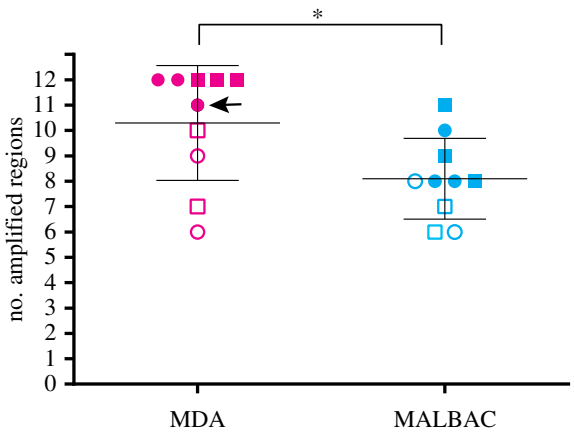

Figure 2. Quality control check following WGA. Pink, MDA WGA with or without L1 primer (MDA); light blue, MALBAC WGA with or without L1 primer (MALBAC). Circles, WGA using random hexamers only (R); squares, WGA using random hexamers and L1 primer (RL). Filled shapes, samples picked for following TIPseq; open shapes, samples that were not selected for the following TIPseq. An arrows indicates an MDA sample that is included in the following vectorette PCR and next-generation sequencing, but has only 11/12 regions amplified by qPCR, while other MDA samples have 12/12 regions amplified. *t-test, $p = 0.0279$.

12 regions (12/12). None of MALBAC-R or MALBAC-RL samples had uniform amplification of all 12 tested regions. Overall, MDA amplifies more regions of interest than MALBAC (t-test p-value 0.0279). These results suggested that MDA-based WGA had superior yield when compared with MALBAC-based WGA.

We picked three high-performing cell samples from each condition (MDA-R, MDA-RL, MALBAC-R or MALBAC-RL) for the subsequent vectorette PCR and next-generation sequencing in order to compare their performance. In the selected MDA samples, 5 out of 6 amplified all 12 regions, and the remaining 1 amplified 11 regions. In the selected MALBAC samples, none showed good recovery at all 12 loci.

(c) L1Hs insertion site amplification and next-generation sequencing

After WGA and quality control, vectorette PCR was performed in those samples with consistently high qPCR amplicon yields. Vectorette PCR is a one-sided, ligation-mediated amplification reaction. For the MDA-R and MDA-RL samples, we tried two different template preparation procedures in advance of the amplification itself.

One preparation consisted of digestion of the amplified genomic DNA by restriction enzymes; ligation of the digested DNA fragments to the vectorette adapters that match the sticky end of the restriction enzymes; and touchdown PCR using the ligated products as the template, and using the L1 primer as forward primer and the vectorette primer as reverse primer. This is analogous to our usual vectorette PCR template preparation from bulk DNA samples. We term these reactions as MDA-R-D or MDA-RL-D, adding the 'D' to connote restriction digest.

The alternative procedure consisted of repairing the ends of whole-genome amplified genomic DNA to add 5′ phosphorylation and 3′ single nucleotide dA tails using a

polymerase lacking 5′–3′ proofreading activity, then ligating the resulting DNA fragments to the vectorette adapter with a complementary dT overhang, followed by the touchdown PCR as above. These samples are called MDA-R-T or MDA-RL-T, the 'T' to connote tailing.

For the MALBAC-R and MALBAC-RL samples, because the last amplification step of MALBAC method adds a dA-tail to the 3′ end of the amplified fragments by *Taq* polymerase, and the sizes of the amplified fragments were only around 100 bp–3 kb, we skipped the digestion or repair steps, and directly proceeded to the single 'sticky' base A/T-mediated ligation of amplified fragments to the vectorette adapter. This was then followed by the touchdown PCR, the same as described above. These samples are referred to here as MALBAC-R-T and MALBAC-RL-T.

Then, for all samples, the vectorette PCR-amplified DNA was sheared to fragments at the size of around 300 bp. Then, DNA sequencing libraries were prepared and sequenced on an Illumina HiSeq 4000 with paired-end 150 bp reads.

(d) Single-cell TIPseq results agree with known GM12878 L1Hs insertions

Before calculating the sensitivity, we developed a list of known L1Hs insertions in GM12878 cell line as 'gold standard' as described here. We began with an encompassing list of known L1Hs insertions, obtained by combining reference L1Hs loci with polymorphic insertions known to be present in GM12878 [20]. Then we applied two exclusionary criteria. First, since our L1 primer (ending with 'ACA') provides some specificity for the Ta subfamily of L1Hs insertions, which are the most active and youngest L1Hs in the human genome, we excluded the pre-Ta subfamily of L1Hs insertions, which have 'ACG' for the corresponding sequence. Using the whole-genome sequencing dataset of GM12878 (SRR622457 sequenced by the 1000 genomes project) [64], we removed insertions from our list that do not have exact matches to the L1 primer binding region. Second, we required that at least one read pair from SRR622457 align with one mate within 500 bp of the 3′ junction and the other covering the primer binding region with no mismatches. This excludes variant reference L1Hs that are missing from GM12878 and any L1 with observed sequence divergences within the primer binding region. After these exclusions, we were left with a list of 468 'gold standard' L1Hs insertions expected in GM12878: 373 reference loci, 14 homozygous non-reference loci and 81 heterozygous non-reference loci. The 'gold standard' list can be found in electronic supplementary material, table S2.

TIPseqHunter2 pipeline was used to identify the L1Hs insertions in all the samples [31]. Insertions within 100 bp were merged as the exact location of an insertion can be hard to pinpoint, especially when L1Hs inserts into an A-rich region that blends with the polyA tail. Using a strict (svm probability > 0.9) cut-off and comparing to our gold standard list, we find single-cell TIPseq sensitivities (the fraction of gold standard insertions that are identified) as high as 90%, with 5/6 MDA-D experiments achieving sensitivity greater than 80% (figure 3*a*). Sensitivity is about 10% lower when only non-reference heterozygous insertions are considered (electronic supplementary material, figure S2). Thirteen of 18 single-cell experiments have PPVs (the fraction of insertions calls that are in the gold standard list) in the 70–80% range (figure 3*a*), with three samples exceeding 80% PPV and two falling below 70%. Because TIPseqHunter2 provides a probability score for each potential insertion, the cut-off can be made more or less stringent, improving either sensitivity or PPV at the expense of the other. The effect of this trade-off is shown in figure 3*c*,*e*,*g*. In our experiments, a strict cut-off (svm probability > 0.9) was used for TIPseqHunter2. The TIPseqHunter2 probability scores for each potential insertion in each experiment can be found in electronic supplementary material, table S2.

(e) Unknown insertions filtering and validation

We next investigated those insertion calls made in an MDA-D experiment (our favoured protocol, see below), but that did not appear in our gold standard list. About one-third (n = 139) are real L1Hs that appear in euL1db [65], but were excluded from our gold standard list. These could be members of the pre-Ta subfamily of L1Hs in GM12878 that were non-specifically amplified (owing to only 1 bp mismatch with the L1 primer). Another sizeable fraction (n = 197) are in or within 25 bp of an L1PA2, L1PA3 or L1PA4 element. These calls are likely mis-amplification from L1 primer binding to highly similar sequences in evolutionarily older LINE-1 elements. Of the remaining 160 calls, 54 are in segmental duplications with greater than 95% similarity, and 79 are within 10 kb of known L1Hs element. The latter seem to reflect intramolecular rearrangements that result in chimeric MDA products [66], and ultimately false-positive calls proximal to true LINE-1 insertions. We incorporated more stringent criteria based on these observations, filtering out calls that are (i) within 25 bp of an L1PA2, L1PA3 or L1PA4 element, (ii) in a segmental duplication, or (iii) within 10 kb of a known L1Hs, and repeated our sensitivity and PPV analysis. These filters removed 34 gold standard L1Hs elements (30 reference, 4 heterozygous non-reference), while increasing PPV above 80% in 13/18 single-cell experiments (figure 3*b*). After applying this filtering across the five MDA-D samples passing quality control, on average 297 (80%) of the gold standard reference, 59 (73%) of the gold standard heterozygous non-reference and 12 (86%) of the gold standard homozygous non-reference insertions were detected. Across the 6 MDA-D experiments, 27 unknown predicted, but likely false, LINE-1 calls passed filtering. Only 2 of these 27 are predicted in bulk. Effects of this filtering strategy on PPV versus sensitivity plots are shown in figure 3*d*,*f*,*h*.

We next performed PCR validations on 4 of these 27 unknown insertions. We tried to amplify the LINE-1 insertion by spanning PCR with primers flanking the insertion sites. Owing to potential difficulty amplifying large LINE-1 of unknown size, we also attempted 3′ junction PCRs pairing the L1 primer in the 3′ of LINE-1 with a primer in unique, downstream flanking sequence. We recovered amplicons from bulk DNA and all whole-genome amplified, single-cell samples from the 3′ junction PCRs, but not the corresponding spanning PCRs. Sanger sequencing of the 3′ junction PCR products indicated that these are all non-specific PCR amplifications aligning to the wrong location of the genome. We also tested three candidates that would be filtered out as they are within 10 kb of known L1Hs, and we found all of them are artefacts caused by WGA. Consistent with our assumption that GM12878 is genomically stable with well-characterized LINE-1 variants,

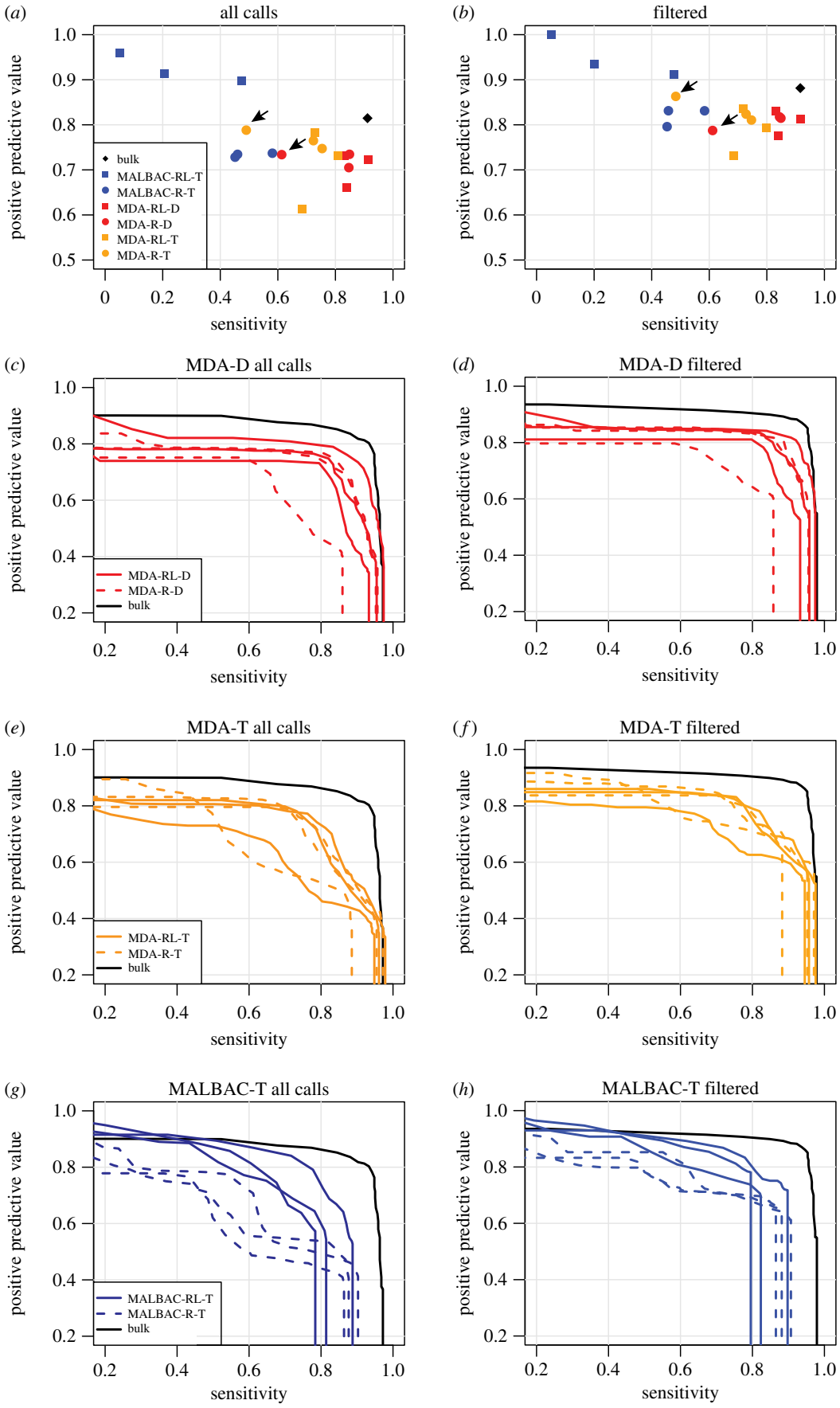

Figure 3. Comparison to 'gold standard' known insertions. Sensitivity and PPV when comparing single-cell TIPseq to a set of known GM12878 insertions with intact primer binding sites. (*a*) Sensitivity and PPV for all experiments, including all insertions and using a probability cut-off of 0.9. (*b*) As (*a*), but including only insertions that pass our three filters. Diamond, bulk DNA TIPseq; circle, WGA using random hexamers only (R); square, whole-genome amplification using random hexamers and L1 primer (RL). Arrows indicate the MDA sample included in the vectorette PCR (both MDA-D and MDA-T) and next-generation sequencing stages that had less than perfect QC (corresponds to the same sample indicated in figure 2). (*c–h*) Sensitivity–PPV curves for each single-cell TIPseq experiment as the probability cut-off is varied from 0 to 1. Black lines, bulk DNA TIPseq; red, MDA WGA followed by restriction enzyme digestion and ligation with vectorette adaptors (MDA-D); orange, MDA WGA followed by end repair, dA tailing and ligation with dT vectorette adaptor (MDA-T); dark blue, MALBAC WGA followed by ligation with dT vectorette adaptor (MALBAC-T).

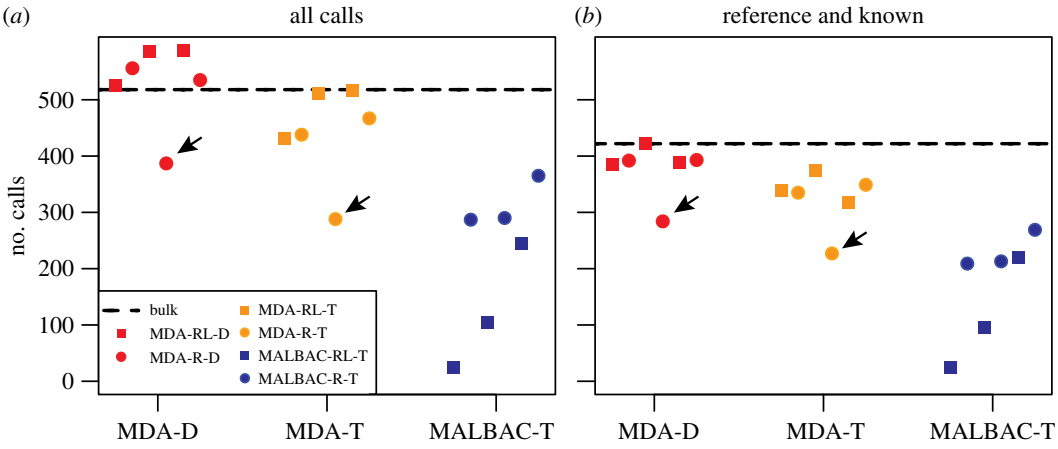

Figure 4. Total number of insertions predicted. (*a*) Including all predictions. (*b*) Including only known L1Hs insertion sites including the reference L1Hs and published polymorphic L1Hs. Black dashed line, bulk DNA regular TIPseq; red, MDA WGA followed by restriction enzyme digestion and ligation with vectorette adaptors (MDA-D); orange, MDA WGA followed by end repair, dA tailing and ligation with dT vectorette adaptor (MDA-T); dark blue, MALBAC WGA followed by ligation with dT vectorette adaptor (MALBAC-T). Circles, WGA using random hexamers only (R); squares, WGA using random hexamers and L1 primer (RL). Arrows indicate the MDA sample included in the vectorette PCR (both MDA-D and MDA-T) and next-generation sequencing stages that had less than perfect QC (corresponds to the same sample indicated in figure 2).

we did not identify novel LINE-1 insertions that could be validated by site-specific PCRs and Sanger sequencing.

Taken together, these findings show that TIPseq and TIPseqHunter data analysis provide near-complete profiles of LINE-1 insertion sites from single cells with infrequent false-positive calls after filtering. Subsequent manual curation and PCR validation of positive calls would still be needed to conclusively demonstrate retrotransposition in samples with somatic mosaicism.

(f) Sensitivity for identifying L1Hs insertions is better in multiple displacement amplification samples than multiple annealing and looping-based amplification cycles samples

To compare the performance of single-cell TIPseq with WGA by MDA and MALBAC, we tested the sensitivity and PPV in detecting L1Hs insertions in both samples. When tested against our 'gold standard' GM12878 insertion list, MALBAC samples had poor sensitivities: average MALBAC sensitivity was only 37% compared to 75% for MDA (*t*-test *p*-value 0.004), although MALBAC did have slightly higher PPV (88 versus 81%, *t*-test *p*-value 0.06).

(g) Restriction digested (MDA-D) templates perform best in single-cell TIPseq

We used two approaches to prepare the template for vectorette PCR. We digested the amplified genomic DNA with restriction enzymes, then ligated the digested DNA fragments with the vectorette adapter that matches the 'sticky ends' of the restriction enzymes (MDA-D); or we converted the amplified genomic DNA to repaired DNA with 5′ phosphorylated and 3′ dA-tailed ends, then ligated the repaired DNA fragments to the vectorette adapter that has dT-tailed ends (MDA-T). For MALBAC WGA products, we did not have an experimental arm to evaluate restriction digestion. Ends of the amplified DNA fragments were used directly in vectorette adapter ligations (MALBAC-T), because MALBAC-amplified fragments had dA tails at the 3′ end, and the size range of amplified fragments is smaller than MDA products, 100 bp–3 kb.

We found that TIPseqHunter2 called more L1Hs insertions in MDA-D samples when compared with MDA-T and MALBAC-T samples (figure 4*a*,*b*) (electronic supplementary material, table S2). When compared with our 'gold standard' GM12878, L1Hs insertion set excluding insertions in or within 25 bp of an annotated non-L1Hs reference L1, MDA-D samples (average sensitivity 82%) performed nearly on a par with a bulk DNA sample (sensitivity 92%), slightly better than MDA-T samples (average sensitivity 70%, *t*-test *p*-value 0.08) and much better than MALBAC-T samples (average sensitivity 37%, *t*-test *p*-value 0.002). PPV was not significantly different between MDA-D and MDA-T (81% versus 81% *t*-test *p*-value 0.88) and was only slightly better for MALBAC-T (81% versus 88%, *t*-test *p*-value 0.06) (figure 3*a*,*b*). In summary, MDA-D samples performed the best with highest sensitivity.

Not surprisingly, the subset of MDA samples that failed to amplify all 12 regions in the QC step but were carried forward to the vectorette PCR stage anyway performed worst in terms of the sensitivity detecting L1Hs insertions in both MDA-D samples (sensitivity 61% compared to average 86% for the other 5 MDA-D samples) and MDA-T samples (sensitivity 48% compared to average 73% for the other 5 MDA-T samples) (figures 2, 3*a*,*b* and 4*a*,*b* arrow pointed). This suggests that samples able to amplify all 12 regions in the QC step could be considered as good quality WGA samples to move forward to the vectorette PCR stage. Thus, QC is a good indicator of the performance potential of an individual sample.

(h) Adding L1Hs-specific primers at whole-genome amplification does not improve sensitivity

We added L1Hs-specific primers into the WGA reaction of both MDA and MALBAC methods to see if this would skew these amplifications towards L1Hs 3′ downstream regions and improve sensitivity. We were not concerned about any chimeric products between the L1Hs-specific

primers and random gDNA fragments, because they should be effectively excluded by TIPseqHunter2.

The quality control test following the WGA step showed no significant difference between the regular MDA whole-genome amplification (MDA-R) samples and those with added L1 primer in the MDA whole-genome amplification (MDA-RL) as far as numbers of loci recovered (*t*-test *p*-value 0.700). Similarly, no difference was appreciated comparing the regular MALBAC whole-genome amplification (MALBAC-R) and preparations using the L1 primer in the MALBAC whole-genome amplification (MALBAC-RL) (*t*-test *p*-value 0.856) (figure 2).

After the complete protocol, we compared the two approaches by evaluating their identification of known GM12878 L1Hs insertions. We found no significant difference between how MDA performed with or without L1 primer (MDA-RL and MDA-R) in sensitivity (86% for MDA-RL-D, 77% for MDA-R-D, *p*-value 0.36; 73% for MDA-RL-T, 71% for MDA-R-T, *p*-value 0.86) or PPV (81% for MDA-RL-D, 81% for MDA-R-D, *p*-value 0.99; 79% for MDA-RL-T, 83% for MDA-R-T, *p*-value 0.27) (figure 3*c–f*). For MALBAC samples, it yields a sensitivity of 24% for MALBAC-RL-T versus 50% for MALBAC-R-T (*t*-test *p*-value 0.17) and PPV 95% for MALBAC-RL-T versus 82% for MALBAC-R-T (*p*-value 0.03) (figure 3*g*,*h*). The inclusion of L1 primer provided modest improvement of PPV for MALBAC-T samples, but not sensitivity.

4. Discussion

LINE-1 is known to retrotranspose in the germline [67], during development [68] and in many human cancers [25–33]. It is possible that increased occurrence of LINE-1 insertions will characterize diseases like Fanconi anaemia [69] or ageing in normal tissues [70]. Single-cell LINE-1 mapping is an emerging tool that can be used to explore somatic mosaicism in benign tissues and genetic heterogeneity in malignancies. Single-cell LINE-1 mapping has been used as a marker of mosaicism in the human brain [51–56,58], and other tissues and disease states may prove important to explore. Despite interest in this topic, there are significant technical challenges inherent in single-cell LINE-1 mapping that have posed a barrier to studies in the field.

Here, we report a new method for single-cell LINE-1 mapping in single cells sorted from a well-characterized HapMap lymphoblastoid cell line. After WGA from single cells, we perform QC by qPCR to decide which amplified well enough to go on the TIPseq protocols. We designed 12 pairs of primers targeting regions 100–800 bp away from the 3′ polyA tail of homozygous L1Hs insertions on 12 different chromosomes. Our findings demonstrate that the samples which yielded amplification for all 12 primer pairs in the QC step showed better overall performance for genome-wide L1Hs insertion site detection. This indicates that the QC step is key to choose well-amplified samples, reducing sequencing cost [53].

In our experiments, we compared MDA- and MALBAC-based WGAs. In the QC step, MDA showed more consistent recovery of genomic sequences downstream of L1Hs than MALBAC. In the complete analysis, MDA followed by TIPseq had higher sensitivity for detecting L1Hs insertions than MALBAC followed by TIPseq. There are several differences in these WGA products: (i) the sizes of MDA fragments are larger, 3–50 kb, while MALBAC produces smaller fragments, only 100 bp–3 kb; and (ii) the amount of DNA produced by MDA is about 40 µg starting from one cell, while the amount of DNA produced by MALBAC is only about 0.1–1 µg. Because of the small fragment sizes, we did not subject MALBAC fragments to restriction digests. Thus, the ligation to vectorette oligonucleotides following digestion depends on a single A/T overhang rather than longer, 'sticky end' ligations, a factor that could reduce the ligation efficiency and reduce the numbers of amplicon templates for the L1Hs-specific vectorette PCR. Consistent with this, we see poorer performance of A-tailed MDA products in vectorette PCRs (MDA-T) when compared with those that have been restriction digested before 'sticky end' ligation (MDA-D).

Another single-cell 3′ focused L1 sequencing, L1 insertion profiling (L1-IP) [16], has been reported [53]. It uses MDA for WGA, followed by a nested PCR using a primer specific to L1Hs ('AC', amplifying Ta and pre-Ta subfamilies) and degenerate primers [16]. Single-cell TIPseq is similar, but uses no nesting in its L1 amplification step; we use an L1 primer more specific to the Ta subfamily of L1Hs (ACA) [9] paired with a specific vectorette primer [42,47]. The latter requires additional steps to ligate sequences corresponding the vectorette primer to the DNA templates. TIPseq also breaks up these amplicons before sequencing, potentially resulting in reads more distributed downstream of L1 insertions. Both TIPseqHunter [31] and the computational analysis performed for single-cell L1-IP [53] rely on classifiers that use features of the LINE-1 polyA and its juxtaposition with unique genomic sequence. Whereas single-cell L1-IP analysis generates training data through an iterative process that uses the result of a previous iteration to train next iteration, we used a training set generated from fixed present insertions and amplified regions that lack evidence for the L1 primer binding region. Furthermore, we also used an svm model with radial basis kernel rather than logistic regression. This allows a nonlinear classification boundary that may perform better when one feature strongly suggests insertion, but another feature does not.

Importantly, this study represents one of the first reports of a single-cell transposon insertion site profiling protocol that compares different conditions and tests a well-studied genomically stable cell line. Testing on a cell line with known LINE-1 insertions allowed us to carefully investigate the basis for artefacts that lead to false-positive calls and filter them out. It allows us to provide a robust estimate of the accuracy of this method.

In the future, the specificity and sensitivity of single-cell LINE-1 mapping could be further improved by harnessing emerging single-molecule, long-read sequencing technologies. In the interim, though, it is clear that deep coverage of the 3′ end of L1Hs insertion sites is possible by coupling WGA with standard TIPseq protocols. This approach is an economical and robust one for resolving the occurrence of retrotransposition events in single human cells.

Data accessibility. All the sequencing data are deposited at NCBI Sequence Read Archive (SRA) with BioProject accession number PRJNA547805. TIPseqHunter2 software is available on github at https://github.com/FenyoLab/TIPseqHunter. The analysed data supporting this article have been uploaded as part of the electronic supplementary material.

Authors' contributions. W.M.: analysis and interpretation of data, drafting and revising the article. Z.T.: analysis of data. J.P.S.: technical

support, revising the article. L.M.P.: technical support, revising the article. J.D.B.: technical support, revising the article. D.K.: revising the article. D.F.: interpretation of data, revising the article. K.H.B.: design and interpretation of data, drafting and revising the article, final approval of the version to be published. C.L.: design, acquisition, and interpretation of data, drafting and revising the article, final approval of the version to be published.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the National Institutes of Health (NIH) (grant no. R01GM124531 to K.H.B.) and (grant no. P01AG051449 subcontracts to K.H.B. and D.F.).

Acknowledgements. The authors acknowledge technical advice from Laura Wood and helpful discussions with Melanie Weigert.

References

1. Lander ES *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. (doi:10.1038/35057062)
2. Burns KH, Boeke JD. 2012 Human transposon tectonics. *Cell* **149**, 740–752. (doi:10.1016/j.cell.2012.04.019)
3. Skowronski J, Fanning TG, Singer MF. 1988 Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8**, 1385–1397. (doi:10.1128/MCB.8.4.1385)
4. Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD. 2000 Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508. (doi:10.1101/gr.149400)
5. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 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. (doi:10.1016/0092-8674(93)90078-5)
6. Cost GJ, Feng Q, Jacquier A, Boeke JD. 2002 Human L1 element target-primed reverse transcription *in vitro*. *EMBO J.* **21**, 5899–5910. (doi:10.1093/emboj/cdf592)
7. Feng Q, Moran JV, Kazazian Jr HH, Boeke JD. 1996 Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916. (doi:10.1016/s0092-8674(00)81997-2)
8. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian Jr HH. 1996 High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927. (doi:10.1016/s0092-8674(00)81998-4)
9. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian Jr HH. 2003 Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285. (doi:10.1073/pnas.0831042100)
10. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010 LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170. (doi:10.1016/j.cell.2010.05.021)
11. Badge RM, Alisch RS, Moran JV. 2003 ATLAS: a system to selectively identify human-specific L1 insertions. *Am. J. Hum. Genet.* **72**, 823–838. (doi:10.1086/373939)
12. Dewannieux M, Esnault C, Heidmann T. 2003 LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48. (doi:10.1038/ng1223)
13. Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian Jr HH. 2011 Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400. (doi:10.1093/hmg/ddr245)
14. Batzer MA, Deininger PL. 2002 Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379. (doi:10.1038/nrg798)
15. Xing J *et al*. 2009 Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* **19**, 1516–1526. (doi:10.1101/gr.091827.109)
16. Ewing AD, Kazazian Jr HH. 2010 High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **20**, 1262–1270. (doi:10.1101/gr.106419.110)
17. Ewing AD, Kazazian Jr HH. 2011 Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* **21**, 985–990. (doi:10.1101/gr.114777.110)
18. Hormozdiari F *et al*. 2011 Alu repeat discovery and characterization within human genomes. *Genome Res.* **21**, 840–849. (doi:10.1101/gr.115956.110)
19. Stewart C *et al*. 2011 A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, e1002236. (doi:10.1371/journal.pgen.1002236)
20. Sudmant PH *et al*. 2015 An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81. (doi:10.1038/nature15394)
21. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015 The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol. Spectr.* **3**, MDNA3-0061-2014. (doi:10.1128/microbiolspec.MDNA3-0061-2014)
22. Kazazian Jr HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988 Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166. (doi:10.1038/332164a0)
23. Hancks DC, Kazazian Jr HH. 2016 Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9. (doi:10.1186/s13100-016-0065-9)
24. Payer LM, Burns KH. 2019 Transposable elements in human genetic disease. *Nat. Rev. Genet.* **20**, 760–772. (doi:10.1038/s41576-019-0165-8)
25. Rodic N *et al*. 2014 Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am. J. Pathol.* **184**, 1280–1286. (doi:10.1016/j.ajpath.2014.01.007)
26. Leibold DM, Swergold GD, Singer MF, Thayer RE, Dombroski BA, Fanning TG. 1990 Translation of LINE-1 DNA elements in vitro and in human cells. *Proc. Natl Acad. Sci. USA* **87**, 6990–6994. (doi:10.1073/pnas.87.18.6990)
27. Ardeljan D, Taylor MS, Ting DT, Burns KH. 2017 The human long interspersed element-1 retrotransposon: an emerging biomarker of neoplasia. *Clin. Chem.* **63**, 816–822. (doi:10.1373/clinchem.2016.257444)
28. De Luca C, Guadagni F, Sinibaldi-Vallebona P, Sentinelli S, Gallucci M, Hoffmann A, Schumann GG, Spadafora C, Sciamanna I. 2016 Enhanced expression of LINE-1-encoded ORF2 protein in early stages of colon and prostate transformation. *Oncotarget* **7**, 4048–4061. (doi:10.18632/oncotarget.6767)
29. Burns KH. 2017 Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424. (doi:10.1038/nrc.2017.35)
30. Nguyen THM *et al*. 2018 L1 retrotransposon heterogeneity in ovarian tumor cell evolution. *Cell Rep.* **23**, 3730–3740. (doi:10.1016/j.celrep.2018.05.090)
31. Tang Z *et al*. 2017 Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proc. Natl Acad. Sci. USA* **114**, E733–E740. (doi:10.1073/pnas.1619797114)
32. Doucet-O'Hare TT *et al*. 2015 LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc. Natl Acad. Sci. USA* **112**, E4894–E4900. (doi:10.1073/pnas.1502474112)
33. Doucet-O'Hare TT, Sharma R, Rodic N, Anders RA, Burns KH, Kazazian Jr HH. 2016 Somatically acquired LINE-1 insertions in normal esophagus undergo clonal expansion in esophageal squamous cell carcinoma. *Hum. Mutat.* **37**, 942–954. (doi:10.1002/humu.23027)
34. Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011 LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215. (doi:10.1146/annurev-genom-082509-141802)
35. Tubio JMC *et al*. 2014 Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343. (doi:10.1126/science.1251343)
36. Lee E *et al*. 2012 Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971. (doi:10.1126/science.1222077)
37. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. 2014 Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063. (doi:10.1101/gr.163659.113)

38. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010 Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261. (doi:10.1016/j.cell.2010.05.020)
39. Solyom S *et al.* 2012 Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2328–2338. (doi:10.1101/gr.145235.112)
40. Doucet TT, Kazazian Jr HH. 2016 Long interspersed element sequencing (L1-Seq): a method to identify somatic LINE-1 insertions in the human genome. *Methods Mol. Biol.* **1400**, 79–93. (doi:10.1007/978-1-4939-3372-3_5)
41. Rodic N *et al*. 2015 Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat. Med.* **21**, 1060–1064. (doi:10.1038/nm.3919)
42. Huang CR *et al*. 2010 Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**, 1171–1182. (doi:10.1016/j.cell.2010.05.026)
43. Shukla R *et al*. 2013 Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111. (doi:10.1016/j.cell.2013.02.032)
44. Sanchez-Luque FJ, Richardson SR, Faulkner GJ. 2016 Retrotransposon capture sequencing (RC-Seq): a targeted, high-throughput approach to resolve somatic L1 retrotransposition in humans. *Methods Mol. Biol.* **1400**, 47–77. (doi:10.1007/978-1-4939-3372-3_4)
45. Streva VA, Jordan VE, Linker S, Hedges DJ, Batzer MA, Deininger PL. 2015 Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. *BMC Genomics* **16**, 220. (doi:10.1186/s12864-015-1374-y)
46. Rahbari R, Badge RM. 2016 Combining amplification typing of L1 active subfamilies (ATLAS) with high-throughput sequencing. *Methods Mol. Biol.* **1400**, 95–106. (doi:10.1007/978-1-4939-3372-3_6)
47. Steranka JP *et al*. 2019 Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome. *Mob DNA* **10**, 8. (doi:10.1186/s13100-019-0148-5)
48. Ho HJ, Ray DA, Salem AH, Myers JS, Batzer MA. 2005 Straightening out the LINEs: LINE-1 orthologous loci. *Genomics* **85**, 201–207. (doi:10.1016/j.ygeno.2004.10.016)
49. Salem AH, Ray DA, Batzer MA. 2005 Identity by descent and DNA sequence variation of human SINE and LINE elements. *Cytogenet. Genome Res.* **108**, 63–72. (doi:10.1159/000080803)
50. Vincent BJ, Myers JS, Ho HJ, Kilroy GE, Walker JA, Watkins WS, Jorde LB, Batzer MA. 2003 Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites. *Mol. Biol. Evol.* **20**, 1338–1348. (doi:10.1093/molbev/msg146)
51. Singer T, McConnell MJ, Marchetto MC, Coufal NG, Gage FH. 2010 LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends Neurosci.* **33**, 345–354. (doi:10.1016/j.tins.2010.04.001)
52. Richardson SR, Morell S, Faulkner GJ. 2014 L1 retrotransposons and somatic mosaicism in the brain. *Annu. Rev. Genet.* **48**, 1–27. (doi:10.1146/annurev-genet-120213-092412)
53. Evrony GD *et al*. 2012 Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496. (doi:10.1016/j.cell.2012.09.035)
54. Evrony GD *et al*. 2015 Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59. (doi:10.1016/j.neuron.2014.12.028)
55. Upton KR *et al*. 2015 Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239. (doi:10.1016/j.cell.2015.03.026)
56. Evrony GD, Lee E, Park PJ, Walsh CA. 2016 Resolving rates of mutation in the brain using single-neuron genomics. *eLife.* **5**, e12966. (doi:10.7554/eLife.12966)
57. Lodato MA *et al*. 2015 Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98. (doi:10.1126/science.aab1785)
58. Sanchez-Luque FJ *et al*. 2019 LINE-1 evasion of epigenetic repression in humans. *Mol. Cell* **75**, 590–604. (doi:10.1016/j.molcel.2019.05.024)
59. Consortium IH. 2003 The international HapMap project. *Nature* **426**, 789–796. (doi:10.1038/nature02168)
60. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)
61. Langmead B, Salzberg SL. 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. (doi:10.1038/nmeth.1923)
62. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, Sermon K. 2006 Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* **1**, 1965–1970. (doi:10.1038/nprot.2006.326)
63. Zong C, Lu S, Chapman AR, Xie XS. 2012 Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626. (doi:10.1126/science.1229164)
64. Auton A *et al*. 2015 A global reference for human genetic variation. *Nature* **526**, 68–74. (doi:10.1038/nature15393)
65. Mir AA, Philippe C, Cristofari G. 2015 euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.* **43**, D43–D47. (doi:10.1093/nar/gku1043)
66. Lasken RS, Stockwell TB. 2007 Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* **7**, 19. (doi:10.1186/1472-6750-7-19)
67. Kazazian Jr HH, Moran JV. 2017 Mobile DNA in health and disease. *N Engl. J. Med.* **377**, 361–370. (doi:10.1056/NEJMra1510092)
68. Faulkner GJ, Garcia-Perez JL. 2017 L1 mosaicism in mammals: extent, effects, and evolution. *Trends Genet.* **33**, 802–816. (doi:10.1016/j.tig.2017.07.004)
69. Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, Wysocka J. 2018 Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* **553**, 228–232. (doi:10.1038/nature25179)
70. De Cecco M *et al*. 2019 L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566**, 73–78. (doi:10.1038/s41586-018-0784-9)