Supplemental Information

Supplemental Note 1 - Single-neuron whole genome sequencing performance analyses

Our large high-coverage whole genome sequencing (WGS) dataset afforded us the opportunity to examine in detail the sequencing read alignment statistics, genome coverage, and chimera mechanisms of single-cell MDA, as an aid for future single-cell genomics research in understanding mechanisms of single-cell amplification and developing future improved amplification methods. This detailed analysis is important as there is no published high-coverage single-cell WGS data from primary human cells, and in prior studies of MDA only 2 high-coverage WGS single-cells from a lymphoblastoid cell line have been previously published (Hou et al., 2012). We included in our coverage and alignment analyses samples from two prior studies of high-coverage single-cell WGS: 2 single cells from a lymphoblastoid cell line (YH) amplified by MDA (Hou et al., 2012) and 5 single cells from a cancer cell line (SW480) amplified by MALBAC (Zong et al., 2012), as well as their respective unamplified bulk DNA samples. Comparison to these samples helps distinguish trends specific to our data versus MDA or single-cell amplification generally.

I. Alignment statistics

A summary of basic sequencing read alignment statistics is presented in **Table S1**. Our unamplified bulk DNA and MDA samples (100-neuron and single-neurons) showed similar insert size distributions, as expected due to processing in the same sequencing library protocol. The fraction of mapped reads was also similar across all samples, with 98-99% of reads mapping to the genome in every unamplified bulk DNA and MDA sample. As expected, mitochondrial genome reads were absent from MDA 100-neuron and single-cell samples, as these samples were amplified from purified nuclei, thereby excluding mitochondrial DNA.

Interestingly, relative to unamplified bulk DNA, MDA samples showed a consistent ~9-fold reduction in the fraction of reads mapping to the decoy contig (hs37d5): 3.7% versus 0.4% of reads mapped to the decoy in bulk versus MDA samples, respectively (**Table S1**). The decoy contig was added by the 1000 genomes project to the human genome reference in order to reduce false positive alignments due to repetitive and unassembled regions of the genome, and consists mostly of repetitive sequence elements. As expected, most of the increase in decoy contig mapped reads in bulk samples was due to non-uniquely mapped reads, indicating they stemmed from repetitive sequences. This in turn also led to a slightly higher overall fraction of uniquely mapped reads in MDA samples compared to bulk DNA samples (92% versus 89%, respectively) (**Table S1**). Further investigation revealed that the repetitive reads mapping to the decoy contig in bulk but not MDA samples derived from satellite DNA sequences, the primary non-coding DNA of constitutive heterochromatin such as centromeres. We therefore performed an analysis of all annotated satellite DNA regions in the genome, and found 1.2% versus 0.1% of reads aligned to satellite DNA regions in bulk versus MDA samples, respectively (**Table S1**). This indicates systematic under-amplification of satellite DNA regions in MDA samples. Under-amplified satellite DNA regions were primarily peri-centromeric and encompassed several families of satellite DNA, including ALR-alpha, HSATII, CATTC repeats, CER (centromeric repeats), and BSR-beta. The decrease in satellite DNA reads was also seen in YH MDA single cells compared to YH bulk DNA, while MALBAC single cell samples did not show this effect.

Two possible explanations for the reduction in satellite DNA reads in MDA samples are: a) highly condensed heterochromatin of satellite DNA in peri-centromeres and centromeres may not be adequately denatured with the standard alkaline lysis most commonly used in MDA single-cell amplification; or, b) due to the repetitiveness of satellite DNA, random hexamers used in MDA may not anneal at sufficient density in these regions to enable exponential amplification. However, the former may be the more likely explanation due to the variety of satellite DNA families that are under-amplified. Furthermore, our analysis did not reveal a correlation of satellite DNA GC-content with read depth, arguing against GC bias as a possible explanation. Since under-amplified satellite DNA regions (i.e. regions with >2-fold coverage difference in bulk versus MDA samples, and at least 1/10 the average normalized coverage depth in bulk samples) are mostly peri-centromeric/centromeric gene-poor regions, and account for a small fraction of the assembled genome reference (11.7Mb, 0.4% of the genome), their under-amplification would not significantly affect studies of functional regions of the genome. Nevertheless, this analysis suggests that alternative denaturation methods such as heat or proteinase K digestion may be necessary to capture these regions, and that part of the genome-wide variability in MDA amplification may be due to differences in initial denaturation efficiency of heterochromatin.

Next, we studied the prevalence of discordant read pairs in our samples. Both true positive structural variants and chimeras (false positive structural variants created during MDA amplification and during sequencing library preparation) appear in the data as 'discordant' read pairs whose reads either: a) do not align with the correct orientation pointing towards each other; b) align too distant from each other, outside the expected size distribution of DNA fragments used to construct the library; c) align to different chromosomes; or d) only one of the two reads of the pair aligns to the genome. In contrast, 'concordant' read pairs align with the correct orientation and at the expected distance from each other in the genome. Interestingly, the average percentage of concordant and discordant reads per sample was similar in both bulk samples and 4 MDA single-neuron samples whose sequencing libraries were prepared in the same batch as the bulk samples (to control for batch differences in sequencing library chimera rates)— bulk cortex and heart, versus MDA single-neurons #2, 3, 6, 77. Concordantly paired reads comprise 95% of reads per sample on average in bulk samples and 95% of reads on average in the 4 MDA single-neuron samples from the same batch (and nearly the same, 96%, across all 16 MDA single neurons regardless of sequencing library preparation batch) (**Table S1**). The average percentage of discordant read pairs per sample was also similar between the bulk and the 4 MDA single neurons prepared in the same sequencing library preparation batch, comprising 4.3% and 3.9% of reads, respectively (and 2.8% across all 16 MDA single neurons) (**Table S1**). Furthermore, the subset of discordant read pairs in which both reads of the pair aligned to the genome was similar between these groups, comprising 2.8% and 2.3% of total reads on average in the bulk and 4 MDA single-neuron samples prepared in the same sequencing library batch, respectively (and 1.9% across all 16 MDA single neurons) (**Table S1**). SW480 MALBAC single-cell samples had a lower percentage of concordant reads, 89%, compared to 98% in their respective SW480 bulk sample (**Table S1**).

An analysis of the distribution of discordant reads showed that most derive from chimeras rather than true structural variants. Discordant reads pile up at higher read depth at the breakpoints of true structural variants, whereas discordant reads arising from chimeras created during MDA amplification or during sequencing library preparation occur randomly across the genome. Discordant reads derived from chimeras would therefore be expected to be randomly

distributed at lower read depth across the genome, in contrast to pile-ups of discordant reads at breakpoints of true structural variants. We therefore roughly estimated the fraction of discordant reads that are chimeras by measuring the fraction of discordant reads that are in regions that have ≤ 3x read depth of discordant reads (analyzing plus and minus strand discordant reads separately, and excluding reads mapped to satellite regions and contigs other than autosomes or sex chromosomes). This estimated that on average 89% and 71% of discordant reads are chimeras in the bulk and 4 MDA single-neuron samples from the same batch, respectively, and 76% across all 16 MDA single neurons. This estimate indicates that most discordant reads in our samples derive from chimeras rather than structural variants. This is in fact a lower bound for the fraction of discordant reads that are chimeras in MDA samples since chimeras occurring early in MDA amplification and regions with a higher propensity to create MDA chimeras would create regions with pile-ups of discordant reads. The lower fraction of discordant reads that are estimated to be chimeras in single-neuron samples may be due to these chimeras occurring in early stages of MDA amplification.

Chimeras and true structural variants also appear as reads that only partially align to the genome ('clipped' reads). Clipped reads were 0.9% and 1.2% of reads in the bulk and 4 MDA single-neuron samples prepared in the same sequencing library batch, respectively (and 1.0% of reads across all 16 MDA single neurons) (**Table S1**). Similarly to discordant reads, most clipped reads also derive from chimeras rather than true structural variants, since the average fraction of clipped reads in regions with ≤ 3x read depth of clipped reads was 89% and 88% in the bulk and 4 MDA single-neuron samples from the same batch, respectively (and 86% across all 16 MDA single neurons).

The similar fraction of both discordant and clipped reads in bulk and MDA samples (especially for samples from the same sequencing library preparation batch) suggests that MDA chimeras account for a minority of the total chimeras in MDA samples, and that most chimeras in both bulk and MDA samples were created during sequencing library preparation. This is because the bulk and MDA samples were processed by an identical sequencing library preparation protocol, and would therefore have the same sequencing library chimera rate. Any MDA chimeras in MDA samples would be in addition to these sequencing library chimeras, so if MDA chimeras accounted for a large fraction of all chimeras we would have expected a higher rate of discordant and clipped reads in MDA single-neuron samples. Chimeras can be created during sequencing library preparation when two DNA fragments are ligated to each other prior to adaptor ligation, due to incomplete end-repair or dA-tailing of DNA fragments. Since sequencing library chimeras occur after DNA fragmentation, they would be expected to occur more often between fragments from different chromosomes or between distant regions on the same chromosome, in contrast to MDA chimeras, which occur between nearby regions on the same chromosome (Evrony et al., 2012; Lasken and Stockwell, 2007) (see also section on 'Whole-genome sequencing analysis of MDA chimeras'). Indeed, in bulk and MDA single-neuron samples from the same batch, 0.8% and 0.7% of all reads, respectively, were from discordant read pairs where both reads of the pair mapped to the same chromosome, whereas 2.1% and 1.6% of total reads, respectively, were from discordant pairs where the reads mapped to different chromosomes (**Table S1**). The larger fraction of discordant read pairs mapping to different chromosomes in both bulk and MDA samples is consistent with most chimeras arising during sequencing library preparation rather than MDA. In fact, a separate detailed analysis of MDA chimeras showed that MDA chimeras comprise only 0.4% of total reads in MDA single cell samples (**Table S1**; see section on 'Whole-genome sequencing analysis of MDA chimeras' for

details). Furthermore, analysis of publicly available high coverage WGS bulk samples created a with different sequencing library protocol (NA12877-ERX069504, NA12878-ERX069505, and NA12882-ERX069506 available on NCBI SRA) (data not shown) revealed fewer discordant pairs mapping to the same chromosome (0.4%) as well as significantly fewer discordant pairs mapping to different chromosomes (0.1%) and clipped reads (0.3%) compared to bulk and MDA samples of this study, indicating that sequencing library preparation for these samples created fewer chimeras. Variability in the amount of chimeras created during library preparation is known to occur due to differences in enzyme preparations' efficiencies in DNA end-repair and dA-tailing, and could be circumvented with improved enzyme preparations and additional DNA size selection prior to ligation (Quail et al., 2008). Nevertheless, since chimeras created during sequencing library ligation are randomly distributed in the genome, they would not be expected to significantly affect somatic structural variant calling in single cell WGS samples, though they would affect the reliability of low-level somatic structural variant calling in bulk samples.

Finally, we studied the effect of the hs37d5 decoy contig on the above general alignment statistics by aligning all the samples to the same human genome reference but without the decoy contig. Samples mapped to a reference with the decoy compared to without the decoy had identical alignment statistics within 1% difference (percentage of total reads), with the following exceptions: in bulk samples, an additional 3% of the total reads were mapped on average, and most of these additional mapped reads were concordantly paired non-uniquely mapped reads. Additionally, when the decoy contig was included, bulk samples had 1% fewer of their total reads mapped to non-decoy satellite regions. Overall, these results are consistent with the greater number of satellite DNA reads in bulk samples described previously, and indicate that the decoy contig leads both to mapping of repetitive satellite DNA reads that otherwise would not have been mapped, as well as to mapping of satellite reads to the decoy that would otherwise have mapped to satellite DNA regions in other regions of the genome.

II. Genome coverage analyses

We studied the genome-wide coverage achieved by our WGS samples, including the fraction of the genome covered, the distribution of read depth coverage across the genome, the relative coverage of certain genomic features such as retrotransposons annotated in the human genome reference, as well as comparisons to the previously published high-coverage WGS of MDA single-cells (YH cell line) (Hou et al., 2012) and MALBAC single-cells (SW480 cell line) (Zong et al., 2012) described above.

Analysis methods

Genome coverage statistics were calculated after filtering PCR duplicate reads, relative to the hs37d5 human genome reference (1000 Genomes Project reference based on the GRCh37 primary assembly) excluding assembly gaps annotated by the Genome Reference Consortium (all 'N' sequences to which reads cannot align), and the mitochondria (MT), decoy (hs37d5), and human herpes virus (NC_007605) contigs. Retrotransposons and Refseq exon annotations of the human genome were obtained from RepeatMasker annotations and the UCSC genome browser.

Plots of the fraction of the genome covered above different normalized read depth cutoffs (**Figure S1A**) were constructed by first extracting the read depth distribution of each sample.

The read depth distribution is defined as the fraction of the genome covered at each read depth. Then, for each integer read depth cutoff *r* (1x, 2x, 3x, …), the fraction of the genome with read depth ≥ *r* is calculated from the read depth distribution. This fraction (y-axis) is then plotted against the normalized cutoff *r/R* (x-axis), where *R* is the genome-wide average read depth of the sample. For example, if a sample was sequenced at a total genome-wide average read depth of 40x (*R = 40)*, and half of the genome was covered at ≥30x read depth (*r* = 30), then a point is plotted with an x-value at a normalized read depth cutoff of 0.75 (i.e. 30/40), and a y-value of 0.5. Specifically, given the read depth distribution of each sample, this plot shows the complementary cumulative distribution function of the read depth distribution, where the read depth is normalized to genome-wide average read depth.

Subsampling analyses (**Figure S1B**) were performed by randomly sampling reads using SAMtools. For each sample, the fraction, *s*, of reads subsampled out of the total number of reads, *T*, to achieve a target genome-wide average read depth*, C*, was calculated as $s = \frac{C}{T \cdot readLength\,(bp) / genomeSize(bp)}$. Subsampling was performed for each integer genome-wide average read depth (1x, 2x, 3x, …) up to the maximum possible total read depth for each sample. Subsampling of bulk and single-cell samples from the SW480 cell line was performed after read trimming, taking into account post-trimming average read lengths and reads eliminated due to short length after trimming, to avoid biasing against these samples (undersampling) due to trimming.

Lorenz curves (**Figure S1C**) were constructed by plotting for each integer genome-wide average read depth *r* (1x, 2x, 3x, …) a point with x-value equal to the fraction of the genome with ≤ *r* read depth and y-value equal to the fraction of reads that are in regions of the genome with ≤ *r* read depth.

Genome-wide coverage distribution

We first studied the fraction of the genome covered and how evenly reads were distributed across the genome in our WGS samples. Bulk DNA, MDA 100-neuron, and MDA single-neuron samples achieved genome-wide average read depths of 32x±0.2 (SD), 32x, and 42x±7 (SD), respectively (**Table S2**). The variability in the average read depth per sample reflects variability in the total number of sequencing runs performed per sample and variability in output of the sequencing instruments. At these genome-wide average read depths, bulk DNA, MDA 100-neuron, and MDA single-neuron samples achieved coverage of 100±0% (SD), 98%, and 98±0% (SD) of the genome at ≥1x read depth, respectively, and 99±0%, 83%, and 81±2% of the genome at ≥10x read depth, respectively (**Table S2**). This represents a locus dropout rate of ~2% (69Mb on average) in single-neuron samples and ~2% (55Mb) in the MDA 100-neuron sample, compared to a bulk DNA locus dropout of 0.2% (5Mb), consistent with prior studies estimating single-cell MDA locus dropout by targeted genotyping (Evrony et al., 2012) and high-coverage sequencing (Hou et al., 2012). The two YH MDA single-cells had 3% and 8% locus dropout (**Table S2**), but this may be due to their relatively lower total read depth compared to our samples. SW480 MALBAC single-cells had an average locus dropout of 15%; the 2 MALBAC single-cells with the best genome coverage (SRX202787 and SRX202978) had a locus dropout of 7%. The increased locus dropout of YH MDA and MALBAC single-cells cannot be accounted for by their cell lines of origin, as their corresponding bulk samples had 0% and 1%

locus dropout, respectively. Therefore, our MDA single-neurons appear to have less locus dropout than that seen in prior single-cell high-coverage WGS studies.

We further investigated the characteristics of locus dropout regions. Satellite regions accounted for 7% and 12% of the locus dropout (percentage of locus dropout bases) in MDA single-neuron and 100-neuron samples, respectively, compared to only 1% of dropout in bulk samples. The average GC-content of all locus dropout regions in MDA samples was 0.49, higher than the genome average of 0.41. There was also an increasing average GC content for locus dropout regions as the number of single-neurons sharing the dropout region increased— regions that were dropped out in all 16 single neurons had an average GC of 0.52 and regions that were dropped out in only one single neuron had an average GC of 0.45. This suggests that satellite regions and GC content explains much of the locus dropout in MDA samples. See 'MDA GC-content amplification bias' section for further GC content analyses.

Since the prior genome coverage statistics do not control for total sequencing read depth, we next performed a set of analyses controlling for the total number of reads in each sample, thereby allowing more direct comparisons of genome coverage between samples and sample types. We plotted the fraction of the genome covered above different minimum read depth cutoffs, while normalizing the read depth cutoffs to the genome-wide average read depth (**Figure S1A**). An ideal sample would have perfectly even coverage across the genome, resulting in a step function at y=1. Deviation from this indicates a wider read depth distribution with a larger fraction of the genome covered at both higher and lower read depths than the genome-wide average read depth. This analysis showed that bulk samples approach the ideal distribution, while MDA samples deviate from this, with more regions of the genome at both higher and lower read depths. Although these plots show a wider read depth distribution for MDA samples, MDA single-neuron samples showed highly consistent distributions (**Figure S1A**; left plot).

Similar plots for the YH MDA and MALBAC single-cells with the best genome coverage (YH-1, SRX202787 and SRX202978) showed similar patterns, but slightly more pronounced deviations than MDA single-neuron samples (**Figure S1A**; right plot). However, their corresponding bulk samples also showed greater deviation than 1465 bulk DNA samples, so part of the deviation of YH MDA and MALBAC single-cell samples may be due to factors causing wider read depth distributions more generally in these studies, such as GC biases arising during sequencing library PCR. Interestingly, the YH MDA single-cell had a relatively higher fraction of the genome covered than MDA single neuron samples at low normalized read depth cutoffs (**Figure S1A**; right plot; YH-1 line above MDA single neurons at normalized read depth cutoffs < 0.3). At increasing cutoffs, the YH MDA single-cell then showed the pattern characteristic of a wider read depth distribution compared to MDA single neurons: less fraction of the genome covered at intermediate cutoffs (between 0.3 and 2.25) and greater fraction of the genome covered at high (>2.25) cutoffs. Further investigation of the read depth distribution of YH MDA single-cell YH-1 (and similarly for YH MDA single-cell YH-2) showed that this sample has a bimodal read depth distribution, with a defined peak below the genome-wide average read depth similar in shape to peaks seen in bulk samples, but also a long tail at higher read depths that is characteristic of MDA samples (data not shown). The mechanism for this bimodal distribution is unclear, and perhaps may reflect different kinetics of amplification in different regions of the single-cell genome in that study's implementation of MDA. Therefore, although the YH MDA single-cell has an overall wider read depth distribution than MDA single neuron samples, the YH MDA single cell has a qualitatively different read depth distribution shape, with a better (more

narrow) read depth distribution at low read depths but a wider read depth distribution at high read depths. Overall, the wide read depth distribution of all analyzed single-cell amplification methods highlights the need for development of new methods with more even genome coverage, which will have significant benefit for downstream mutation analyses.

Next, we subsampled reads from each sample to different integer (1x, 2x, 3x, …) genome-wide average read depths. At each subsampled genome-wide average read depth we then evaluated the fraction of the genome covered at ≥1x and ≥10x read depth (**Figure S1B**). Subsampling analyses provide a view of the total sequencing read depth necessary to obtain a desired genome coverage and assist in determining at which point additional sequencing yields diminishing returns in terms of coverage. Both of these performance measures are determined by the underlying variability in coverage across the genome (i.e. the read depth distribution). Since subsampling in all samples was performed at the same intervals of genome-wide average read depths (i.e. 1x, 2x, 3x…), it was possible to directly compare performance across all the samples regardless of their original total sequencing read depth.

Bulk samples achieved 95% coverage of the genome at ≥1x with a subsampled genome-wide average read depth of only 4x. In contrast, at the same subsampled read depth, MDA single-neurons on average cover 80% of the genome at ≥1x read depth (**Figure S1B**). MDA single-neurons require sequencing up to 18x read depth before 95% of the genome is covered at ≥1x. Note, however that these plots do not reflect the overall read depth distribution as in **Figure S1A**; MDA single-neuron samples have not only less of the genome covered at the genome-wide average read depth compared to bulk samples, but also more of the genome covered at higher read depths than bulk samples. For example, at a subsampled genome-wide average read depth of 4x, MDA single neurons have less of the genome covered at ≥1x but more of the genome covered at ≥10x than bulk samples (**Figure S1B**). The YH MDA single-cell (YH-1) achieves 95% genome coverage at ≥1x with a subsampled read depth of 11x, less than the read depth necessary to achieve the same coverage by MDA single neurons. This reflects the qualitatively different read depth distribution of the YH MDA single-cell as discussed previously. At this same subsampled read depth, MDA single-neurons cover more of the genome at ≥10x than the YH MDA single-cell. The two SW480 MALBAC single-cells with the best coverage (SRX202787 and SRX202978) reach 93% genome coverage at ≥1x with their maximum subsampled read depth of 31x. Overall subsampling plots indicate that MDA single-cells achieve greater overall genome coverage at lower read depths compared to MALBAC. However, note that in terms of other performance measures, MALBAC still maintains significant advantages versus MDA, specifically in terms of remarkable consistency in genome coverage performance between single-cells (SD is nearly 0 in genome coverage between the 2 best MALBAC single-cells, **Figure S1A**) as well as better evenness of genome coverage at large genomic scales (see 'Coverage variability analyses' section for further details).

Lorenz curves for each sample type were plotted as in Zong, et al (2012) to provide an additional view of the read depth distribution in a way that controls for different total sequencing read depths between samples (**Figure S1C**). In Lorenz curves, the cumulative fraction of reads is plotted as a function of the cumulative fraction of the genome covered at increasing read depths. Perfectly even coverage across the genome would approximate the y = x line. Bulk samples of individual 1465 show a nearly ideal Lorenz curve. MDA 100-neuron and MDA single-neuron samples each deviate farther, respectively, from the bulk sample curve. YH MDA and SW480 MALBAC single-cell samples are deviated more than MDA single-neuron samples, however,

their corresponding bulk DNA samples also deviate more than bulks samples of individual 1465. Therefore, as in the analysis of **Figure S1A**, some of the increased deviation from uneven genome coverage in YH MDA and SW480 MALBAC single-cells relative to MDA single-neurons may be due to additional biases in library preparation of these prior studies. Additionally, Lorenz curves were plotted after subsampling to equal subsampled read depths across all samples (matching the sample with the lowest total read depth) (**Figure S1C**, right panel). This confirmed that Lorenz curves control well for variable total sequencing read depths among samples, since the same trends seen in Lorenz curves constructed without subsampling (i.e. using all reads for each sample) (**Figure S1C**, left panel) were seen after subsampling (**Figure S1C**, right panel).

Coverage of annotated retrotransposons

Average genome coverage of all retrotransposons annotated in the human genome reference ('known' reference insertions) for each active retrotransposon family (L1Hs, AluY, SVA) is presented in **Table S2**, as well as for older L1Pa elements, all L1 elements, and Refseq exons for comparison. The average coverage of each genomic feature relative (normalized to) the genome-wide average coverage was calculated to measure the degree to which each feature type was over- or under-amplified on average. Additionally, the 500bp flanks of L1Hs, AluY, and SVA were separately analyzed since coverage of the sequences flanking retrotransposon insertions would impact detection by WGS. A subset of this analysis is plotted in **Figure 1C**. Note, however, that this analysis is performed on annotated reference insertions, so these analyses also reflect insertion site sequence biases resulting from the effects of selection over evolutionary time (Ovchinnikov et al., 2001; Szak et al., 2002). Evolutionarily recent insertions have a significantly less biased distribution in the genome compared to older insertions (Boissinot et al., 2000; Cordaux et al., 2006; Ovchinnikov et al., 2001). Therefore, somatic insertions likely distribute more randomly in the genome so that their local regions would more closely approximate the genome-wide average coverage.

In MDA samples, L1Hs, AluY, and SVA insertions annotated in the human genome reference are over- or under-amplified in a manner that correlates with the GC content of their sequences and flanks (**Table S2**). MDA single-neuron samples over-amplify known reference L1Hs insertions and their flanks by 10% and 14%, respectively; AluY insertions and their flanks are under-amplified by 26% and 18%, respectively; SVA insertions and their flanks are under-amplified by 42% and 31%, respectively. The GC contents of L1Hs, AluY, and SVA are progressively higher (0.42, 0.54, and 0.63, respectively) compared to the genome-wide average GC content of 0.41, while their flank sequences show the same trend but at significantly lower GC contents (0.37, 0.40, 0.46, respectively) (**Table S2**). These results are consistent with relative over- and under-amplification of low and high GC content regions, respectively. Refseq exons were also under-amplified by 35% on average in MDA single-neurons, consistent with the exome's higher GC content (0.49) relative to the genome-wide average. Bulk samples also show biases correlating with GC content, though to a lesser degree than MDA samples, while SW480 MALBAC single cells show greater biases than MDA samples. For example, SW480 bulk samples under-amplify SVA insertions and their flanks by 11% and 14%, respectively, while SW480 MALBAC single cells under-amplify these features by 89% and 75%, respectively. Overall, these correlations with GC content suggest that GC content accounts for some or perhaps most of the relative amplification biases of retrotransposon sequences annotated in the

human genome reference. However, as described above, these biases may not hold for the insertion sites and flanks of somatic insertions, due to their more random insertion sites. Still, the high GC content of AluY and SVA elements themselves would lead to some under-amplification regardless of the local GC content of the region surrounding the insertion site, since the phi29 polymerase of MDA would need to traverse them in order to achieve adequate exponential amplification of the region. Overcoming GC biases is highlighted by these analyses as a major goal for future single-cell genomics research.

III. MDA GC-content amplification bias

Since the above analyses and prior studies (Evrony et al., 2012; Hou et al., 2012) indicate that GC-content is a prominent source of bias in MDA amplification, we performed a detailed analysis of MDA GC bias in our high-coverage WGS samples. These analyses quantify the degree to which variation in coverage across the genome is explained by GC bias versus other factors and stochastic amplification noise. Distinguishing these sources of variation quantitatively helps understand the underlying sources of noise in MDA as an aid for development of future single-cell genome amplification methods.

Analysis methods

Genome coverage is affected by the underlying abundance of DNA from that genomic region in the sample and by ancillary factors relating to the amplification, sequencing and mapping of DNA fragments from the region. In standard bulk WGS, most coverage variation can be traced to differences in local GC content and differences in mappability across the genome (Benjamini and Speed, 2012). We compare the effect of these two factors between MDA and bulk DNA samples. Each sample was analyzed separately, and samples were also analyzed in sets grouped by sample type: MDA single-neuron (16 samples), MDA 100-neuron (1 sample), and bulk DNA (2 samples).

Genome coverage: We measure genome coverage in 100 kb equally-sized tandem bins across the genome by counting the number of DNA fragment 5' ends mapped to each bin. Only uniquely mapping reads with mapping quality score ≥20, and only reads mapping to the positive strand (to count each DNA fragment once), were counted. We model the mapping probability of each bin using simulated paired-end WGS data generated by the tool 'dwgsim' (http://sourceforge.net/projects/dnaa/files/dwgsim/) with an equivalent insert size to our experimental samples. Low-mappability bins, defined as bins with less than half of the genome-wide median coverage in the simulated WGS data, are excluded from the analysis. Coverage values for the X-chromosome were doubled for male samples to have the same copy number as autosomes so that the lower copy number of chromosome X does not bias coverage dispersion statistics; Y-chromosome bins were discarded due to low mappability. Overall, 85% of 100 kb bins (26,379 bins) were retained. Read coverage in remaining bins was median-corrected to stabilize comparisons across samples with different sequencing depths. In analyses of the single-neuron set and bulk DNA set, the median-corrected coverage for each bin was averaged across samples in the set.

GC curves: For each sample and sample set, we fit a smoothed regression curve, *f*, relating GC content (fraction of bases that are G or C) to coverage of each bin using the loess function in R

(span = 0.16). We quantify the magnitude of each GC curve by measuring how the curve deviates from the mean line. In the absence of GC content effects, the curve would be horizontal at the mean. Small deviations indicate a weak relation between GC content and coverage, whereas large deviations indicate a strong relation. We measure deviation of each curve by sampling at 100 points corresponding to the 0.5% ,…, 99.5% quantiles of the GC distribution of the 100kb bins $[q_{gc}(0.5\%) \approx .33, \dots, q_{gc}(99.5\%) \approx .57]$. The deviation of the GC curve at each of these points from the average across all 100 points, $\bar{f} = avg_{i=0.5}^{99.5} f(q_{gc}(i))$, was estimated by standard-deviation (SD) and total variation (TV) metrics (Benjamini and Speed, 2012):

$$SD_{GC} = \sqrt{\frac{1}{100}\sum_i \left(f\left(q_{gc}(i)\right) - \bar{f}\right)^2}, \quad TV_{GC} = \frac{1}{2\,\bar{f}}\frac{1}{100}\sum_i \left|f\left(q_{gc}(i)\right) - \bar{f}\right|.$$

SD measures dispersion in normalized counts, whereas TV measures the proportion of reads influenced by the stratification to GC categories. We measure the residual dispersion r_b for bin b by removing the GC effect from the normalized count. For (C_b, GC_b), the normalized count and GC of bin *b,* the residual $r_b = C_b - f(GC_b)$. SD_{resid} measures the dispersion of the residuals, and the proportion of coverage variability (out of the total variance) explained by GC equals $SD_{GC}^2/(SD_{GC}^2 + SD_{resid}^2)$.

Correction of coverage biases: Two methods to correct genome coverage variability are evaluated: a) a 'non-paired' correction method based solely on GC and mappability modeling of the sample being studied; this method is called 'non-paired' because it does not require any additional samples other than the sample being corrected; and b) a 'paired' correction method that uses additional different sample(s) of the same type as a reference.

In the non-paired correction method, the coverage correction factor for each bin in sample *s* was derived from the GC-curve $f_s()$, and further adjusted for the bin's mappability. The bin's mappability was estimated using the median-corrected fragment count of the bin in simulated WGS data (see above). Specifically, the correction factor $E_{s.b}^{gc}$ for bin *b* of sample *s* with GC-content GC_b and mappability M_b is $E_{s.b}^{gc} = f_s(GC_b) \cdot M_b$.

For the paired (reference sample-based) correction method, we evaluated two references for correction of single neuron MDA samples: a) the MDA 100-neuron sample, and b) a pool of all other MDA single neuron samples (15 neurons), excluding the single neuron being corrected. For the MDA 100-neuron reference, we use the median-corrected coverage, $E_{s.b}^{100n} = C_{100n.b}$, as the correction factor. For the pooled 15 MDA single-neuron reference, the correction factor is the average of median-corrected coverage across all 16 single neurons, after removing the count for the sample being corrected: $E_{s.b}^{pool} = \frac{16}{15}\left(avg\{C_{n1.b}, \dots, C_{n16.b}\} - \frac{1}{16}C_{s.b}\right)$.

For analyses in which samples were not corrected (i.e. uncorrected), the median coverage of the sample was used as the correction factor $E_{s.b}$ for all bins.

Corrected copy numbers for each bin *b* are then obtained by dividing the bin count in the sample *s*, $C_{s.b}$, by its correction factor $E_{s.b}$, stabilized by a small constant: $CN_{s.b} = \frac{C_{s.b}+\epsilon}{E_{s.b}+\epsilon}$, $\epsilon = 0.05$. Corrected copy numbers for each bin are then median-corrected by dividing by the median of all bins with CN ≥ 0.5 and $\log_2$-transformed. The median of all bins with CN ≥ 0.5 is used for this median-correction since in the presence of a large number of dropout bins, the median of non-dropout bins increases; excluding most dropout bins (CN < 0.5) prevents them from biasing the median correction so that the median of non-dropout bins is centered at CN = 1 ($\log_2$CN = 0).

Genome coverage variability statistics: Genome coverage variability after correction is evaluated using two metrics: median absolute pairwise deviation (MAPD) (Cai et al., 2014) and median absolute deviation from the median (MDAD). MAPD measures variability of bin copy numbers between adjacent bins, and is calculated as the median of the absolute differences in $\log_2$-copy number between every pair of adjacent bins across the genome. MDAD measures variability of bin copy numbers relative to the genome-wide median, and is calculated as the median of all absolute deviations of $\log_2$-copy numbers from the genome-wide median $\log_2$-copy number. MAPD and MDAD measure different aspects of genome coverage variability. MAPD captures stochastic bin-to-bin (i.e. between adjacent bins) noise of single-cell read depth variability, and is robust to true copy number changes and systematic noise on scales larger than the bin size, since large-scale systematic noise has less effect on copy number differences between adjacent bins. MDAD, on the other hand, evaluates each bin independently relative to the genome-wide median, so it is less sensitive to biases in coverage that are correlated across adjacent bins; it is therefore better suited to evaluate any residual noise after normalization, both stochastic and systematic. However, it is affected by true copy number variants. For both MAPD and MDAD, lower values indicate less variability.

Analysis of coverage variability and GC bias

Variation in coverage across different GC content is considerably larger in MDA samples compared to bulk DNA (**Figure S2A**): the average bin was 0.50 and 0.55 away from the mean in MDA 100-neuron and MDA single-neuron sets, respectively, but only 0.04 away in the bulk DNA set (in median-corrected coverage units). The increased variation can be mostly attributed to stronger GC effects. The standard deviation of the fitted GC curve from the mean (SD_{GC}) was 0.52 and 0.56 in MDA 100-neuron and MDA single-neuron sample sets, respectively, but only 0.04 in the bulk DNA set. This represents an approximately 15 times greater GC effect (measured by either SD or TV) in MDA samples than in bulk DNA. For example, in the MDA single-neuron sample set, we observe a 4.4x decrease in mean coverage between the 10% and 90% GC quantiles (GC content of 0.35 and 0.48, respectively). The MDA GC curves show negative slopes, and for high GC ranges (GC>0.55) coverage is almost 0. In contrast, the bulk DNA sample set has a 1.1x decrease in coverage over a similar range of GC content.

Furthermore, GC bias accounts for >70% of the total variation in coverage in MDA 100-neuron and MDA single-neuron sample sets, compared to 22% of the variation in bulk DNA. However, GC bias accounts for less of the total variation in individual MDA single-neuron samples (45% on average) than in the MDA single-neuron sample set (71%), since individual single-neurons have a greater proportion of their total variation stemming from non-GC (mostly stochastic noise) effects ($SD_{resid} \approx 55\%$ on average) that are averaged out when creating the MDA 16 single-neuron sample set (**Figure S2A**). Residual variance around the GC curve is greater in MDA single-neuron samples compared to bulk DNA and MDA 100-neuron samples (**Figure S2A**). Nevertheless, although coverage variation is higher in single-neuron MDA samples compared to the MDA 100-neuron sample, their GC curves are strikingly similar. This suggests that amplification of larger numbers of cells in an MDA reaction reduces stochastic variation, but not GC-related biases. Curves for individual MDA samples are also similar (data not shown), indicating that use of an MDA reference sample could correct for the systematic MDA amplification biases.

Mappability/GC-modeling (non-paired) correction versus reference sample-based (paired) correction

We compared non-paired coverage correction based solely on the GC and mappability modeling shown in the previous section, versus paired coverage correction that uses sample(s) of the same type as a reference without explicit GC or mappability modeling. In contrast to GC and mappability modeling-based correction, paired corrections using reference samples are less susceptible to modeling errors, and their accuracy improves as more samples are available to build a reference. In MDA 100-neuron and single neuron samples, both non-paired and paired corrections produce reasonable correction for variability at the 100 kb bin level (**Figures S2B-C**). This confirms GC as a prominent source of systematic bias in MDA WGS data. Additionally, we see improved correction for MDA single neurons with an MDA single neuron reference versus an MDA 100-neuron reference, though this may be due to the larger number of samples (n=15) used to build the single neuron reference than the MDA 100-neuron reference (n=1). However, the MDA 100-neuron reference outperforms any single-neuron reference consisting of only 1 neuron (data not shown). Both non-paired and paired correction methods reduced dispersion by >40% compared to uncorrected coverage. Nevertheless, paired correction consistently outperformed non-paired correction for all single neuron samples (**Figures S2B-C**). MDAD dispersion of single neurons decreases by 16% and 9% on average using a pooled single neuron and MDA 100-neuron reference, respectively, relative to GC/mappability correction. For the MDA 100-neuron sample, MDAD scores are comparable for both correction methods, but paired correction outperforms non-paired correction in terms of MAPD dispersion. We therefore choose to use paired corrections in the subsequent analysis.

IV. Coverage variability analyses

Since variability in single-cell genome coverage (i.e. deviations from even coverage of genomic loci) is a major factor affecting detection of somatic mutations, we performed further analyses of genome coverage variability in MDA samples to better understand the factors, both GC and non-GC, causing this variability. These factors were evaluated by testing the ability of additional correction methods employing various reference samples and bin sizes to correct for the variability, as well as the dependency of coverage variability on overall sequenced read depth. Finally, we compared genome coverage variability of MDA samples to MALBAC samples and studied the genomic spatial scales at which read depth variability manifests in each single-cell amplification method. The findings of these analyses aided in interpretation of sources of error and sensitivity loss during subsequent analyses of somatic retrotransposition, but can be broadly applied to other single-cell genomics analyses.

Analysis methods

The previous analysis of MDA GC-sequence bias ('*MDA GC-sequence bias*') showed that non-paired corrections based only on GC-content and mappability modeling do not perform as well as paired corrections that use MDA reference sample(s) for correction. The latter capture the effects of any shared systematic read-depth variability between the sample and the reference,

including both GC and non GC-related factors. Therefore, subsequent analyses used reference sample-based paired correction.

The previous analysis of GC-bias also counted read depth in tandem equally-sized bins. This was necessary for an unbiased analysis of GC bias but leads to significant variability in the expected number of reads per bin (i.e. variable Poisson sampling λ for each bin) due to GC content, mappability and other systematic factors affecting read depth. Maintaining an equal expected number of reads (λ) per bin is important so that each bin has the same expected variability in read depth by Poisson sampling. Therefore, in all the subsequent analyses in this section, bin boundaries were determined so that each bin has an equal number of mapped reads in the reference sample (Evrony et al., 2012; Navin et al., 2011). The resulting bins differ in size but maintain an equal expected read count (λ) for each bin, thereby controlling for mappability, GC, and any other systematic biases shared by the test sample and the reference sample. Specifically, we first filter both the test and the reference samples to keep only uniquely mapping reads with minimum quality score of 20, and only reads mapping to the positive strand (to avoid counting both reads of a read-pair that derive from the same DNA fragment). For each analysis, one chooses the desired total number of bins across the genome, B. The total number of post-filter reads, R, in the reference sample is then divided by the total number of bins, B, to determine the number of reads per equal-read bin, $r = R/B$. Reads are then consecutively counted in the reference sample across the genome beginning from chromosome 1, and bin boundaries are placed after every r reads to create equal-read bins. The average size of the equal-read bins is inversely proportional to the total number of bins, B, chosen for the analysis. For example, an analysis with B=6,000 equal-read bins has an average bin size of ~500 kb, and an analysis with B=60,000 equal-read bins has an average bin size of ~50 kb.

After defining equal-read bin boundaries using the reference sample, the number of reads mapping to each bin is counted for the test sample (e.g. the single-cell sample). The bin counts are then divided by the genome-wide median bin count across all bins of the sample to obtain a normalized relative copy number between the test sample and the reference sample, for each bin. Normalized relative copy numbers are then median corrected by dividing by the median of all bins with a relative copy number ≥ 0.5. This step is necessary, as described previously, due to bins with low relative copy number (dropout bins) artificially raising the genome-wide median above a relative copy number of 1. Relative bin copy numbers are then $\log_2$-transformed. The same two statistical measures of read depth variability used in the 'MDA GC-content amplification bias' section, MAPD (median absolute pairwise difference) and MDAD (median absolute deviation), were used to quantify the effects of different normalization methods. Each statistical measure reflects different aspects of variability as described above.

Analysis of reference sample choice for normalization of genome coverage variability

We evaluated the performance of the following WGS samples as references for genome coverage normalization in ~100kb (30,000 total) equal-read bins across the genome: a) simulated reads from the reference genome (simulated with the 'dwgsim' tool), which controls only for mappability; b) unamplified bulk DNA; c) MDA-amplified caudate nucleus 100-neurons; d) reads pooled from 8 MDA-amplified cortex single neurons; e) reads pooled from 15 MDA-amplified cortex single neurons (this analysis was performed only for chromosome 1 bins due to high computation and storage requirements), each time excluding the single neuron used as the

test sample. Note that in every analysis, the test sample was excluded from the set of samples used for the reference. MAPD and MDAD scores, as well as representative genome-wide coverage plots, for various sample type versus reference combinations are shown in **Figure S3**.

The results in **Figure S3** show that unamplified bulk DNA exhibited very even coverage and the least coverage variability among all sample types, with slightly lower variability using an unamplified bulk DNA reference compared to a simulated reads reference. In contrast, the MDA 100-neuron sample showed marked coverage variability with simulated read and unamplified bulk DNA references, but this variability was largely corrected using a pooled 8 MDA single-neuron reference. Interestingly, as was seen for unamplified bulk DNA samples, the MDA 100-neuron sample had lower variability as measured by MDAD (though less so MAPD) using an unamplified bulk DNA reference compared to a simulated read reference. This suggests some of the coverage variability in the MDA 100-neuron sample was due to differences in copy number (and perhaps other factors affecting mappability such as SNVs) specific to individual 1465 relative to the simulated human genome reference, and/or due to biases introduced during sequencing library preparation. Importantly, the MDA 100-neuron sample had significantly greater reduction in MAPD and MDAD variability measures when using an MDA single-neuron reference, indicating MDA itself accounted for most of the systematic variability in coverage in the MDA 100-neuron sample.

Finally, MDA single-neuron samples exhibited a similar pattern as MDA 100-neuron samples in terms of improved normalization (decreased coverage variability) with MDA versus unamplified bulk DNA or simulated read references. The pooled MDA single-neuron reference performed slightly better than the MDA 100-neuron reference, suggesting either the presence of additional systematic biases due to single-neuron amplification itself that are not fully controlled for by an MDA 100-neuron reference. Alternatively, this may be due to lower overall stochastic noise due to construction of the single-neuron reference from 15 pooled single-neurons versus only one 100-neuron sample, but this appears less likely since the pooled set of all 16 single neurons had a greater residual dispersion (SD_{resid}) than the MDA 100-neuron sample in previous GC bias analyses (see previous 'MDA GC-content amplification bias' section). Nevertheless, single-neuron samples retain a significant stochastic noise component, as seen by their increased MAPD and MDAD compared to the 100-neuron sample, when both are compared to an MDA reference. Overall, the above results indicate a significant systematic MDA noise bias, likely primarily related to GC-content based on the above GC analyses, that can be corrected to a large degree by MDA references. At the same time, single-neuron samples, and to a lesser degree MDA 100-neuron samples, suffer residual coverage variability after normalization, likely due to stochastic noise biases that cannot be controlled for by any reference. These results with high-coverage WGS data are consistent with previous low-coverage single-neuron sequencing data (Evrony et al., 2012). The MDA 100-neuron sample was used as the reference for all subsequent single-neuron coverage variability analyses.

Analysis of bin size for normalization of genome coverage variability

The effect of bin size on genome-wide coverage variability in each of the 16 WGS single-neurons was evaluated to assess at what size scales the systematic biases and stochastic noise of MDA manifest. Not surprisingly, equal-read bin analyses each using a different average bin size of ~10, 50, 100, 500, or 1,000 kb revealed decreasing dispersion from the median (MDAD) with

increasing bin size (**Figures S4A-B**). This is expected, as larger bins reduce the variability due to stochastic uneven amplification of the genome by averaging read depth across larger genomic scales; larger bins also better estimate systematic MDA biases that manifest at larger scales such as large-scale GC-content variation (see also '*Power spectral density analysis of genome coverage variability*' below). Note that in this analysis, decreased coverage variability with increasing bin sizes is not due to the higher read count per bin, λ, with increasing bin sizes. If the variability had been purely due to Poisson sampling error (i.e. Poisson-limited), then larger bin sizes with correspondingly larger read counts per bin, λ, would lead to decreased coverage variability due to a lower Poisson coefficient of variation (estimated as $1/\sqrt{\lambda}$). However, MDAD is stable at a given bin size across markedly different sequencing read depths, from 0.1x to 30x subsampled read depths (see '*Effect of total sequencing read depth on genome coverage variability*' below). This implies that the decreased variability with increasing bin size is not due to increasing read depth per bin, since just increasing read depth has little effect on coverage variability. Therefore, the variability in genome coverage in single-neuron samples is well above Poisson sampling variability, and the decreases in variability with increasing bin size are instead due to smoothing of stochastic MDA noise and better normalization for large-scale systematic MDA biases.

Interestingly, while MAPD estimates of coverage variability increased with decreasing bin size from 1,000 kb to 100 kb, MAPD decreased rather than increased with smaller bin sizes of 50 kb and 10 kb. Since MAPD measures variability between adjacent pairs of bins, bin sizes that are smaller than the size-scale at which the predominant systematic biases and stochastic noise manifest would in fact preserve concordance in copy number between adjacent bins. This indicates that a major component of single-neuron MDA coverage bias takes place at scales greater than ~50-100 kb. This is consistent with the power spectral density analysis below ('*Power spectral density analysis of genome coverage variability*').

Effect of total sequencing read depth on genome coverage variability

Each of the 16 WGS single-neuron samples was subsampled at genome-wide average read depths of 0.1x, 0.5x, 1x, 5x, 10x, 15x, 20x, 25x, and 30x, at both ~50 kb and ~500 kb equal-read bins. At each bin size, MAPD and MDAD measures of genome coverage variability were remarkably stable across all subsampled read depths, except for an increase in MAPD at the lowest, 0.1x, read depth (**Figures S4C-D**). The stability of variability across a large range of subsampled read depths indicates that in MDA single-cell samples, the stochastic noise component of coverage variability, as well as the variability inherent to the systematic MDA bias, are well above Poisson sampling variability. This leads to the important conclusion that systematic single-cell MDA bias and stochastic noise variability in genome coverage cannot be mitigated by increased sampling/sequencing above ~0.5x average genome coverage. MAPD and MDAD were reduced with 500 kb versus 50 kb bins at each subsampled read depth as seen in the previous '*Analysis of bin size for normalization of genome coverage variability*'.

Genome-wide coverage plots of all 16 WGS single neurons at full read depth, using the MDA 100-neuron reference and ~500 kb equal-read bins, are provided in **Figure S5**.

Comparison of MDA and MALBAC genome coverage variability

Genome coverage variability of the 16 MDA single neurons was compared to 3 MALBAC-amplified single cells (SRX202978, SRX204745, and SRX205035) at large genomic scales using ~100kb and ~500kb equal-read bins. In this analysis, each sample used a reference created by the same amplification method as the sample: the MDA 100-neuron sample was used as a reference for MDA single neurons, and merged data from 2 other MALBAC single cells (SRX202787 and SRX204744) was used as a reference for the 3 analyzed MALBAC single cells. MAPD and MDAD dispersion statistics, as well as coverage plots, showed clearly less coverage variability and more even coverage for MALBAC samples compared to MDA samples with both 100kb and 500kb bins (**Figure S6A**). MALBAC single-cells showed variability in quality, however, the best MALBAC single-cell sample (SRX202978) achieved the same dispersion statistics as the MDA 100-neuron sample (**Figures S3A and S6B**). An improved high-coverage MALBAC WGS reference based on a larger set of high-quality samples would likely further improve MALBAC coverage variability.

Importantly, although MALBAC samples show improved genome coverage variability at large scales (> 100kb), they exhibit significant coverage variability at small scales (< 10kb) with large peaks and troughs of coverage (**Figure S6C**). In contrast, MDA samples show significantly more even coverage at these scales (**Figure S6C**). The peaks and troughs of coverage are highly consistent in location and depth between MALBAC samples, explaining how MALBAC's highly variable coverage at small scales manifests as low variability at larger scales. Therefore, while MALBAC's even coverage at large scale makes it better suited than MDA for calling copy number variants by read depth analysis, MDA appears better suited for high coverage WGS structural variant analyses, such as detection of retrotransposon insertions, which relies on detection of discordant and breakpoint spanning reads that could be missed in troughs of MALBAC coverage. Overall, MALBAC appears to offer a trade-off versus MDA, with more evenness of coverage at large scales and more consistent capture of specific regions at higher coverage, but with reduced overall genome coverage (see previous section on 'Genome coverage') and more variability of coverage at small scales. Therefore, MDA and MALBAC offer different benefits depending on the application and are in fact complimentary approaches. Further high-coverage WGS of each method will be helpful in understanding their relative benefits and building new amplification methods able to achieve the advantages of both.

Power spectral density analysis of genome coverage variability

In order to further characterize genome coverage variability at different genomic scales, we performed a power spectral density analysis of genome coverage variability across chromosome 1 in one representative sample from each sample type of individual 1465 (bulk DNA, MDA 100-neuron, MDA single-neuron), as well as a MALBAC single-cell and its corresponding SW480 bulk DNA. These power spectra show the degree to which variability in read depth is distributed over different genomic scales (frequencies) in a more comprehensive manner than the above genome-wide coverage plots. A similar analysis was previously performed for MALBAC samples (Zong et al., 2012), and is useful in comparing different sequencing technologies' ability to maintain read depth evenly over different genomic scales.

The power spectra for all sample types (**Figure S7**) revealed greater read depth variability at larger genomic scales (i.e. smaller frequencies) compared to smaller genomic

scales. This is likely due to large-scale variations in sequence composition such as GC-content (for example, isochores and chromosome bands) that bias read depth because of GC-bias of sequencing library preparation. Furthermore, at these large genomic scales, MDA samples had greater variability in read depth compared to bulk DNA samples, likely due to the greater influence of GC content on MDA read depth variability compared to sequencing library preparation alone. At large genomic scales, the MALBAC single-cell sample had a read depth variability intermediate between bulk and MDA samples, consistent with MALBAC's better read-depth stability at large genomic scales (**Figures S6A-B**) and better performance than MDA in calling large (megabase-scale) copy number variants (Hou et al., 2013). At the larger genomic scales, the MDA 100-neuron sample also showed less variability than the MDA single-neuron sample, as expected due to less stochastic read depth variability in 100-cell samples compared to single-cell samples.

At smaller genomic scales, below a frequency of ~$3.5 \cdot 10^{-5}$ bp (i.e. a scale of ~30kb), MDA samples showed significantly less read depth variability than the MALBAC single-cell sample, which had a distinct peak of increased variability at these scales. Below a scale of ~10kb, MDA samples were mostly concordant with bulk DNA. These results are consistent with the high-resolution coverage visualization at smaller genomic scales of MDA and MALBAC single-cell samples that show more even coverage for MDA samples compared to MALBAC samples, which have periodic peaks of very high coverage separated by low or absent coverage troughs (**Figure S6C**).

V. Whole-genome sequencing analysis of MDA chimeras

Chimeras are false positive structural variants created during single-cell amplification, which pose a challenge for single-cell sequencing studies of somatic structural variants, including retrotransposons (Evrony et al., 2012; Lasken and Stockwell, 2007). Although they are generally present at a lower signal level than true structural variants, they can in some cases be difficult to differentiate from true variants (Evrony et al., 2012). Our large WGS dataset of single-cell MDA samples presented a unique opportunity to study in detail the chimeras created by MDA in order to further understand their mechanism of formation, the fundamentals of which have been previously elucidated by Lasken and Stockwell (Lasken and Stockwell, 2007). A better understanding of MDA chimeras will in turn assist future single-cell studies to model and control for their effects, as well as aid in developing improved single-cell amplification methods.

Analysis methods

We studied MDA chimeras by extracting non-PCR duplicate discordant read pairs aligning to the autosomes and sex chromosomes, in which both reads of the pair aligned to the same chromosome within 100kb of each other. This excluded the large majority of chimeras that arise during sequencing library preparation, since sequencing library preparation chimeras occur after DNA fragmentation such that they are much more likely by chance to be due to ligation of fragments from different chromosomes (inter-chromosomal chimeras) or fragments from the same chromosome but from large mapping distances from each other. In contrast, MDA chimeras occur prior to DNA fragmentation and have been shown to occur predominantly in an

intra-chromosomal, local manner (Evrony et al., 2012; Lasken and Stockwell, 2007). Further supporting the assumption that most library preparation chimeras are inter-chromosomal and more numerous than MDA chimeras, are: a) inter-chromosomal discordant reads are more numerous than intra-chromosomal discordant reads in both bulk and MDA samples; b) the percentage of discordant reads is similar in bulk and MDA single-neuron samples from the same sequencing library preparation batch; and c) the percentage of inter-chromosomal discordant reads is similar between these sample groups (see above 'Alignment statistics' section for details; and **Table S1**). As described below, we also subtract bulk sample chimera rates from MDA sample chimera rates in our analysis to further correct for local intra-chromosomal sequencing library chimeras in MDA samples. Therefore, our analysis limited to local intra-chromosomal chimeras is expected to account for the large majority of MDA chimeras and excludes most sequencing library chimeras. This was confirmed definitively by the subsequent analyses (see below) that revealed the characteristic signatures of MDA chimeras, as well as new information regarding their properties and mechanisms of formation.

Next, the extracted discordant read pairs were grouped into inversion, deletion, and duplication chimeras, according to their reads' orientations (**Figure S8A**). Discordant read pairs of each chimera type were then grouped by the distance between the 3' ends of the reads of the pair (in 50bp bins). The distance between the 3' ends of the reads serves as the best estimate of the chimera breakpoint distance, though not a precise estimate since the true insert sizes of chimera read pairs are not known. One read of each read pair was included in the analysis to avoid double-counting chimeras. Inversion chimeras where both reads overlap with breakpoint distance < 3bp were excluded as these derive from sequencing artifacts. The number of read pairs for each chimera type and bin was normalized to the total number of extracted chimera read pairs plus concordantly mapping read pairs (non-PCR duplicate, aligning to the autosomes or sex chromosomes). This normalization excludes most library chimeras from the denominator.

Histograms of each chimera type versus breakpoint distance are plotted in **Figure S8B**. Initial plots revealed peaks of deletion and duplication chimeras in unamplified bulk DNA samples with periodicity of ~1800bp. Further investigation revealed these peaks derive from highly repetitive satellite regions of the genome, mostly centromeric ALR/Alpha and HSATII satellites. The repetitive and polymorphic nature of these regions, some of which include tandem repeats with periods of 1868bp (e.g. chr8:43,820,851-43,838,887, hg19) leads to these peaks. Not surprisingly, these peaks are absent in MDA samples, consistent with our prior finding that single-neuron MDA does not amplify well heterochromatic satellite regions (see 'Alignment statistics' section above). In order to obtain a more accurate analysis of MDA chimeras without discordant reads deriving from satellite regions, we excluded read pairs aligning to the 9,566 satellite regions in the genome annotated by RepeatMasker in hg19 (13.4Mb of the genome). As predicted, the periodic peaks in deletion and duplication chimeras disappeared after filtering satellite reads (**Figure S8C**), proving that these derived from reads aligning to satellite regions.

Finally, the unamplified bulk DNA baseline was subtracted from single-neuron and 100-neuron plots, in order to eliminate remaining discordant reads that stem from germline polymorphisms, library preparation chimeras, and alignment artifacts. The resulting plots reveal the breakpoint distance distribution of local MDA chimeras (**Figure S8D**).

Chimera quantification and distribution

Inversion chimeras accounted for 96% of all MDA chimeras, exhibited a peak at 250bp, and exponentially decreased to background at a breakpoint distance of ~10kb (**Figure S8D**). Inversion chimeras comprise 0.35% ± 0.03% (SD) and 0.23% of the total read pairs in single-neuron and 100-neuron samples respectively (**Figure S8D**; **Table S1**). The ratio of inversion chimeras to other chimeras (duplications and deletions) is 23.6 and 25.8 in single-neuron and 100-neuron samples, respectively.

Deletion chimeras were far fewer in number, accounting for 0.6% and 0.4% of MDA chimeras in single-neuron and 100-neuron samples, respectively. Deletion chimeras exhibited a modest peak from ~1-3kb, gradually decreasing to background at ~10kb. Deletion chimeras comprise 0.002% ± 0.0005% (SD) and 0.0009% of all read pairs in single-neuron and 100-neuron samples respectively (**Figure S8D**; **Table S1**).

Duplication chimeras were also much less frequent than inversion chimeras, accounting for 3.4% and 3.3% of MDA chimeras in single-neuron and 100-neuron samples, respectively. Duplication chimeras peak at ~1.7-1.8kb and decrease exponentially to background by ~30kb. Duplication chimeras comprise 0.012% ± 0.001% (SD) and 0.008% of all read pairs in single-neuron and 100-neuron samples respectively (**Figure S8D**; **Table S1**).

The estimated frequency of chimeras per 100kb of DNA produced by MDA (calculated as [fraction of chimera reads] / [average library insert size] * 100,000) was 1.2 chimeras/100kb and 0.8 chimeras/100kb in single-neuron and 100-neuron samples, respectively. The frequency of inversion chimeras was 1.1/100kb and 0.8/100kb, in single-neuron and 100-neuron samples, respectively. The frequency of deletion chimeras was 0.007/100kb and 0.003/100kb, in single-neuron and 100-neuron samples, respectively. The frequency of duplication chimeras was 0.04/100kb and 0.03/100kb, in single-neuron and 100-neuron samples, respectively. See **Table S1** for per sample statistics.

Comparison to prior studies

Overall, these results are remarkably consistent with both our prior study of chimeras in targeted L1 insertion-profiling (L1-IP) of single-neurons amplified by MDA (Evrony et al., 2012), and a study of chimeras by Lasken and Stockwell in single bacteria amplified by MDA (Lasken and Stockwell, 2007). In the L1-IP study, chimeras were quantified by counting the number of reads in the 20 kb flanks of germline L1Hs insertions. Lasken and Stockwell quantified chimeras by searching for 454 sequencing reads with 2 segments aligning to non-contiguous locations in the genome.

The fraction of reads that were chimeras in L1-IP was 0.3% and 0.2% in single-neuron and 100-neuron samples, respectively, and the fraction of reads that were chimeras in single-bacteria amplified by MDA was 0.45% (Evrony et al., 2012; Lasken and Stockwell, 2007). These are strikingly similar to each other and to the single-neuron WGS estimate of 0.35% and 0.23% in single-neuron and 100-neuron samples. The slight increase in single bacteria MDA chimeras may be due to larger insert sizes in 454 sequencing libraries, since the fraction of chimera reads would increase with larger insert size. Moreover, the chimera frequency estimated in single bacteria MDA was 0.87 per 100kb (Lasken and Stockwell, 2007) (calculated assuming an average 454 library insert size of 500bp; note, the authors' estimate of 1 per 22kb was calculated using the 100bp read length rather than a calculation using insert size), compared to

1.2 per 100kb in single-neuron WGS samples. An additional finding replicated in single-neuron WGS that was previously seen in L1-IP (Evrony et al., 2012), is the slight increase in total chimera reads in single-neuron versus 100-neuron samples (**Figure S8D**). The reason for this is unclear, since if MDA chimeras occur at a given rate per length of amplified DNA (i.e. ~1 per 100kb), then the proportion of chimeras as a fraction of total DNA should remain constant regardless of the initial input DNA. One hypothesis is that this difference may be due to the relative molar increase in random hexamer per amount of input DNA in single-neuron samples, leading to an increased number of amplicons simultaneously replicating a given template, thereby leading to an increased rate of chimeras.

Similarly to single-neuron WGS, both prior studies also found a predominance of inversion chimeras. 85% of single-neuron L1-IP chimeras and 85% of single-bacteria MDA chimeras were inversions (Evrony et al., 2012; Lasken and Stockwell, 2007), compared to 96% in single-neuron WGS. The reason for the relatively higher proportion of inversion chimeras out of all chimeras in WGS data is unclear, but may be due to loss of sensitivity for deletion chimeras smaller in size than the minimum detectable by WGS discordant read analysis, since read pairs falling within the observed variation in insert size are annotated as concordant by the alignment software. Moreover, both L1-IP and single-neuron WGS found that duplication chimeras occur more frequently than deletion chimeras. Of single-neuron L1-IP chimeras, 12% were duplications versus 3% that were deletions (Evrony et al., 2012), an ~4-fold difference. Single-neuron WGS found 3.4% and 0.6% duplication and deletion chimeras, respectively, an ~6-fold difference. The differences between L1-IP and single-neuron WGS may be due to some loss of sensitivity for deletion chimeras in single-neuron WGS, as described above.

The distributions of chimeras versus breakpoint distance are also similar among the studies, though WGS provides a much clearer picture of these distributions due to its higher throughput. L1-IP chimera distributions (aggregated across all chimera types) decrease to background by ~15kb breakpoint distance, and single-bacteria MDA inversion chimeras decrease to background by ~10kb (Evrony et al., 2012; Lasken and Stockwell, 2007). As described above, single-neuron WGS inversion and deletion chimeras similarly decrease to background by ~10kb, but interestingly duplication chimeras have a longer tail and decrease to background by ~30kb. Moreover, WGS revealed that the peaks of inversion chimera distributions versus deletion and duplication chimera distributions occur at different breakpoint distances, with the latter peaking at larger breakpoint sizes. These WGS results are in contrast to single-bacteria chimera analyses that found no dependence on distance for local deletion and duplication chimeras, termed 'direct' chimeras in that study (Lasken and Stockwell, 2007). The discrepancy is likely due to the increased sensitivity of WGS, with more chimera reads captured compared to the lower throughput of 454 sequencing. Consistent with WGS, both L1-IP and single-bacteria MDA showed local (<20kb) enrichment of deletion and duplication chimeras (Evrony et al., 2012; Lasken and Stockwell, 2007).

One aspect of chimera formation that L1-IP resolves, which the current WGS and prior single-bacteria sequencing studies cannot address, is the proportion of inversions occurring to a location upstream versus downstream of the 3' single stranded DNA end that is priming the chimera. The current WGS sequencing libraries and 454 sequencing libraries used in the single-bacteria MDA study (Lasken and Stockwell, 2007) clone DNA fragments in a random orientation relative to the sequencing adaptors. Although the prior single-bacteria study performed a comparison between upstream and downstream inversions, standard sequencing

library preparations do not allow one to distinguish upstream from downstream inversions. This is because a read appearing as an upstream inversion when sequenced in one direction would appear as a downstream inversion when sequenced in the opposite direction, and vice versa. In contrast, L1-IP amplifies DNA fragments in a directional manner due to its L1-specific primers. Re-analysis of prior L1-IP chimera data (Evrony et al., 2012) revealed that 45% of inversions occur to an upstream locus and 55% of inversions occur to a downstream locus in single-neuron samples. This is close to an equal probability of upstream versus downstream inversions.

Refined model for MDA chimera formation

Overall, the above results propose a refined model for local MDA chimera formation (**Figure S9**) that is based on the model first described by Lasken and Stockwell (Lasken and Stockwell, 2007). The features of this model are:

1) 3' single strand ends (termed source strands) are liberated from the template strand to which they are annealed by reannealing of the downstream displaced single-stranded amplicon (termed the reannealing strand) via the branch migration mechanism proposed by Lasken and Stockwell (Lasken and Stockwell, 2007) (**Figure S9A**).

2) Branch migration between the source strand and the reannealing strand is likely a stochastic process able to proceed in both directions, either towards or away from the source strand. Additionallly, branch migration would be able to proceed back towards the source strand only up to the point where the reannealing strand is single-stranded. Branch migration is not energetically favorable beyond the point where the reannealing strand has itself been annealed to by a random hexamer initiating replication of the reannealing strand into a double-stranded form. This is because branch migration beyond this point would require denaturation of two double stranded DNAs with annealing of only one double stranded DNA (**Figure S9A**).

3) About once every 100 kb of synthesized DNA, a source strand (free 3' single stranded end) primes on a nearby single strand (termed the target strand), likely mostly mediated by microhomologies (Evrony et al., 2012; Lasken and Stockwell, 2007). Since random hexamers are significantly more numerous and more free to diffuse than source strands, most nearby displaced strands would be expected to be double-stranded, and therefore not available for annealing of the source strand. This, in addition to the need for microhomology, means that source strand annealing to single-stranded target strands is the rate-limiting step in chimera formation and explains why chimeras are relatively rare events.

4) The distance that branch migration extends determines the length of the displaced source strand and in turn the distribution of chimera breakpoint distances, since the farther branch migration extends, the farther the source strand can reach to more distant target strands (**Figure S9B**). The branch migration distance distribution is related to the inversion chimera distribution, which has a peak at ~250bp and maximum extension to ~10kb. However, the true branch migration distance cannot be determined from this data since bending of the template strand can allow source strands to reach target strands more distant than the length of the displaced source strand.

5) The reannealing strand is not preferentially used as the target strand, because if this were the case, upstream inversions would be more frequent than downstream inversions. Similarly, the source strand itself is not preferentially used as a target strand (either by a hairpin mechanism or by annealing to its own displaced 5' end). This aspect of the model, also previously noted by Lasken and Stockwell (Lasken and Stockwell, 2007), can be explained by the same reasoning as in (3). Namely, source strand annealing is less frequent than random hexamer annealing, so the reannealing strand and displaced source strand would usually be converted to double-stranded DNA before source strand priming can occur. Note also, that as described above, only L1-IP (Evrony et al., 2012) data is able to exclude preferential priming to the reannealing strand or to the source strand, since sequencing library cloning of DNA fragments in random orientation in the current WGS study and the single-bacteria MDA study (Lasken and Stockwell, 2007) cannot distinguish upstream from downstream inversions.

6) Most chimeras are inversions, which form by priming to upstream or downstream target strands that were displaced off the same template strand to which the source strand molecule is annealed (**Figure S9C**). There is about equal probability of annealing to upstream versus downstream target strands, for the same reasons as described above: most displaced single-strands are double-stranded due to the lower efficiency of source-target strand annealing relative to random hexamer annealing.

7) Deletion chimeras are the least frequent and could occur either by: a) priming to target single strands formed by displacement of amplicons of downstream amplicons of the template strand; or b) priming downstream on the template strand to single-stranded DNA (**Figure S9D**). The latter mechanism would be less frequent because most of the downstream template DNA would be double stranded.

8) Duplication chimeras are less frequent than inversions, and could occur either by: a) priming to target single strands formed by displacement of amplicons of upstream amplicons of the template strand; or b) priming upstream on the template strand to single stranded DNA (**Figure S9E**). The former mechanism can explain the duplication chimera breakpoint distance distribution peaking at larger distances (~1,800bp) and extending farther (up to ~30kb) compared to inversion chimera breakpoints, because the source strand is an amplicon of the template strand whereas the target strands are amplicons of amplicons of the template strand. The latter mechanism would be less frequent because most of the upstream template DNA would be double stranded.

Unresolved aspects of the model include:

1) The reason for the higher frequency of duplication versus deletion chimeras is unclear, as both could occur by the same mechanisms differing only in that the former have upstream target strands and the latter have downstream target strands. An asymmetry stemming from the fact that the branch migration mechanism liberates more single-stranded source strand as it progresses upstream, could explain the preference for upstream target strands. However, if this were the case, then upstream inversion chimeras would be more frequent than downstream inversion chimeras, which is not observed by L1-IP.

2) The relationship between the branch migration distance distribution and the observed chimera breakpoint distance distribution requires confirmation and further clarification.

3) Related to this, will be determining why the peak of inversion and duplication/deletion chimeras are specifically at 250bp and ~2kb, respectively.

4) Confirmation of our model's proposal that self-priming (hairpin) and priming to the reannealing strand do not occur frequently due to the higher efficiency of random hexamer priming and extension relative to branch migration, thereby rapidly converting free single strands to double strands that prevent chimera formation.

Supplemental Note 2 - Somatic mutation of poly-A microsatellites

Several general features of microsatellite mutation processes have been outlined by prior studies (Arcot et al., 1995; Brinkmann et al., 1998; Chakraborty et al., 1997; Ellegren, 2004; Kelkar et al., 2011; Kelkar et al., 2008; Manley et al., 1999; Parsons et al., 1995; Pearson et al., 2005; Sun et al., 2012; Whittaker et al., 2003) that are relevant to interpretation of the somatic retrotransposon poly-A microsatellite mutations we observed. Retrotransposon-derived poly-A microsatellite mutation in particular has also recently been reviewed (Grandi and An, 2013). The general features of microsatellite mutation are: a) mutation rates increase exponentially with increasing microsatellite length (Brinkmann et al., 1998; Ellegren, 2004; Grandi and An, 2013; Kelkar et al., 2008; Sun et al., 2012; Whittaker et al., 2003); b) long microsatellites are generally biased towards truncating mutations (Ellegren, 2004; Grandi et al., 2012; Sun et al., 2012; Whittaker et al., 2003), consistent with the observation that L1 and SVA poly-A tails truncate over evolutionary time with older elements having shorter poly-A tails (Ovchinnikov et al., 2001); c) shorter repeat motifs and motifs with lower GC-content have higher mutation rates (though some triplet repeats can have high rates of mutation with a preference towards expansion for reasons that are not fully understood) (Chakraborty et al., 1997; Ellegren, 2004; Grandi and An, 2013; Kelkar et al., 2008); d) most mutations are single-step mutations (i.e. "stutter" mutations of one repeat unit), while a minority of events are multi-step mutations (i.e. mutations of > 1 repeat unit); however, multi-step mutations become increasingly common for shorter repeat motifs (Brinkmann et al., 1998; Ellegren, 2004; Sun et al., 2012; Whittaker et al., 2003); e) polymerase slippage in conjunction with mismatch repair machinery is the major mechanism of mutation (Ellegren, 2004; Manley et al., 1999; Parsons et al., 1995; Pearson et al., 2005); f) mutation rates can differ between genomic loci due to differences in local sequence contents, chromatin structures, local point mutation rates that can disrupt microsatellite repeats, presence inside retrotransposons, and other factors (Ellegren, 2004; Kelkar et al., 2008). Interestingly, retrotransposons themselves are major generators of microsatellites in the genome, having created almost all poly-A tails in the genome (see also analysis below), and likely significant contributors to other classes of microsatellites as well (Arcot et al., 1995; Grandi and An, 2013; Kelkar et al., 2011).

Taking the above features of microsatellite mutation into account suggests that poly-A microsatellites, and in particular long poly-A microsatellites created during somatic retrotransposition, are expected to have one of the highest mutation rates of any class of microsatellites (Grandi and An, 2013), with a bias towards truncating mutations and a relatively increased rate of large multi-step mutations. It is therefore not surprising that all pure poly-A tails in the human genome reference are shorter than the original poly-A tails created by the somatic brain retrotransposon insertions found in this study (see analysis below). Furthermore, since microsatellites themselves have some of the highest mutation rates of any type of sequence element in the genome (Ellegren, 2004; Sun et al., 2012), long poly-A tails of somatic retrotransposon insertions may in fact have the highest mutation rate of any sequence in the genome. These general features of microsatellite mutation suggest the following model for the mutational path of L1#1's poly-A tail: the originating poly-A tail may have been 250 bp (the largest poly-A tail we cloned; **Figure 4D**) or larger, which due to its very large size underwent several early, large multi-step truncating mutations giving rise to the main sub-lineages represented by peaks in the poly-A length distributions. The sub-lineages created by early large multi-step mutations were then further diversified by numerous single-step and other smaller multi-step mutations. The poly-A tail of L1#2 also underwent many large multi-step mutations

that were subsequently diversified by stutter and smaller multi-step mutations, but at significantly lower rates relative to L1#1, likely due to a shorter poly-A tail of the originating insertion (since mutation rates decrease exponentially with decreasing microsatellite length) and/or due to differences in local genomic features influencing the mutation rate.

We performed two additional analyses in order to further characterize the relationship between poly-A microsatellites and retrotransposons, and to understand the factors determining the poly-A tail lengths of the originating retrotransposon insertion events: an analysis of poly-A microsatellite distribution and lengths in the genome relative to retrotransposons, and an analysis of the poly-A lengths of retrotransposon RNA transcripts.

Poly-A microsatellite distribution and lengths in the human genome reference

Over one million microsatellite loci are present in the human genome, accounting for ~3% of the genome (Ellegren, 2004; Lander et al., 2001). An analysis of all microsatellites in the human genome reference with repeat motifs of lengths 1 to 6, as annotated by Tandem Repeats Finder (TRF)(Benson, 1999) in the UCSC genome browser (inclusion criteria: loci with minimum score of 50, where score = 2×match - 7×mismatch - 7×indel, i.e. minimum poly-A included is 25bp), shows that poly-A microsatellites are the most abundant microsatellite when ranking by total number of loci (**Figure S19A**), with ~70,200 loci. When ranking by the total number of bases, poly-A microsatellites account for ~2.2 Mb of the genome, ranking 3rd after AC and AT microsatellites (**Figure S19B**). Note that these are underestimates of the number of loci and genomic bases due to the minimum score cutoff used by TRF. Importantly, a comparison of the locations of poly-A microsatellites to L1, Alu, and SVA retrotransposon loci (including all subfamilies for each) showed that 86% of poly-A loci overlap an L1, Alu, or SVA element (defined as overlap over at least half the length of the poly-A), with 3.5%, 82%, and 0.2% overlappping L1, Alu, and SVA, respectively. Additionally, 80% of TRF-annotated poly-A bases in the genome reside within L1, Alu or SVA elements, supporting retrotransposons as the source of most poly-A microsatellites (Grandi and An, 2013).

The distributions of poly-A lengths in the genome for loci overlapping L1, Alu, or SVA elements are similar, though L1- and SVA-derived poly-A tails show trends for slightly longer and more variable sizes, respectively, compared to Alu-derived poly-A tails (**Figures S19C-D**). The size distribution of all poly-A tails extends to larger sizes compared to the distribution of a subset of more pure poly-A tails that have ≥95% A's (**Figures S19C-D**). The largest annotated poly-A in the human genome reference is 415 bp, and the largest poly-A loci overlapping L1, Alu and SVA (over at least half the length of the poly-A) are 174 bp, 84 bp, and 44 bp, respectively. The maximum size poly-A microsatellites with ≥95% purity in the human genome reference are 90 bp (genome-wide), and 83bp, 72bp, and 38bp overlapping L1, Alu, and SVA, respectively. This is consistent with pure poly-A loci undergoing more rapid shortening. Overall, this analysis shows that the poly-A tails of L1#1 and L1#2 are significantly longer compared to the distribution of genome-wide poly-A lengths and in fact longer than any annotated pure poly-A tails in the human genome reference.

Retrotransposon transcript poly-A tail lengths

The poly-A tail length inserted into the genome during retrotransposition is to some degree determined by the poly-A tail length of the original retrotransposon RNA transcript. The distribution of retrotransposon transcript poly-A tails has to our knowledge not previously been studied. To obtain a view of the poly-A lengths of retrotransposon transcripts, we analyzed data from the first transcriptome-wide poly-A length (PAL) profiling study (Subtelny et al., 2014). Retrotransposon transcripts were successfully identified in PAL data of human cell lines profiled in this study (HeLa and HEK293T), with 8, 1495, and 34 L1Hs, AluY, and SVA transcripts found, respectively. The 8 observed L1Hs poly-A lengths were: 26, 42, 69, 80, 116, 123, 170, and 174 bp. The median poly-A lengths of all L1Hs, AluY, and SVA transcripts were 98, 41, and 95bp, respectively. 12% of Alu poly-A lengths were >150bp, and 6% were >200bp. 26% of SVA poly-A lengths were >150bp, and 3% were >200bp. The large number of AluY transcripts allowed us to further compare the AluY poly-A length distribution to the poly-A length distribution across the entire transcriptome of these cell lines (~4.8 million poly-A tails profiled from ~10,000 genes). The 25% and 75% quantiles of AluY poly-A lengths were 15bp and 85bp, respectively. In contrast, the median poly-A length across the entire transcriptome for these cell lines was 75 bp, with the 25% and 75% quantiles at 38 and 126bp, respectively. 17% of poly-A lengths across the entire transcriptome were >150bp, and 6% were >200bp. The smaller median and quantiles of AluY versus transcriptome-wide poly-A lengths, as well as histograms of their poly-A length distributions (data not shown), show a shift of the AluY distribution towards shorter poly-A lengths, though with a similar proportion of transcripts with large (>200bp) poly-A tails. The small number of L1Hs and SVA transcripts preclude conclusions regarding how their distributions compare to the transcriptome-wide distribution.

The largest genomic poly-A tails cloned for L1#1 and L1#2 in our study were 250bp, and 116bp, respectively, which correspond to the 98th and 71st percentiles, respectively, relative to the transcriptome-wide poly-A length distribution. Therefore, the original transcript of L1#1 likely had a significantly longer poly-A tail than the transcriptome-wide median, although transcriptome-wide PAL has not yet been performed on neuronal progenitors. The original transcript of L1#2 was also likely above the median transcriptome-wide poly-A length, though still within the size range observed for L1Hs transcripts in PAL profiling. Prior studies have indicated that poly-A tails play important roles in L1 and Alu retrotransposition (Grandi and An, 2013), and have shown that Alu activity increases with increasing poly-A length (Dewannieux and Heidmann, 2005). Therefore, it is possible that retrotransposon transcripts with longer poly-A tails, such as L1#1 and L1#2, may have higher retrotransposition activity and preferentially integrated into the genome so that poly-A tails of somatic insertions would tend to be larger than the average retrotransposon transcript poly-A tail. However, this hypothesis will require further testing. Additionally, the size of the retrotransposon transcript poly-A tail does not necessarily equal the size of the poly-A tail inserted into the genome, as internal priming and slippage can occur during reverse transcription to either lengthen or shorten the poly-A tail (Srikanta et al., 2009; Wagstaff et al., 2012). Moreover, the poly-A tail length of any given transcript is dynamic, so that although transcripts are believed to begin in the nucleus with a poly-A tail of ~250bp (Kuhn et al., 2009) that is similar to the largest genomic poly-A we found for L1#1, the poly-A tail length of the retrotransposon transcript when it re-enters the nucleus is likely different. Therefore, while the above results suggest that the somatic L1Hs retrotransposons we found originated from L1Hs transcripts with longer than average poly-A tails relative to the transcriptome-wide average, we cannot estimate for certain the original transcripts' poly-A tail lengths. The above questions regarding the role of poly-A tail length in somatic retrotransposon

activity may begin to be addressed as more somatic retrotransposon insertions are identified in future studies, with concomitant profiling of transcriptome poly-A tail lengths from matched cell types.

Supplemental Experimental Procedures

I. Human tissues and DNA samples

Post-mortem tissues used in this study from individual UMB1465 (left and right cerebral hemispheres, caudate, cerebellum, spinal cord, heart, lung, liver) were obtained from the NIH NeuroBioBank at the University of Maryland (Baltimore, MD). UMB1465 was a 17 year-old male who died in a motor vehicle accident and was one of the three individuals profiled in our previous single-neuron L1 insertion-profiling (L1-IP) study (Evrony et al., 2012). All UMB1465 tissues we studied, except for the right cerebral hemisphere, were obtained, processed, frozen, and stored at -80°C without fixation within 4 hours of death, according to a standardized protocol: http://medschool.umaryland.edu/btbank/method2.asp. The left cerebral hemisphere was sectioned prior to freezing in 17 coronal sections ~1cm each in thickness. The right cerebral hemisphere was formalin-fixed, similarly sectioned, and stored at room temperature. All studies of UMB1465 brain were performed on frozen unfixed tissues, except for attempts to detect L1#1 and L1#2 in the formalin-fixed right cerebral hemisphere using nested PCR (see 'Assays for somatic L1s in formalin-fixed tissues'). Right hemisphere formalin-fixed tissue was not amenable to other DNA assays, such as ddPCR, due to low efficiency of DNA extraction (data not shown).

Bulk DNA was extracted from frozen tissues with the QIAamp DNA Mini kit with RNase A treatment (Qiagen). Bulk DNA was extracted from formalin-fixed paraffin-embedded tissues of the right hemisphere with the QIAamp DNA FFPE Tissue kit, with 2 additional xylene incubations (one incubation until paraffin dissolves and a second 5 min. incubation) and an additional 100% ethanol wash, which increased DNA yield as assayed by PCR for control genomic regions. The cerebral cortex bulk DNA from individual 1465 used for whole-genome sequencing was obtained from location D (**Figures 3B and S14**). Genomes of the 16 cerebral cortex single neuron samples and the caudate nucleus 100-neuron sample were amplified by MDA (Dean et al., 2002) as part of our previous targeted L1 insertion-profiling (L1-IP) study (Evrony et al., 2012), which provided a large amount of DNA suitable for whole genome sequencing. The neurons were sorted from the brain in the previous study (Evrony et al., 2012) using NeuN as a neuronal marker. The 16 single neurons were originally sorted from location D of the middle frontal gyrus (**Figures 3B and S14**). The first 4 WGS single neurons we sequenced were chosen to include single neurons 2 and 77 as positive controls for identifying somatic retrotransposon insertions (i.e. both were found to have L1#1 in our prior targeted L1 insertion profiling (Evrony et al., 2012)) and 2 other neurons were randomly chosen from the set of 6 cells previously sequenced by low coverage sequencing (Evrony et al., 2012). The remaining 12 WGS single neurons were chosen among the single neurons with the highest number of known reference L1 insertions with score >0.8 identified by L1 insertion profiling in our prior study (Evrony et al., 2012). Nuclei of non-neuronal (NeuN-negative) cells from location D were purified and amplified by MDA as previously described (Evrony et al., 2012).

Unamplified bulk DNA from a breast cancer primary tumor (ID: TCGA-E1-A15E-01A), lymph node metastasis (ID: TCGA-E1-A15E-06A), and normal blood (ID: TCGA-E1-A15E-10A) from an individual were obtained with permission from The Cancer Genome Atlas (TCGA) project. The primary tumor was an $ER^{+}PR^{+}HER2^{+}$ invasive ductal carcinoma, grade 3, of the left breast. The metastasis was in the sentinel left axillary lymph node with a pathologic diagnosis of metastastic carcinoma. Eight additional lymph nodes dissected from the left axilla were negative for metastasis.

II. Whole-genome sequencing

500ng of DNA from each sample was sheared on a Covaris E210 Focused Ultra-sonicator. Paired-end, barcoded whole-genome sequencing libraries were then prepared with the NEXTflex DNA Sequencing Kit (Bioo Scientific) with 8-cycles of PCR amplification. Paired-end sequencing (100bp x 2 or 101bp x 2) was performed on HiSeq 2000 sequencers (Illumina) at the Harvard Biopolymers Facility (Harvard Medical School) and Axeq (Seoul, South Korea). Sequencing reads are deposited in the NCBI SRA with accession SRP041470.

High coverage whole-genome sequencing data from a study of MALBAC amplification of single cancer cells from the SW480 cancer cell line and corresponding SW480 unamplified bulk DNA (Zong et al., 2012) were obtained from the NCBI SRA under accessions: a) MALBAC-amplified SW480 single cells (SRX202787, SRX202978, SRX204744, SRX204745, SRX205035); b) Unamplified bulk SW480 DNA (SRX202980). MALBAC and Illumina sequencing adaptors were trimmed from MALBAC single-cell data using the 'phacro' toolkit (Hou et al., 2013) (http://sourceforge.net/projects/phacro/) with default settings, with the following adaptors: SRX204744 and SRX204745-GTGAGTGCTGGAGTGAGGTAGTGTGGAG; SRX205035, SRX202787 and SRX202978-GTGAGTGATGGTTGAGGTAGTGTGGAG. MALBAC adaptor trimming slightly improved alignment and coverage statistics, primarily for SRX202787, SRX202978, and SRX205035 (data not shown). An additional 10bp was trimmed from the 3' end of read 1 of SW480 bulk DNA, SRX202787, and SRX202978, due to trailing MALBAC adaptor not removed by 'phacro' (SRX202787 and SRX202978) and low quality random bases not aligning to the genome in most reads of all 3 samples. This further improved coverage and alignment statistics of these 3 samples (data not shown). High coverage whole-genome sequencing data of MDA-amplified single cells from a lymphoblastoid cell line and matching unamplified bulk DNA from a prior study (Hou et al., 2012) were obtained from the NCBI SRA under accessions: a) MDA-amplified single cells: BGI_YH1 (SRX121614, and SRX121616-SRX121620) and BGI_YH2 (SRX121610-SRX121613, and SRX121615); b) Unamplified bulk DNA (SRX121621-SRX121628). High-coverage whole-genome sequencing data for breast cancer primary tumor, metastasis, and normal blood samples from individual TCGA-E1-A15E were downloaded from cgHub (https://cghub.ucsc.edu).

III. Read alignment

All analyses were performed after alignment of sequencing reads to hs37d5 (1000 Genomes Project human genome reference based on the GRCh37 primary assembly) using bwa (Li and Durbin, 2009) version 0.6.2-r126 with settings: aln -l 40 -k 2; sampe -N 100. PCR duplicates were removed with Picard MarkDuplicates for all analyses of WGS data, except for *scTea* which performed PCR duplicate removal within its own pipeline.

IV. Whole genome sequencing coverage and performance analyses

WGS coverage and performance analyses, including all analyses presented in **Supplemental Note 1**, were performed with SAMtools (Li et al., 2009), BEDTools (Quinlan and Hall, 2010), R(R Development Core Team, 2011), custom scripts, and Microsoft Excel. Some plots were

created with ggplot (Wickham, 2009). Further details of the analysis methods can be found in **Supplemental Note 1**.

V. Single-cell analysis of somatic retrotransposition

scTea (Single-cell Transposable element analyzer)

Tea (Transposable element analyzer), which was originally developed to detect insertions of transposable elements (TEs) in cancer genomes (Lee et al., 2012), was adapted and modified for single-neuron whole genome analysis. *Tea* identifies a TE insertion from paired-end sequencing data at single-nucleotide resolution along with its mechanistic signature such as target-site duplication and poly-A tail. The detection is based on the occurrence of: a) flanking clusters of "repeat-anchored mate" (RAM) reads, which are reads uniquely mapped to the reference genome whose mates (RAM mates) map to a custom TE sequence library; and b) partially-aligned reads spanning the insertion breakpoint ("clipped reads"), whose unaligned tail sequences match the inserted TE (Lee et al., 2012). Note that *Tea* can detect only non-reference TE insertions (i.e. those absent in the human genome reference) since only read pairs mapping discordantly to the human genome reference (indicating structural rearrangement) are subject to the analysis.

To address the challenges arising from MDA amplification such as chimeras, and to further improve detection sensitivity, we developed a revised version of *Tea (scTea)*, specifically for MDA single-cell WGS. Major changes include: a) a scoring scheme assigning a score to each call, taking into account amplification noise; b) improved handling of poly-A signals; c) copy number genotyping of insertion calls; d) local read assembly to detect transduced sequences; e) a revised TE sequence library using only known active TE subfamilies; f) rigorous sensitivity analyses to establish call criteria; and g) specificity analyses using independent PCR validation.

Scoring of insertion calls

An initial analysis of MDA-amplified single-neuron WGS data using the original *Tea* pipeline and default settings for high coverage WGS (>30x) of bulk DNA predicted an average of 742 somatic insertions per neuron, the vast majority of which were found to be false predictions caused by MDA chimeras. This motivated us to develop a better scoring scheme to separate true TE insertion signals from MDA chimera noise. The signal level was measured by the number of RAMs supporting an insertion breakpoint and the noise level was measured by the number of RAMs in the local genomic region near the breakpoint that do not support the breakpoint. Specifically, we counted the number of plus strand RAMs on the left side of a predicted insertion (*d1*), and the number of minus strand RAMs on the right side of the insertion (*d2*). We also counted RAMs that do not support the predicted insertion: minus strand RAMs on the left side of the breakpoint (*w1*), and plus strand RAMs on the right side of the breakpoint (*w2*). For each predicted insertion, a score, $s = 2\sqrt{d_1 d_2} - (w_1 + w_2)$, was calculated. The score is maximized when there is the same number of plus and minus strand supporting RAMs and no chimeric RAMs. Scores were further normalized by a correction factor for the total number of mapped reads in each sample, since samples with more total sequencing reads would tend to have higher scores for the same insertion events.

We also tested normalization of scores by local GC-content (in 10 kb bins) in order to correct for MDA GC-amplification bias, using the same GC correction methods evaluated in the 'MDA GC-content amplification bias section' of Supplemental Note 1. Correction of scores for GC-content led to more uniform (reduced variance) score distributions and some separation of insertions according to copy number (one-copy versus two-copy insertions) (data not shown). However, GC correction of scores led to less separation between false positive calls and true positive germline insertions. We therefore did not use GC-content score correction for subsequent analyses.

Processing of poly-A reads

The original *Tea* pipeline identified poly-A sequences in clipped reads but was not optimized to detect poly-A sequences in RAM mate reads. This led to a reduction in sensitivity for retrotransposons with long poly-A tails (> ~150bp) during RAM cluster analysis, since given the average insert size, fewer DNA fragments would span from the genomic flank to the body of the inserted TE sequence beyond the poly-A tail. Similarly, it led to a reduction in sensitivity for TEs with long 3' transductions, as reads mapped to the genomic flank of the TE insertion would have their mates mapped to the genomic flank of the source TE rather than the TE sequence library. To increase detection signal for TEs with long poly-A tails and 3' transductions, we revised the pipeline in four ways: a) a 200 bp poly-A sequence was added to the TE sequence library; b) to ensure that pure poly-A reads are only mapped to the 200 bp poly-A sequence and not AluY, L1Hs, or SVA elements in the TE sequence library, we trimmed the poly-A tails of AluY, L1Hs, and SVA sequences to have at most 27 bp poly-A tails (the size was determined based on the shortest read length generated internally for mapping by *scTea* and bwa mismatch alignment parameters); c) RAM cluster definition allowed for a mixture of RAMs mapping either to TE sequences or to the poly-A sequence; d) in the pairing of plus-strand and minus-strand RAM clusters, a RAM cluster of poly-A reads was allowed to be paired with a RAM cluster of any TE family (Alu, L1, and SVA). These revisions allowed us to capture the signal of all poly-A reads of a TE insertion. Moreover, it enabled us to detect somatic L1#2, which had a long poly-A tail (114bp) and a long 3’ transduction (614bp), by capturing its poly-A tail as a RAM cluster. This demonstrates the improved sensitivity of *scTea* in detecting TE insertions with long poly-A tails and/or 3’ transductions.

Copy number (zygosity) genotyping of insertion calls

The reference alignment coordinates of each read pair in the bulk cortex sample were extracted to recover the full span of the DNA fragment of each read pair. For every insertion being genotyped, we count the number of DNA fragments fully spanning the insertion breakpoint and an additional 5bp beyond both sides of the breakpoint. Homozygous insertions (two-copies per genome) and hemizygous insertions (single copy per genome) would be expected to not have such breakpoint-spanning fragments, while heterozygous insertions (single copy per genome) would have breakpoint-spanning fragments deriving from the allele without the insertion. In order to maintain a high confidence set of true positive insertions for subsequent evaluations of *scTea* single-copy insertion sensitivity, we only genotyped insertions identified in both cortex and heart bulk samples that were also detected in prior studies of retrotransposon polymorphism (see 'Sensitivity analyses' below for details). The distribution of the number of breakpoint-

spanning fragments was plotted separately for insertions of each TE family and showed a clear separation between insertions with 0 or few breakpoint-spanning fragments (homozygous and hemizygous) and a population of insertions with numerous breakpoint-spanning fragments (heterozygous; with 23 fragments on average per insertion). Using a cutoff that confidently separates these two populations, we assigned copy number to insertions. Insertions on chromosomes X and Y all had 0 breakpoint-spanning fragments and were defined as single-copy (hemizygous) insertions, since individual 1465 is a male.

Local read assembly to detect transduced sequences

We revised *scTea* to allow *in silico* assembly of transductions of TE insertions if present to aid review of insertion calls and allow identification of the source TE. The original *Tea* pipeline assembled insertion sequence contigs using only RAM mates mapping to TE sequence libraries; therefore, the resulting sequence contigs would not contain transduced sequences. *scTea* performs the same sequence contig assembly, but also assembles an additional 'extended RAM mate' sequence contig using mates of all discordant reads mapping to the RAM cluster region with the expected orientation relative to the insertion, regardless of whether the mates mapped to the TE sequence library. Contigs were assembled using 'CAP3' (Huang and Madan, 1999). *scTea* successfully detected the 5' and 3' transductions of L1#1 and L1#2, respectively. These extended RAM mate contigs are generated by *scTea* for all predicted insertions and labeled as 'eprammate' and 'enramamte' for plus- and minus-strand oriented discordant reads.

Transposable element sequence library

The TE sequence library in the original *Tea* pipeline contained canonical TE sequences from Repbase (Jurka et al., 2005) and all TE families and subfamilies sequences with divergence < 30% annotated in the human genome reference by RepeatMasker (Smit, 2010), including old inactive TE subfamilies. However, MDA chimeras involving old TE subfamilies such as AluS and L1PA generate many false positive predictions. Therefore, *scTea* uses a TE sequence library that includes active young retrotransposon subfamilies, but not inactive subfamilies. The subfamilies were selected based on previous studies of polymorphic TEs and retrotransposon activity in humans (Hancks and Kazazian, 2012; Stewart et al., 2011). The sequences included in the *scTea* TE sequence library were as follows:

AluY:

- Repbase consensus sequences of AluY, AluYa5, AluYb8, AluYb9, AluYg6.
- All AluY insertions annotated in hg19 by RepeatMasker with size >300 bp and divergence < 5% from the consensus (substitutions, deletions, and insertions).

L1Hs:

- Repbase L1Hs consensus and consensus variants created by diagnostic nucleotide substitutions for Ta-1d, Ta-1nd_G1, Ta-1nd_C, Ta-0, and Pre-Ta_ACG_G subfamilies.
- All L1Hs insertions annotated in hg19 by RepeatMasker with size > 6 kb and divergence < 5% from the consensus (substitutions, deletions, and insertions).

SVA:

- Repbase consensus sequences of SVA, SVA_A/B/C/D/E/F.
- All SVA insertions annotated in hg19 by RepeatMasker with size >2 kb divergence < 10% from the consensus (substitutions, deletions, and insertions).

Sensitivity analyses

We first evaluated the absolute genome-wide sensitivity of the *scTea* computational pipeline using reads simulated from the HuRef (J. Craig Venter) genome (Levy et al., 2007). The HuRef genome was used for this purpose since it is a Sanger-sequenced, fully-assembled diploid genome, which therefore provides a high quality annotation of TE insertions across an entire genome that is not available for any other individual genome. Next-generation WGS data is not available for the HuRef genome, so we simulated 555 million paired reads of WGS data using the tool 'dwgsim' to provide a similar read depth and DNA insert size distribution (300±65 bp) as our bulk WGS data. Half of the read pairs were simulated from the ABBA00000000 assembly (HuRef, high-scoring allele) and half from the ABSL01000000 assembly (HuRefPrime, low-scoring allele). Since 'dwgsim' simulates each chromosome with equal read depth, half of chromosome X and Y read pairs were removed to simulate a 46XY karyotype. HuRef-specific TE insertions (i.e. absent from the human genome reference) annotated by Xing, et al (Xing et al., 2009) were used as the gold standard set to estimate *scTea* sensitivity for HuRef simulation data. This annotation contained 584 AluY, 52 L1 (including 49 L1Hs), and 16 SVA insertions, and sensitivity was defined as the fraction of these insertions detected by *scTea* within a 50bp margin. We further investigated insertions called by *scTea* but absent from the annotation of Xing, et al and found many additional insertions (139 AluY, 35 L1Hs, 6 SVA) insertions. Most of these insertions were missed by Xing, et al. due to absence of a poly-A tail or target-site duplication, or incomplete annotation of HuRef insertions by Levy, et al (2007).

Sensitivity of *scTea* for TE detection in MDA-amplified (100-neuron and single-neuron) WGS data was evaluated using high-confidence non-reference (i.e. absent from the human genome reference) germline insertions as a reference set. The high-confidence reference set was defined as all insertions detected in both heart and cortex bulk WGS of individual 1465 (without score cutoff to avoid biasing against low-scoring germline insertions) that were also independently identified in prior studies of TE polymorphism within a 500bp distance (due to uncertainty in precise location of many insertions in prior studies). The set of 'known non-reference' TE insertions detected in prior studies was obtained from dbRIP (Wang et al., 2006), Beck et al (2010), Ewing and Kazazian (2010), Ewing and Kazazian (2011), Hormozdiari et al (2011), Huang et al (2010), Iskow et al (2010), and Stewart et al (2011).

Based on the sensitivity analyses and examination of false positive calls and missed true germline insertions with different parameter cutoffs, we generated a final call set for each sample with the following criteria: a) score ≥ 9; b) ≥ 2 RAM reads on each side of the breakpoint; c) ≥ 4 clipped reads supporting the insertion call; d) estimated target-site duplication or deletion ≤ 50 bp in size in the absence of a poly-A tail, or ≤ 250 bp in size if a poly-A tail was detected; e) at least half of clipped reads at the insertion site aligned to ±2 bp of the insertion breakpoint. In order to capture any brain-specific somatic insertions, we performed germline insertion filtering using insertions detected in the heart bulk sample with the following criteria: a) > 1 RAM read in the RAM cluster genomic region; b) > 1 clipped read within 2 bp of the predicted insertion breakpoint in the heart bulk sample. We also tested less stringent germline filtering allowing up

to 2 RAM reads and up to 2 clipped reads in the heart bulk sample, which produced only 4 additional somatic insertion candidates across all 16 single-neuron samples that were false positives by manual review of sequencing data.

Specificity analysis

We evaluated the specificity of *scTea* by attempting to validate by PCR a total of 80 candidate insertions called in the 1465 heart bulk DNA sample, including both unknown (i.e. not identified in prior studies of retrotransposon polymorphism) and known non-reference insertions (i.e. independently identified in prior studies of retrotransposon polymorphism, as described in the previous section). The number of candidates selected for validation from each TE family were: 48 AluY (16 unknown and 32 known non-reference), 24 L1Hs (8 unknown and 16 known non-reference), and 8 SVA (4 unknown and 4 known non-reference). Candidates were randomly selected from all the *scTea* insertion calls in 1465 heart bulk DNA with score ≥ 9, separately for each family and insertion type (unknown and known non-reference). Insertion candidates for which PCR validation primers could not be designed (due to repetitive genomic sequence) were replaced by other randomly selected candidates. Specific numbers of insertions were selected separately from unknown and known non-reference candidates to ensure validation attempts of adequate numbers of candidates from each category, since unknown and known non-reference insertions may have different specificities. In order to account for possible differences in specificity of unknown (UNK) and known non-reference (KNR) candidates, the final specificity for each TE family was calculated as a weighted average of the specificity for unknown and known non-reference insertions, weighted by the relative fraction of each out of all insertion candidates. Specifically, for each TE family, $specificity_{family} =$

$$\frac{\#\ UNK\ insertion\ candidates}{total\ \#\ of\ insertion\ candidates} \cdot \frac{\#\ validated\ UNK\ insertions}{\#\ UNK\ insertion\ candidates\ tested}$$

$$+ \frac{\#\ KNR\ insertion\ candidates}{total\ \#\ of\ insertion\ candidates} \cdot \frac{\#\ validated\ KNR\ insertions}{\#\ KNR\ insertion\ candidates\ tested}$$

For example, $specificity_{AluY} = \frac{189}{1136} \cdot \frac{13}{16} + \frac{947}{1136} \cdot \frac{32}{32} = 97\%$.

Validation results and sequences of validation primers used for each candidate insertion are in **Table S3**. Validation PCR was performed using cerebellum bulk DNA from individual 1465, since cerebellum tissue was an abundant source of available DNA. All validated insertions were confirmed by Sanger sequencing (Genewiz). Of the 77 validated insertions, 74 were insertion of a new retrotransposon element, and 3 were calls due to a deletion in the hg19 reference but not in individual 1465 of part of a retrotransposon element that is present in hg19 (**Table S3**). See the next section ('Validation and cloning of retrotransposon candidates') for details of the PCR validation methods.

VI. Validation and cloning of retrotransposon candidates

Validation PCR protocols

Validation of germline and somatic insertion candidates predicted by *scTea* was attempted by: 1) full-length PCR (FL-PCR) with genomic primers designed to flank the candidate (for Alu and L1

candidates), and 2) 3'-junction PCR (3'PCR) with a primer designed downstream of the 3'-end of the candidate (relative to the candidate insertion's predicted orientation) paired with an internal primer specific to the 3' sequence of the retrotransposon (for L1 and SVA candidates).

FL-PCR and 3'PCR primer design, full-length TOPO cloning, and Sanger sequencing were performed as previously described (Evrony et al., 2012). See below for further details. A set of positive and negative control germline retrotransposon insertions were included in every validation experiment to confirm proper setup of the assays. Sequences of validation primers used for each candidate insertion can be found in **Table S3**. Positive validation reactions were confirmed by Sanger sequencing (Genewiz). The insertion allele PCR product of heterozygous insertions was gel purified prior to sequencing if necessary.

Alu insertion candidate validation

Validation of Alu insertion candidates was performed by FL-PCR. Alu FL-PCR reaction mix and PCR cycling conditions were identical to L1 3'PCR (Evrony et al., 2012) since Alu insertion cloning does not require the specialized long-range PCR used in L1 FL-PCR. To confirm Alu FL-PCR assay sensitivity and specificity, 16 population-polymorphic germline Alu insertions were assayed: 8 Alu insertions present in the human genome reference but predicted to be absent from individual 1465 based on bulk WGS, and 8 Alu insertions predicted to be present in individual 1465 based on bulk WGS but absent from the human genome reference. Alu FL-PCR confirmed all 16 Alu insertions correctly as present or absent in individual 1465, and all 16 insertions were detected by Alu FL-PCR among 8 additional human DNA samples and confirmed by Sanger sequencing.

L1 insertion candidate validation

Validation of L1 insertion candidates was performed with both FL-PCR and 3'PCR in case one validation method failed. L1 FL-PCR and 3'PCR reaction mixes and PCR cycling conditions were as previously described(Evrony et al., 2012). The 3'PCR L1Hs-specific primer was L1Hs-AC-22 (TATACCTAATGCTAGATGACAC) (Evrony et al., 2012).

Primer sequences for FL-PCR of the source L1Hs on chr13: 30,215,844-30,221,843 (hg19) from which L1#2 derived are:
Left TGGGCAAGTGTTGAAAGCTT
Right AGGACTAAAAGCCTTTCCCCT

Additional cloning experiments of the source L1Hs from which L1#2 derived were performed with Q5 high-fidelity DNA polymerase (NEB) with the following primers:
Left GGATCTTAAGGTTGAAGGTTTGG
Right AAAGTAGTTCTCGAGCTCCGGT

SVA insertion candidate validation

Validation of SVA insertion candidates was performed by 3'PCR. SVA 3'PCR reaction mix and PCR cycling conditions were identical to L1 3'PCR (Evrony et al., 2012). FL-PCR was not performed as PCR of full-length SVA insertions is challenging due to their high GC content. The 3'PCR SVA-specific primer used was either SVA-1 (TCACTTGTTTATCTGCTGACCTTC) or

SVA-2 (CCTTCCCTCCACTATTGTCCTA) (Stewart et al., 2011), chosen for each candidate based on the *scTea* contig assembled for the insertion so that the SVA-specific primer sequence is present in the insertion without mismatches.

Assays for somatic L1s in formalin-fixed tissue

Efficiency of each DNA extraction from formalin-fixed paraffin-embedded tissue of the right cerebral hemisphere was assayed by PCR for control genomic regions with amplicons ranging in size from 60 bp to 350 bp. Three independent DNA extractions were performed from each location. Most fixed tissue DNA samples successfully amplified most, though not all, control amplicons, although at low levels based on band intensity versus non formalin-fixed control DNA. DNA remaining for each sample after genomic PCR controls was aliquoted into a series of nested 3'-junction PCR assays designed to detect L1#1 and L1#2, along with positive and negative control reactions (i.e. single-cell DNA harboring the somatic L1, unrelated human, and water controls). Nested 3'PCR assays were designed with shorter amplicon sizes than digital nested 3'PCR assays used for poly-A tail cloning in order to increase the probability of detection from formalin-fixed DNA, which is fragmented and damaged. PCR reaction cycling is identical to the protocol in the '*Digital nested 3'PCR'* section below.

L1#1 primer sequences:

Round 1: TGTTATTTGGCCCTTTAAGGAA (targets genomic flank)

TATACTAATGCTAGATGACAC (targets L1 3' UTR)
Note: Primer modified for L1#1 since L1#1 and its source L1 contain a 1 bp deletion relative to the L1Hs consensus of TATACCTAATGCTAGATGACAC

Round 2: AGCCCTTGCAGAGGAATCA (targets genomic flank)

CACATGTACCCTAAAACTTAG [FAM labeled] (targets L1 3' UTR)

L1#2 primer sequences:

Round 1: CCCTTTCCAAGTCCATTGAG (targets 3' transduction)

CCCAAATCATCAACTAATCCTAATTT (targets genomic flank)

Round 2: TTGATTGTGTCATTTTTCTTCTTTG (targets 3' transduction)

AATTGTTAGTAATTGATAAGGACATGG (targets genomic flank)

VII. Droplet digital PCR (ddPCR)

Custom ddPCR assays for L1#1 and L1#2 were performed with the QX100 Droplet Digital PCR System (Bio-Rad) per manufacturer's instructions. Each L1 assay was multiplexed with an assay for RNaseP serving as a genomic copy number reference (copy number = 2). L1 and RNaseP assays were labeled with 6FAM and HEX, respectively. Assays were performed at least in duplicate for each tissue and location. Confirmation of ddPCR single-copy sensitivity and linear concentration measurement (**Figure S13B**) was performed with a stock solution of synthetic

oligos of the L1#1 amplicon (IDT Technologies) diluted to a target concentration of 20,000 copies/ul; the actual concentration measured by ddPCR was 6,740 copies/ul. Both ddPCR assays were tested with multiple unrelated human control samples that confirmed assay specificity (**Table S4**). Standard PCR with left+right ddPCR primers and PCR product Sanger sequencing confirmed the correct amplicons are amplified. Note that L1#2 primers also amplify one off-target band that is not measured by the internal fluorescent probe during the ddPCR assay. Primers and probes were ordered from IDT.

Percent mosaicism (% of cells) was calculated with the QX100 QuantaSoft software. ddPCR assays exhibit reduced L1 signal in double-positive ($L1^{+}RNaseP^{+}$) versus single-positive ($L1^{+}RNaseP^{-}$) droplets due to relatively higher PCR efficiency of the shorter RNaseP amplicons. Double-positive droplets are seen only in samples harboring the L1 and appear in the proportion expected based on the fraction of single-positive ($L1^{+}RNaseP^{-}$ and $L1^{-}RNaseP^{+}$) droplets, both in experiments with bulk DNA from individual 1465 and in dilution experiments with known input copy numbers of synthesized oligos of L1#1 (data not shown). Reduced L1 signal in double-positive droplets does not affect quantification since double-positive droplets are still detected in a distinct distribution from the $L1^{-}RNaseP^{+}$ population.

The presence or absence of L1#1 and L1#2 in unamplified bulk DNA from every location and tissue was independently verified by bulk (non-digital) nested 3'-junction PCR and was fully concordant with ddPCR results (**Figure S15**). Bulk nested 3'PCR is identical to digital nested 3'PCR except that DNA is not diluted for single-copy cloning (see 'Poly-A tail cloning and sizing' below for nested 3'PCR protocol).

Primer sequences for RNaseP reference:
Left GATTTGGACCTGCGAGCG
Right GCGGCTGTCTCCACAAGT
Probe TTCTGACCTGAAGGCTCTGCGC (Hex-IowaBlackZen labeled)
(Product size: 62bp)

Primer sequences and cycling conditions for L1#1:
Left AGGCACAATCTGTGAAGCAG (targets 5' transduction)
Right AAAAGGCTGAATTAAACCTAACACA (targets genomic flank)
Probe ATGATTCCTGGCCCTCTGCATTGTCT (6FAM-IowaBlackZen labeled)
95°C 10 min; [94°C 30 sec, 60°C 1 min] x 40; 98°C 10 min; 12°C forever
(Product size: 132bp)

Primer sequences and cycling conditions for L1#2:
Left TGACAAAGGGCTAATATCCAGAA (targets transposon sequence at 5' junction)
Right AGGTCAGTGTTGCACTAGCAAT (targets genomic flank)
Probe AGGTGCTGGAGGGATCATCCCT (6FAM-IowaBlackZen labeled)
95°C 10 min; [94°C 30 sec, 58°C 1 min] x 40; 98°C 10 min; 12°C forever
(Product size: 221bp).

ddPCR reaction mix (per reaction):

Component	Volume (ul)

Water	4.4
2x ddPCR Supermix (Bio-Rad)	10.5
20x L1 primer-probe mix*	1.05
20x RNaseP primer-probe mix*	1.05
DNA	4
Total volume	**21**

*20x primer-probe mix: 18 uM left+right primers (each), and 5 uM probe.

Estimation of total number of cells harboring the insertions:

The total number of cells in the brain with each insertion (L1#1 and L1#2) was estimated by calculating the total number of insertions present in DNA extracted from sampled brain regions, and then correcting this number by the fraction of the brain tissue sampled out of the insertion's full geographic distribution in the brain. Specifically, for each insertion we first calculated the total number of insertion copies present in DNA extracted from each brain location. This was calculated for each brain location by multiplying the total volume of extracted DNA solution from that location by the number of insertion copies per unit volume as measured by ddPCR for that location. The sum of the number of insertion copies across DNA samples from all locations is the total number of cells with the insertion that we extracted from the brain (since each somatic insertion is present in one copy per cell). Because we sampled/extracted DNA from only a fraction of the brain regions harboring each insertion, we next estimated the fraction of sampled brain regions out of the total brain regions estimated to harbor the insertion (i.e. L1#1 subregion of the left middle frontal gyrus; L1#2 left brain). This fraction was based on the sizes of cortical tissue samples from which DNA was extracted and photographs of sampled brain regions (**Figure S14**). The fraction of an insertion's distribution in the brain that was sampled was then used to correct the total number of insertion copies extracted, in order to obtain the total number of cells estimated to harbor the insertion in the brain. Nevertheless, these estimates are only rough extrapolations since they rely on an assumption of the same average mosaicism in unsampled regions as in sampled regions. Better estimates of clone sizes will require development of high-throughput technologies for whole-brain genetic profiling.

Mapping of locations to representative brain image:

Coronal sections of the frozen left cerebral cortex of individual 1465 were photographed before and after sampling. Sampled locations were them mapped to a representative brain using measured section thicknesses and anatomy of gyri as seen on the section photographs. Since an image of the complete brain of individual 1465 prior to sectioning was not available, sampled locations are illustrated on a representative brain image obtained with permission from the University of Wisconsin and Michigan State Comparative Mammalian Brain Collections (http://brainmuseum.org), supported by the US National Science Foundation. Because section thicknesses changes during processing and freezing, measured section thicknesses were scaled proportionately during mapping to match the total rostral-caudal length of the representative brain. Final sample locations on the representative brain are the best mapping possible using all available information, but are not definitive due to variability among individual brains.

VIII. Poly-A tail cloning and sizing

Digital nested 3'PCR

In order to accurately measure the distribution of poly-A tail lengths present in a sample, we developed a digital nested 3'PCR approach (dnPCR) in which single copies of poly-A tails are cloned directly from unamplified bulk DNA. dnPCR single-copy cloning of poly-A tails directly from unamplified bulk DNA is necessary in order to: a) avoid potential artifacts of MDA; and b) avoid skews in the distribution of poly-A tail sizes that arises in bulk poly-A tail cloning due to differential amplification efficiency of differently sized poly-A tails. dnPCR addresses the first issue since it clones individual poly-A tails directly from unamplified cortex, obviating the need for MDA single-cell amplification which could potentially introduce poly-A tail mutation. Furthermore, single-copy (digital) cloning by dnPCR is essential, since shorter poly-A tails amplify during PCR with significantly greater efficiency relative to longer poly-A tails (data not shown). Therefore, bulk (non-digital) nested 3'PCR in which numerous poly-A tails are amplified in the same reaction, yields a skewed distribution of product intensities with increased intensities at smaller lengths. This means that bulk (non-digital) nested 3'PCR cannot be used to measure the distribution of poly-A tail lengths. (Nevertheless, due to its single-copy sensitivity, bulk non-digital nested 3'PCR was still useful as an independent confirmation of ddPCR results for presence/absence of L1#1 and L1#2 in every tissue/location; see 'Droplet digital PCR (ddPCR)' above).

Since dnPCR avoids both the use of MDA and avoids biases in the poly-A size distribution that would arise in bulk poly-A cloning, the only potential remaining source of error would be dnPCR itself. As described below, dnPCR of clones of known poly-A tail length from previous dnPCR experiments can be performed to estimate an upper-bound for the rate at which dnPCR itself mutates the poly-A tail. Furthermore, poly-A tail sizes cloned by dnPCR in tissues with a predominant single poly-A tail size provide a more accurate absolute measure of the fidelity/mutation rate of dnPCR.

In dnPCR, each DNA sample is first diluted to a target retrotransposon insertion concentration of 0.3 copies/µl based on the absolute concentration of the somatic retrotransposon insertion as measured by ddPCR. By Poisson statistics ($\lambda = 0.3$), there would be < 5% (≈3.7%) chance that the 1µl of diluted DNA input into a dnPCR reaction would contain >1 poly-A tail. This is important since having many reactions with >1 poly-A tail would confound building an accurate distribution of poly-A tail lengths in situations with low poly-A tail length variability. After each experiment, refined estimates of λ for each dilution were calculated using the formula $\lambda = -\ln(k)$ where k is the fraction of reaction wells without a product. These estimates of λ were used to further fine-tune dilutions for subsequent experiments. Experiments with $\lambda > 0.5$ (i.e. $k < 0.61$, with >9% chance of >1 copy in a reaction well by Poisson statistics) were excluded from downstream analysis.

The diluted DNA is then run through a 2-round nested PCR targeting the 3' junction (containing the poly-A tail) of the somatic retrotransposon insertion. At least 1 out 16 reactions in every experiment were water input negative controls. PCR recipes and cycling conditions are as follows:

Round 1 PCR (per reaction)

Component	Volume (ul)
Water	12.68
5x colorless GoTaq Flexi buffer (Promega)	4

$MgCl_2$ (25 mM)	1.2
dNTP (10 mM each)	0.4
DMSO	0.2
Primer 1 (100uM)	0.16
Primer 2 (100uM)	0.16
GoTaq Hot Start polymerase (Promega)	0.2
Diluted DNA	1
Total volume	**20**

Round 1 cycling conditions:

95°C 5 min
[95°C 30 sec, 59°C 30 sec; 72°C 1 min] x 15
72°C 5 min
8°C forever

Round 2 PCR (per reaction)

Component	Volume (ul)
Water	12.68
5x colorless GoTaq Flexi buffer (Promega)	4
$MgCl_2$ (25 mM)	1.2
dNTP (10 mM each)	0.4
DMSO	0.2
Primer 1 (100uM)	0.16
Primer 2 (100uM)	0.16
GoTaq Hot Start polymerase (Promega)	0.2
Round 1 PCR product	1
Total volume	**20**

Round 2 cycling conditions:

95°C 5 min
[95°C 30 sec, 59°C 30 sec; 72°C 1 min] x 35
72°C 5 min
8°C forever

L1#1 primer sequences:

Round 1: CAATCAAGATTGGGGAGGTG (targets genomic flank)

TATACTAATGCTAGATGACAC (targets L1 3' UTR)
Note: Primer modified for L1#1 since L1#1 and its source L1 contain a 1 bp deletion relative to the L1Hs consensus of TATACCTAATGCTAGATGACAC

Round 2: ACATGGTGGAGGGGACATAG (targets genomic flank)

CACATGTACCCTAAAACTTAG [FAM labeled] (targets L1 3' UTR)

L1#2 primer sequences:

Round 1: CCAAATTTCAGCCATTTTGC (targets genomic flank)

GGCTGTAGGTTTTTGGTGGA (targets 3' transduction)

Round 2: AGAATGCATAACTACCCAAATCA (targets genomic flank)

CCCCCACTTCCTTCCTGTAT [FAM labeled] (targets 3' transduction)

Breast cancer somatic L1 primer sequences:

Round 1: TGAAATTTTGTGATTTGGGTGT (targets genomic flank)

TATACCTAATGCTAGATGACAC (targets L1 3' UTR)

Round 2: TTGAACATTGCCAAAACTCAAC (targets genomic flank)

CACATGTACCCTAAAACTTAG [FAM labeled] (targets L1 3' UTR)

Screening and sequencing positive digital nested 3'PCR reactions

After round 2 PCR, all reaction wells are screened by 2% agarose gel electrophoresis for wells containing a product. These are picked for downstream analysis. A small subset of dnPCR reaction wells had 2 or 3 products of different size visible on gel and concordantly on capillary electrophoresis, due to cloning of >1 poly-A tail in the same reaction. Poly-A tails from these multi-product reactions were counted individually similarly to reactions cloning only one poly-A tail. At least 100 poly-A tails were cloned from each studied tissue/location (see **Table S5** for full list), with the exception of L1#1 in cerebral cortex (location I) that had very low mosaicism allowing cloning of only 4 poly-A tails.

A subset of positive dnPCR reactions from each tissue and location were Sanger sequenced (Genewiz, Inc.) and confirmed that dnPCR amplifies the 3' junction of the targeted retrotransposon insertion with 100% specificity (data not shown). Correlation of Sanger sequencing traces with poly-A tail sizes measured by capillary electrophoresis revealed that the poly-A tail length corresponds to the point on Sanger sequencing where the poly-A signal begins to decrease from its plateau intensity.

Sizing of digital nested 3'PCR products

All reactions containing a product were sized by capillary electrophoresis on 3130 or 3730 DNA Analyzers (Life Technologies) with standard settings for fragment size analysis. dnPCR products are detected by capillary electrophoresis by the FAM-labeled primer incorporated into final products in round 2 of dnPCR .

Sizing reaction mix:

Component	Volume (ul)
Hi-Di formamide (Life Technologies)	8.7
GeneScan 500 LIZ size standard (Life Technologies)	0.3

Final dnPCR product	1
Total volume	**10**

The sizing reaction mix was denatured at 95°C for 3 min, then cooled to 4°C for 3 min, prior to sizing.

GeneMapper 4.0 software (Life Technologies) was used to analyze electrophoresis traces. Due to inevitable slippage of polymerase during amplification of homopolymers, dnPCR products appear as stutter peaks in electropherograms (**Figure S15A**). The highest intensity peak at the center of a stutter was selected as the product size. Since there can be two similarly intense peaks in the center of a stutter, there is a ±1 bp uncertainty in picking the correct product size. The size of the poly-A tail was then calculated as: [measured dnPCR product size] - [known amplicon sequence length excluding the poly-A tail] - 1. The amplicon sequence lengths excluding the poly-A tail are 340bp, 211bp, and 113bp for L1#1, L1#2, and the breast cancer somatic L1, respectively. The additional 1 bp was subtracted to account for the 3' terminal dA added by Taq during final extension of PCR.

Digital nested 3'PCR of poly-A tails of known length

In order to confirm that dnPCR itself does not cause the highly polymorphic poly-A tails we observed, we input poly-A tails of known length into dnPCR. Poly-A tails of known length were obtained from a prior dnPCR experiment for L1#1 from cerebral cortex (location D). Three poly-A tails with different lengths were selected from this dnPCR experiment: 62bp, 74bp and 148bp. Round 1 PCR product from these reaction wells was diluted to 0.3 copies/ul in the same manner as dnPCR of unamplified bulk tissue DNA. The diluted control poly-A DNA was then run through the same 2-round dnPCR assay for L1#1 and sized as described above. The resulting poly-A length distributions (**Figure S16B**) illustrate that dnPCR cannot cause the significant poly-A tail polymorphism we observe in tissues of individual 1465. There was a trend for a wider size distribution with increasing poly-A tail length for the control poly-A tails (**Figure S16B**). This is expected due to increased stutter of longer poly-A tails during PCR.

Importantly, the recovered poly-A tail lengths of control poly-A tails is a significant *overestimate* of the variability introduced by dnPCR itself. Control poly-A tails were sampled from prior round 1 PCR product, and therefore underwent an additional 15 cycles of PCR relative to dnPCR performed on bulk DNA. These 15 additional cycles of PCR introduced mutations that are reflected in the final dnPCR. Nevertheless, these control poly-A tails were used to obtain an initial upper bound for dnPCR fidelity, because there is no current method to artificially synthesize a control sample with perfectly homogenous poly-A tails. Despite this limitation, our subsequent dnPCR experiments from bulk DNA proved that the precision of dnPCR is significantly greater than the initial upper bound suggested by the control experiments. This evidence includes: a) the sharp peaks of poly-A tail lengths in dnPCR of bulk tissue DNA, including a sharp (±1 bp) peak at 148bp for L1#1 in cerebral cortex-location D (**Figure 4D**) and sharp L1#2 poly-A peaks between 112 to 114bp across all tissues (**Figure S17B**); b) the 148bp peak in cerebral cortex-location D has a significantly smaller distribution width relative to the distribution of the 148bp poly-A control of known length (**Figures 4D and S16B**); and c) bulk nested 3'PCR and dnPCR of the breast cancer-specific L1 measured 81bp and 80bp poly-A tail peaks in primary tumor versus metastasis in multiple independent experiments (**Figures S18D-E**,

and data not shown). These results indicate that dnPCR achieves a precision of at least ±1 bp across a wide range of poly-A tail sizes.

Additionally, we assessed the degree to which amplification by single-cell MDA mutates poly-A tails. In the event of an MDA mutation by the phi29 polymerase, the dominant form of the amplicon should still reflect the original true poly-A tail length since MDA artifacts would be present at relatively lower levels. Because each of the 2 single DNA strands of a heterozygous variant are copied independently by MDA, even an MDA mutation created in the first replication of the original genomic DNA would be present on average in 1/4 of the final amplicons, assuming the original and MDA-mutant genotype amplify with equal efficiency. Indeed, in every one of the 15 single-neurons harboring the somatic L1 insertions (2 L1#1 neurons and 13 L1#2 neurons), only one consistent peak (with the usual stutter pattern) of L1#1 and L1#2 poly-A tail length was seen on 3'PCR agarose gels and non-digital nested 3'PCR agarose gel and capillary electrophoresis experiments (**Table S5** and data not shown). Moreover, we performed dnPCR of L1#1 in single-neuron 77 to construct a distribution of L1#1 poly-A tail lengths resulting from MDA amplification. Single-neuron 77 has an L1#1 poly-A tail length of 40bp as measured by non-digital nested 3'PCR, and dnPCR showed 25/27 (93%) poly-A tail clones between 39bp and 41bp in length, i.e. ±1 bp of the previously measured 40bp length (**Table S5**). The remaining 2/27 clones were 43bp and 10bp, presumably mutated by either MDA or dnPCR. Altogether, the above results suggest that MDA faithfully amplifies poly-A tails at least up to 115bp in length (the poly-A tail length in most L1#2 single neurons). However, this assessment of MDA amplification of poly-A tails cannot be extrapolated to longer poly-A tail lengths that are likely more susceptible to MDA mutation.

Supplemental References

Arcot, S.S., Wang, Z., Weber, J.L., Deininger, P.L., and Batzer, M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. Genomics 29, 136-144.

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.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40, e72.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-580.

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.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62, 1408-1415.

Cai, X., Evrony, G.D., Lehmann, H.S., Elhosary, P.C., Mehta, B.K., Poduri, A., and Walsh, C.A. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell reports 8, 1280-1289.

Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94, 1041-1046.

Cordaux, R., Lee, J., Dinoso, L., and Batzer, M.A. (2006). Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 373, 138-144.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J.*, et al.* (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA 99, 5261-5266.

Dewannieux, M., and Heidmann, T. (2005). Role of poly(A) tail length in Alu retrotransposition. Genomics 86, 378-381.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nature reviews Genetics 5, 435-445.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A.*, et al.* (2012). Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483-496.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20, 1262-1270.

Ewing, A.D., and Kazazian, H.H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res 21, 985-990.

Grandi, F.C., and An, W. (2013). Non-LTR retrotransposons and microsatellites: Partners in genomic variation. Mobile genetic elements 3, e25674.

Grandi, F.C., Rosser, J.M., and An, W. (2012). LINE-1 Derived Poly(A) Microsatellites Undergo Rapid Shortening and Create Somatic and Germline Mosaicism in Mice. Mol Biol Evol 30, 503-512.

Hancks, D.C., and Kazazian, H.H., Jr. (2012). Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22, 191-203.

Hormozdiari, F., Alkan, C., Ventura, M., Hajirasouliha, I., Malig, M., Hach, F., Yorukoglu, D., Dao, P., Bakhshi, M., Sahinalp, S.C., and Eichler, E.E. (2011). Alu repeat discovery and characterization within human genomes. Genome Res 21, 840-849.

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X.S., and Qiao, J. (2013). Genome analyses of single human oocytes. Cell 155, 1492-1506.

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.

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.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res 9, 868-877.

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.

Jurka, J., Kapitonov, V., and Pavlicek, A. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462-467.

Kelkar, Y.D., Eckert, K.A., Chiaromonte, F., and Makova, K.D. (2011). A matter of life or death: how microsatellites emerge in and vanish from the human genome. Genome Res 21, 2038-2048.

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., and Makova, K.D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18, 30-38.

Kuhn, U., Gundel, M., Knoth, A., Kerwitz, Y., Rudel, S., and Wahle, E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J Biol Chem 284, 22803-22814.

Lander, E., Linton, L., Birren, B., Nusbaum, C., and Zody, M. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol 7, 19.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K*., et al.* (2012). Landscape of Somatic Retrotransposition in Human Cancers. Science 337, 967-971.

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.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

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.

Manley, K., Shirley, T.L., Flaherty, L., and Messer, A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat Genet 23, 471-473.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D.*, et al.* (2011). Tumour evolution inferred by single-cell sequencing. Nature 472, 90-94.

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.

Parsons, R., Li, G.M., Longley, M., Modrich, P., Liu, B., Berk, T., Hamilton, S.R., Kinzler, K.W., and Vogelstein, B. (1995). Mismatch repair deficiency in phenotypically normal human cells. Science 268, 738-740.

Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6, 729-742.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods 5, 1005-1010.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Smit, A.H., R & Green, P. (2010). RepeatMasker Open-3.0.

Srikanta, D., Sen, S.K., Conlin, E.M., and Batzer, M.A. (2009). Internal priming: an opportunistic pathway for L1 and Alu retrotransposition in hominins. Gene 448, 233-241.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P.*, et al.* (2011). A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 7, e1002236.

Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H., and Bartel, D.P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. Nature 508, 66-71.

Sun, J.X., Helgason, A., Masson, G., Ebenesersdottir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., and Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. Nat Genet 44, 1161-1165.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol 3, research0052.

Wagstaff, B.J., Hedges, D.J., Derbes, R.S., Campos Sanchez, R., Chiaromonte, F., Makova, K.D., and Roy-Engel, A.M. (2012). Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion. PLoS Genet 8, e1002842.

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.

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R.M. (2003). Likelihood-based estimation of microsatellite mutation rates. Genetics 164, 781-787.

Wickham, H. (2009). Ggplot2 : elegant graphics for data analysis (New York: Springer).

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A., and Jorde, L.B. (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res 19, 1516-1526.

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.

Supplemental Figures and Legends

Figure S1. Single-neuron WGS genome coverage, Related to Figure 1. (A) Fraction of genome covered above different read depth cutoffs. Read depth cutoffs were normalized to the average genome-wide read depth for each sample in order to allow comparison between samples, since each has a different total read depth. Note, each sample is plotted individually rather than averaging across samples of each type, illustrating the high reproducibility of the 16 MDA single neurons. The left panel shows only samples sequenced in this study for better visualization; the right panel shows samples from this study as well as single-cell and corresponding bulk samples from prior high-coverage MDA and MALBAC amplification studies (Hou et al., 2012; Zong et al., 2012). The single-cells with the best genome coverage from each of these prior studies were plotted (YH-1, SRX202787, and SRX202978). (B) Fraction of genome covered at ≥1x and ≥10x read depth at different subsampled genome-wide average read depths. Reads were randomly subsampled to obtain different total subsampled read depths, allowing comparison between samples regardless of their original total read depths. Lines show the average across samples of each sample type (shading ± SD). Increasing total sequencing leads to increasing genome coverage, with progressively diminishing returns (plateau). Plots for each sample set are shown up to the maximum read depth that was possible to subsample equivalently across all samples in a sample set (i.e. limited by the sample with the lowest read depth in the sample set). Bulk and single-cell samples from prior MDA and MALBAC studies (Hou et al., 2012; Zong et al., 2012) are included for comparison as in Figure S1A. (C) Lorenz curves as in Zong, et al (2012) showing the cumulative fraction of reads as a function of cumulative fraction of the genome, averaged across samples of each sample type. A sample with perfectly even genome coverage would appear on the diagonal y = x line. The left panel is plotted using all sequencing reads of each sample. Bulk and single-cell samples from prior MDA and MALBAC studies (Hou et al., 2012; Zong et al., 2012) are included for comparison as in Figures S1A-B. The right panel is plotted after subsampling to the same total read depth across all samples (i.e. the total read depth of YH MDA single-cell YH-1, which has the lowest total read depth of all samples), showing the same trends as the left panel. Subsampling to other total read depths across all samples always showed the same trends between sample types (data not shown).

Figure S2. MDA GC-content amplification bias, Related to Figure 1. (A) Coverage versus GC content in 100 kb tandem bins across the genome. Each point represents a bin, showing the average of its median-normalized coverage across samples in the set versus its GC content. Note that the global mean and median across all bins may not equal 1 after averaging the median-normalized coverage of each bin across samples in a sample set, as seen in MDA sample sets. GC curves (colored lines) were fit to all mappable 100kb bins across the genome (n=26,379 bins) and show the estimated expected coverage as a function of GC content. In all samples, coverage decays with increasing GC content, with significantly greater GC bias in MDA samples relative to unamplified bulk samples. Only a random subset of 1,000 bins is graphed in each plot for visualization purposes. Total variation, TV_{GC}, and standard-deviation, SD_{GC}, measure dispersion of the GC curve from the global mean across all bins (dashed line). In the absence of a GC effect, the GC curve would coincide with the mean line. Residual standard-deviation (SD_{resid}) measures dispersion of coverage from the GC curve. See **Supplemental Note 1** for further details. (B) MAPD and MDAD statistics of genome-wide coverage variability in (equal-size) 100kb tandem bins of different sample types, for uncorrected, non-paired GC/mappability

modeling correction, and paired correction using a reference sample or pooled sample set. Both non-paired and paired corrections reduce coverage variability to a large degree, with paired correction performing better. (C) Genome-wide coverage plot in 100kb tandem bins of a representative MDA single neuron (#46), showing the effect of non-paired GC/mappability correction versus paired correction using the MDA 100-neuron sample as a reference. Orange lines denote ±1 copy. See **Supplemental Note 1** for further details.

Figure S3. Evaluation of different reference sample corrections of MDA coverage variability, Related to Figure 1. (A) MAPD and MDAD statistics of genome coverage variability of different sample types, after correction using different references to define the boundaries of ~100 kb equal-read bins. All analyses were performed genome-wide (30,000 bins), except for MDA-amplified single-neurons versus pooled 15 MDA-amplified single neurons, which was performed only for chromosome 1. (B) Representative genome-wide coverage plots of different sample types versus different references. Orange lines denote ±1 copy. Purple points are off scale. See **Supplemental Note 1** for further details.

Figure S4. Effect of bin size and total read depth on MDA coverage variability, Related to Figure 1. (A) MAPD and MDAD statistics of genome coverage variability across all 16 high-coverage WGS single cortical neurons using the MDA 100-neuron sample as a reference to define equal-read bins with 5 different average sizes ranging from ~10 to ~1,000 kb. MAPD and MDAD increase with decreasing bin size, except for MAPD which decreases for 50kb and 10kb bins. See **Supplemental Note 1** for further details. (B) Representative genome-wide coverage plots for one single-neuron sample using different bin sizes. Orange lines denote ±1 copy. Purple points are off scale. (C) MAPD and MDAD measures of genome coverage variability across all 16 high-coverage WGS single cortical neurons at different subsampled genome-wide average read depths. All analyses used an MDA 100-neuron reference. (D) Representative genome-wide coverage plots for one single-neuron sample (neuron 46) at 0.1x and 30x subsampled read depths, illustrating stability of genome coverage variability at low read depths.

Figure S5. Genome-wide coverage plots of all 16 cortex single neurons in this study, Related to Figure 1. Genome-wide coverage in ~500kb equal-read bins for each single neuron in this study, normalized using the MDA-amplified caudate 100-neuron sample as a reference. MAPD and MDAD dispersion statistics are shown for each sample. Orange lines denote ±1 copy. Purple points are off scale.

Figure S6. Comparison of MDA and MALBAC single-cell coverage variability, Related to Figure 1. (A) MAPD and MDAD statistics of genome coverage variability across all 16 MDA single neurons and 3 previously published MALBAC single cells (SRX202978, SRX204745, and SRX205035) (Zong et al., 2012). (B) Representative genome-wide coverage plots of 2 MDA single-neurons and 2 MALBAC single-cells, in ~100kb and ~500kb equal-read bins. Orange lines denote ±1 copy. Purple points are off scale. (C) High-resolution read depth plot of a 15kb region (chr3:74,156,494-74,172,005) centered at the location where the somatic L1#2 insertion was identified in 1465 cortex single-neuron 6 (**Figure 2**). Note the relatively even coverage of

MDA at this scale compared to large, yet consistent peaks (stars) and troughs in MALBAC samples. The two MALBAC single cell samples with the best genome coverage are shown. Regions with no read coverage are annotated with black bars beneath plots.

Figure S7. Power spectral densities of read depth variability, Related to Figure 1. Plots of power spectral density (y-axis), which reflects read depth variability, versus genomic spatial frequency (x-axis) in inverse base-pair (bp) units. Smaller frequencies (left side of plot) reflect larger genomic scales. MDA samples have greater read depth variability at larger genomic scales, while MALBAC has greater read depth variability at smaller genomic scales. Power spectral density was calculated by the 'spectrum' function in R on normalized read coverage at single base-pair resolution across all non-gap regions of chromosome 1, followed by spline smoothing to smooth the plots. See **Supplemental Note 1** for further details.

Figure S8. Whole-genome sequencing analysis of MDA chimeric reads, Related to Figure 1. (A) Schematic of discordant read pairs used for quantification of inversion, deletion and duplication chimeras. Inversion chimera read pairs face the same direction; deletion chimera read pairs face each other but with an insert size larger than expected distance (insert size); duplication chimera read pairs face away from each other. Chimera breakpoint distances were estimated by the distance between the 3' ends of the reads in the discordant read pair. (B) Fraction of total WGS read pairs that are inversion, deletion, or duplication chimeras, at different chimera breakpoint distances (in 50bp bins). Plots are averages across the 16 single-neuron WGS and 2 unamplified bulk WGS (cortex and heart), as well as the 100-neuron WGS sample, from individual 1465. Inversion plot of unamplified bulk DNA exhibit peaks at 0-250bp and 500-2,000bp. Both are likely artifacts of library preparation and sequencing: the former are randomly distributed across the genome with no association with specific sequences, and the latter are low-level inversion reads between adjacent Alus in the genome, a phenomenon also present in publicly available Hapmap WGS samples prepared and sequenced by Illumina. Deletion and duplication plots exhibit periodic peaks in unamplified bulk DNA due to reads mapping to tandem repeats present in satellite regions of the genome (red stars). These peaks are not present in MDA samples, because MDA under-amplifies satellite DNA. Note that the deletion plot y-axis is scaled 10x relative to the inversion plot, and the duplication plot y-axis is scaled 10x relative to the deletion plot. See **Supplemental Note 1** for further details and analysis methods. (C) Fraction of total read pairs, excluding reads aligning to satellite regions, that are inversion, deletion, or duplication chimeras, at different chimera breakpoint distances (in 50bp bins). Deletion chimera plots exhibit peaks at 450-500bp and at 6,050-6,300bp in all sample types, corresponding to germline Alu and L1 deletion polymorphisms relative to the human genome reference (verified by analysis of reads, data not shown). Duplication plots exhibit a peak at 200bp in all sample types, corresponding to true germline tandem repeat polymorphisms (verified by analysis of reads, data not shown). (D) Fraction of total WGS read pairs, excluding read pairs aligning to satellite regions and after subtraction of unamplified bulk DNA traces, that are inversion, deletion, or duplication chimeras, at different chimera breakpoint distances (in 50bp bins). The total percentage of WGS read pairs are shown for each chimera type, calculated after excluding satellite reads and subtraction of the unamplified bulk DNA baseline. The Alu deletion peak was not completely normalized by the unamplified bulk DNA subtraction and was excluded from the total chimera calculation.

Figure S9. Model of chimera formation during MDA based on single-neuron WGS, Related to Figure 1. (A) Branch migration between the source strand and the reannealing strand liberates a free single-stranded 3' end of the source strand. Branch migration can progress in either direction, but in reverse direction can proceed only while the reannealing strand is single-stranded. (B) The distribution of branch migration distances is related to the distribution of chimera breakpoint distances, though the exact relationship may not be linear or simple due to flexibility of DNA. (C) Model of inversion chimera formation. (D) Model of deletion chimera formation. (E) Model of duplication chimera formation.

Figure S10. *scTea* detection sensitivity, Related to Figure 2. (A) Reproducibility of *scTea* insertion scores from cortex and heart bulk samples. (B) *scTea* non-reference insertion calls (with score ≥9) from cortex and heart bulk WGS samples show very high overlap. (C) Evaluation of *scTea* sensitivity using simulated WGS reads from the HuRef genome assembly. *scTea* sensitivity at different score cutoffs was evaluated for each retrotransposon family using HuRef-specific retrotransposon insertions reported by Xing, et al (2009) as a gold standard set. *scTea* also detected an additional 139 AluY, 35 L1Hs, and 6 SVA insertions that were not detected by Xing, et al (2009), mostly due to a lack of a poly-A tail or target-site duplication required for detection by Xing, et al and incomplete annotation of HuRef insertions by Levy, et al (2007). (D) Sensitivity of *scTea* for AluY and L1Hs germline known non-reference insertions at different score cutoffs. SVA sensitivity is not plotted due to the small number of insertions. Sensitivity was evaluated relative to a gold standard set consisting of all non-reference insertions detected in both heart and cortex bulk samples (i.e. high-confidence germline insertions) that were independently reported in prior studies of retrotransposon polymorphism (i.e. 'known' insertions, see Supplemental Experimental Procedures for details). Dashed line indicates score cutoff ≥9 used to call the final set of germline and somatic insertions in bulk and MDA samples. (E) Average sensitivity of *scTea* for germline known non-reference insertions at score cutoff ≥9 for each sample type (SD, error bars). Sensitivity for bulk, MDA 100-neuron, and MDA single neuron (all insertions) corresponds to plots shown in Figure S10D. *scTea* sensitivity was separately assessed for insertions present in a single-copy per genome (heterozygous and hemizygous), as a more accurate sensitivity estimate for somatic insertions. Copy number (zygosity) of insertions was determined by counting the number of read pairs spanning the insertion breakpoint in bulk WGS data (see Supplemental Experimental Procedures for details).

Figure S11. Score distribution of *scTea* insertion calls, Related to Figure 2. (A) Score distribution of all insertion calls from all retrotransposon families (AluY, L1Hs, SVA) in bulk samples (top panel), MDA 100 neuron sample (middle panel), and all 16 MDA single neuron samples (bottom panel). Score distributions are plotted separately for calls corresponding to known non-reference insertions reported in prior studies of retrotransposon polymorphism (see Supplemental Experimental Procedures for details) and calls not previously reported (unknown). The latter consists of both true positive insertions not previously reported and false positive insertion calls. Most known insertions have scores ≥9 used for calling the final set of germline and somatic insertions, while scores of many unknown insertion calls are below the score cutoff, indicating they are more likely false positives. Note in MDA samples the increased number of

unknown insertions at low scores deriving from MDA chimeras, as well as the wider distribution of known insertion scores mostly due to MDA GC-amplification bias. (B) Distribution of scores for all germline known non-reference insertions (AluY, L1Hs, SVA) found across all samples of each sample type (bulk, MDA 100-neurons, MDA single neurons). Distributions are plotted separately for insertions present in a single-copy per genome (heterozygous and hemizygous) and homozygous insertions present in two copies per genome. *scTea* score separates insertions of different copy number in bulk samples, whereas the separation is less prominent in MDA samples due to uneven amplification mostly driven by GC content. Nonetheless, the majority of known non-reference insertion calls in MDA samples have score ≥9 used for calling the final set of insertions. Germline insertions in these plots and analyses were defined as insertions detected in both cortex and heart bulk samples, as defined in the sensitivity analyses of *scTea* (**Figures S10D-E**), in order to obtain a high-confidence set of true-positive insertions. Copy number genotyping was determined by counting the number of read pairs spanning the insertion breakpoint (see Supplemental Experimental Procedures for details).

Figure S12. Somatic insertion L1#2 structure and single-neuron screening, Related to Figure 2. (A) Schematic of the structure of somatic insertion L1#2 as determined by cloning and sequencing of the full-length insertion from single neurons (**Figures 2F and S12D**; see **Table S3** for full sequence). L1#2 harbors an inversion, with the inversion point at position 5,233 relative to the RepBase L1Hs consensus sequence, a 5' truncation at position 3,520, a 614bp 3' transduction from the source L1, a poly-A tail, and an 18bp TSD. Features are not drawn to scale. Blue arrows and dashed lines illustrate PCR assays used for validation. The source L1 is a full-length, intact L1Hs retrotransposon on chromosome 13 (chr13: 30,215,844-30,221,843, hg19) that is heterozygous in individual 1465 and polymorphic in the population (data not shown). Red shading in the beginning of the 3' transduction sequence (orange) represents transduction of the germline poly-A tail from the source L1. The full-length source L1 was cloned from individual 1465 bulk cortex, cerebellum, and heart, and single-neuron 6 (data not shown), and its sequence was compared to the sequence of L1#2. The consensus sequences of the source L1 and L1#2 were identical except for the inversion and truncation in L1#2. Additionally, the source L1 sequence was identical among all the tested 1465 samples, but it harbored 4 silent mutations relative to the human genome reference, 2 of which were retrotransposed and present as expected in L1#2 (see full L1#2 sequence document in **Table S3** for details). Furthermore, the germline poly-A tail of the source L1 locus and the transduced copy of this poly-A tail in L1#2 exhibited variability and were on average 22bp in size, which is shorter than the 24bp in the human genome reference. The variability seen in sequencing of this germline poly-A tail and its transduced copy may be due to somatic mosaicism, MDA, PCR, or TOPO cloning. (B) Sequences of the 5' and 3' junctions of the L1#2 insertion site. (C) Representative gels and results of 3'PCR screening of all single-neurons sorted and amplified from the cerebral cortex, caudate nucleus, and cerebellum. L1#2 was found in cerebral cortex single-neurons #6, 18, 22, 32, 45, 79, 248, 271, 278, 289, 326, 370, and 531. 3'PCR did not detect L1#2 in caudate or cerebellum single-neurons, even though it is present in these tissues, due to low mosaicism and sample size. (D) Full-length cloning of L1#2 from all 13 cortical single-neurons identified in the 3'PCR screen. Variability in intensity of PCR products in single neurons is due to variable amplification of alleles by MDA. The band below the insertion allele in the second to last single-neuron is an off-target band (data not shown). The L1#2 insertion was cloned and sequenced in its entirety from all 13 single-neurons. Comparison of L1#2 sequence

among all the neurons showed no differences except for a G insertion in a 3'UTR short poly-G tract (poly-G_6 to poly-G_7) in single-neuron 531. This may be a true somatic mutation in neuron 531, but due to limited numbers of full-length clones from this neuron it is not possible to exclude MDA or PCR artifact. See L1#2 sequence (**Table S3**) for details.

Figure S13. Droplet digital PCR (ddPCR) assays of somatic L1 insertions, Related to Figure 3. (A) Somatic L1 insertion structures showing ddPCR primer and probe locations. L1 insertions and sequence features not drawn to scale. (B) ddPCR with different known input amounts of a synthetic L1#1 oligo shows highly linear measurement of L1#1 concentration (R^2 = 0.998). Greater than 98% of L1-containing droplets contain only 1 copy of the L1#1 oligo by Poisson sampling statistics, confirming single-copy sensitivity of ddPCR. (C) Representative somatic L1 insertion ddPCR assay plots in MDA-amplified single neurons (quantifying genomic copy number after MDA amplification), and unamplified bulk cerebral cortex, caudate nucleus, and cerebellum samples (quantifying mosaicism) from individual 1465. $L1^+$ droplets in bulk samples are plotted with larger points for better visualization.

Figure S14. Brain sections, locations of the cortex, and tissues sampled from individual 1465. (A) Diagram of coronal sections of the brain of individual 1465. Sections and locations were scaled and mapped to a representative brain image of a different individual, since an image of the brain of individual 1465 prior to sectioning was not available. Mapping of sections and locations was based on photographs and measurements of the brain sections of individual 1465. Sections sampled for this study are underlined and sampled locations within them are labeled with letters (A to Z). The caudate nucleus sample was obtained from the interior of section 9 (dashed triangle). Location 'D' (underlined) was the source of single-neuron samples. Locations in which L1#1 was detected are highlighted in blue. A representative photograph of a brain section (section 4) is shown on the right, with a corresponding lucida tracing. See Supplemental Experimental Procedures for further details. (B) Lucida tracings of all cerebral cortex sections sampled from individual 1465, traced from photographs of sections. Sections 2, 3 and 4 are also shown in Figure 3C. Tracings show the posterior surface of each section, except sections 6, 10, and 12 whose anterior surfaces were traced and then mirror imaged to match the orientation of the other sections. Dashed lines indicate regions that were not present in photographs of sections due to sampling prior to this study. Anatomy of these regions was extrapolated based on records of sampled locations, adjacent sections, photographs of right hemisphere formalin-fixed sections, and atlases of normal brain anatomy. Locations in which L1#1 was detected are highlighted in blue.

Figure S15. Bulk nested 3'PCR confirms presence/absence of somatic L1 insertions measured by ddPCR, Related to Figure 3. Bulk nested 3'PCR differs from digital nested 3'PCR used later for poly-A tail sizing in that DNA is not diluted for single-copy cloning. Variability in bulk nested 3'PCR product sizes is due to polymorphism among the poly-A tails amplified in each reaction, but is not a reliable measure of the true poly-A tail distribution (see Supplemental Experimental Procedures for details). (A) Somatic L1 insertion structures showing bulk nested 3'PCR primer locations. L1 insertions and sequence features not drawn to scale. (B) L1#1 bulk nested 3'PCR of unamplified bulk DNA from all tissues and locations sampled from

individual 1465. Correct PCR product of all positive reactions was confirmed by Sanger sequencing. Additional replicate assays of these tissues were concordant with these results (data not shown). Numbered subscripts indicate independent DNA extractions from the same tissue location. Spinal cord 'L' and 'R' subscripts indicate left and right. (C) L1#2 bulk nested 3'PCR of unamplified bulk DNA from all tissues and locations sampled from individual 1465. Cerebral cortex locations O and Y were repeated, as these did not show PCR product in the initial assay due to their lower mosaicism. Product is not detected in every cerebellum assay due to extremely low mosaicism. Additional replicate assays of tissues shown here were concordant with these results, with additional detection of L1#2 in cerebellum sample A_1 (data not shown). Nested PCR assays failed to detect L1#1 and L1#2 in 3 independent DNA extractions from each of 4 locations of the formalin-fixed right cerebral cortex (**Table S4**; gels not shown), though lack of detection may be due to low efficiency of DNA extraction and PCR from formalin-fixed tissue. Correct PCR product of positive cerebellum reactions was confirmed by Sanger sequencing. Correct PCR product in other cortex and caudate locations was confirmed by Sanger sequencing of digital nested 3'PCR assays (**Figure S16A**). Numbered subscripts indicate independent DNA extractions from the same tissue location.

Figure S16. Digital nested 3'PCR cloning of single somatic retrotransposon poly-A tails, Related to Figure 4. (A) Schematic of digital nested 3'PCR. Primers are the same as used for bulk nested 3'PCR (**Figure S15A**), except that round 2 PCR is performed with a FAM-labeled primer to allow precise sizing of poly-A tails by capillary electrophoresis. A subset of poly-A tails cloned from every tissue and location was Sanger sequenced and confirmed 100% specificity of the assay to clone the target retrotransposon poly-A tail (i.e. the genomic flank sequence breakpoint always matched the target retrotransposon insertion site). (B) dnPCR histograms (in 3 bp bins) of L1#1 poly-A tails of known size. Y-axis is the fraction of all poly-A tails cloned from each poly-A tail of known size. Note that these control experiments overestimate the degree to which dnPCR mutates the poly-A tail, as the input poly-A tails previously underwent round 1 of dnPCR (see Supplemental Experimental Procedures for details).

Figure S17. L1#2 poly-A tail size distributions in 12 brain locations determined by digital nested 3'PCR, Related to Figure 4. (A) Size histogram (in 1 bp bins) of all L1#2 poly-A tails (n=1,562) cloned by dnPCR from all 12 locations in the cerebral cortex, caudate nucleus, and cerebellum. Inset y-axis zoom shows smaller peaks of somatic polymorphism. (B) Size histograms (in 1 bp bins) of L1#2 poly-A tails cloned by dnPCR from each of 12 locations in the cerebral cortex, caudate nucleus, and cerebellum. Dashed lines in cerebral cortex-location D, from which single-neurons were isolated, indicate poly-A tail sizes seen in 13 single-neurons with the L1#2 insertion (12/13 neurons: 113-115bp; 1/13 neurons: 38bp; see **Table S5**).

Figure S18. B**reast cancer-specific L1 insertion poly-A tail sizes are distinct in primary tumor versus metastasis, Related to Figure 4.** (A) Schematic of the primary tumor and metastasis locations in the left medial breast and left axilla, respectively. Normal blood DNA was also obtained from the individual. (B) Schematic of the tumor-specific L1Hs insertion, 3.3kb in size and in an intergenic location, containing an inversion, 5' truncation, TSD, and poly-A tail.

Primers used for full-length cloning and nested 3'PCR (bulk and digital) are shown. Features are not drawn to scale. (C) Full-length PCR confirms the insertion is present in both primary tumor and metastasis, but not normal blood DNA from the individual. The full insertion was cloned and sequenced (see **Table S3** for full sequence). (D) Bulk (non-digital) nested 3'PCR (i.e. DNA was not diluted to < 1 copy per reaction) reveals an additional product in primary tumor corresponding to a shorter poly-A tail (36bp). Precise sizing by capillary electrophoresis further reveals that the larger poly-A tail peak is 81bp in primary tumor versus 80bp in the metastasis. Bulk nested 3'PCR sizing results were confirmed in 2 independent experiments (data not shown). Note, as described in Supplemental Experimental Procedures, due to increased amplification efficiency of shorter poly-A tails, the relative intensities of PCR products in bulk nested 3'PCR are not an accurate measure of the true poly-A tail size distribution. (E) Size histograms (in 1 bp bins) of the tumor-specific L1 poly-A tails cloned by digital nested 3'PCR. A peak is seen at 36 bp (starred) only in primary tumor. Note again the 1 bp difference between peaks in primary tumor (81bp) and metastasis (80 bp), respectively. This difference is reliable as discussed above, since these sizes were measured in 2 independent bulk nested 3'PCR experiments (**Figure S18D** and data not shown).

Figure S19. Poly-A microsatellite length distributions in the human genome, and relationship to retrotransposon loci, Related to Figure 4. (A) Number of microsatellite loci with motif lengths between 1 to 6 annotated by Tandem Repeats Finder in the human genome reference. Motifs of lengths 1 to 3 are shown individually. Serial permutations and reverse complements of each motif are grouped (i.e. ACT = ACT, CTA, TAC, AGT, GTA, TAG). (B) Number of genome bases covered by microsatellite loci of each motif as in Figure S19A. (C) Length distributions of all poly-A loci annotated by Tandem Repeats Finder in the human genome (minimum 25bp poly-A size), for all loci, and separately for loci overlapping L1, Alu, or SVA elements by at least half their length. (D) Length distributions of poly-A loci of at least 95% purity, analyzed as in Figure S19C. See **Supplemental Note 2** for further details.

Table S1. Sequencing and alignment statistics of all samples in this study, Related to Figure 1. All fraction statistics were calculated after filtering PCR duplicate reads using as the denominator the total number of non PCR-duplicate reads, except for: a) the fraction of PCR duplicates, which was calculated out of the total number of reads; and b) chimera fractions, which were calculated out of the total number of non PCR-duplicate chimera plus concordant read pairs aligning to autosomes and sex chromosomes (excluding satellite DNA regions). Samples from prior high-coverage WGS single-cell studies are included for comparison. All statistics were highly similar when calculated prior to PCR duplicate removal (data not shown). See **Supplemental Note 1** for further details.

Table S2. Genome coverage statistics of all samples in this study, Related to Figure 1. The 'Coverage by sample' sheet shows coverage (read depth) statistics for each sample for different annotated genomic features relative to the genome-wide average coverage. The 'Coverage by category' sheet shows summary of these statistics for each category of samples. The 'GC-content of features' sheet shows average GC content of each genomic feature across the genome. Statistics for each genomic feature include the fraction of bases of that feature covered by at least

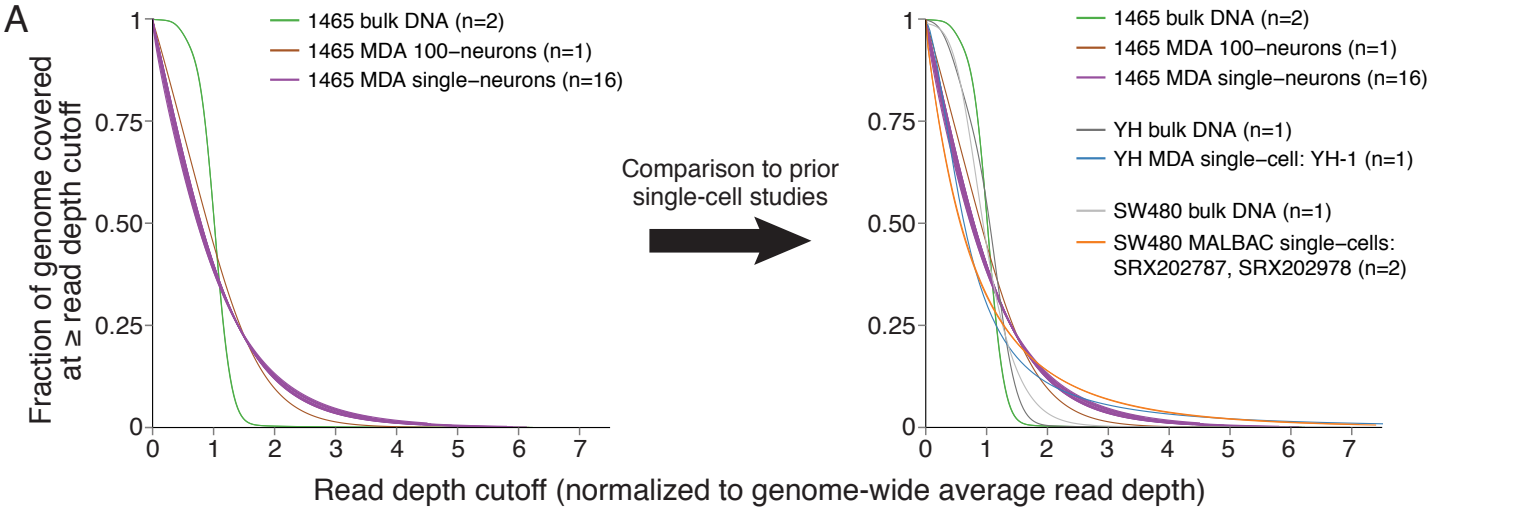
1, 5 or 10 reads (i.e. 1x, 5x, or 10x coverage), as well the fraction of bases not covered by any reads (0x, locus dropout). Samples from prior high-coverage WGS single-cell studies are included for comparison. See **Supplemental Note 1** for further details.

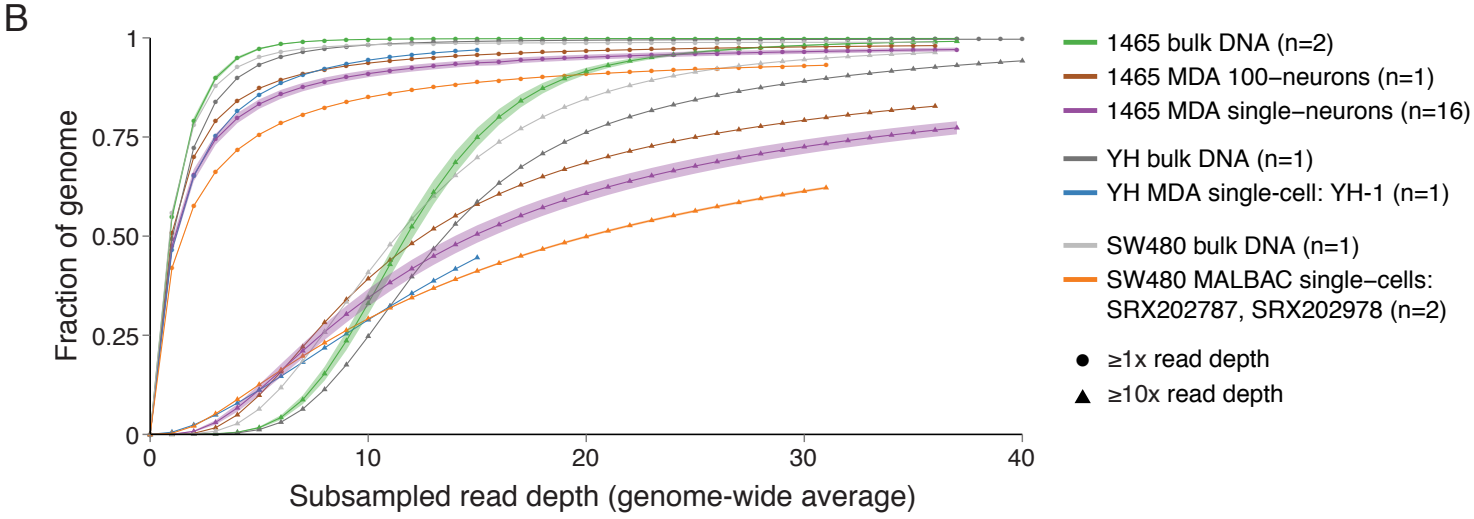
Table S3. Bulk sample and somatic single-cell retrotransposon insertion candidates, Related to Figure 2. Retrotransposon insertion candidates called in 1465 heart bulk DNA used for *scTea* specificity measurement ('Bulk sample candidates' sheet), and somatic insertion candidates called by *scTea* ('Somatic single-cell candidates' sheet). Includes primers used for validation and full-length sequences of validated somatic insertions. Below the list of insertion candidates are further descriptions of columns.

Table S4. Somatic L1 insertion mosaicism measured by ddPCR in tissues from individual 1465, Related to Figure 3. Also includes locations assayed in formalin-fixed right hemisphere by nested 3'PCR and unrelated human DNA controls.

Table S5. Sizes of all cloned somatic L1 poly-A tails, Related to Figure 4. Sizes of all poly-A tails cloned by digital nested 3'PCR (dnPCR) and measured by capillary electrophoresis for L1#1, L1#2, the breast cancer-specific somatic L1, and control L1#1 poly-A tails of known size. Also included are L1#1 and L1#2 poly-A tail sizes of all single-neurons with the insertions as measured by non-digital (i.e. DNA was not diluted to <1 copy per reaction) nested 3'PCR. Non-digital nested 3'PCR of single-neuron DNA measures poly-A tails with the expectation that the predominant species amplified by single-neuron MDA is the true genotype. See 'Supplemental Experimental Procedures' for details.

Figure S1


A
Fraction of genome covered at ≥ read depth cutoff
1465 bulk DNA (n=2)
1465 MDA 100–neurons (n=1)
1465 MDA single–neurons (n=16)
Comparison to prior single-cell studies
YH bulk DNA (n=1)
YH MDA single–cell: YH-1 (n=1)
SW480 bulk DNA (n=1)
SW480 MALBAC single–cells: SRX202787, SRX202978 (n=2)
Read depth cutoff (normalized to genome-wide average read depth)


B
Fraction of genome
1465 bulk DNA (n=2)
1465 MDA 100–neurons (n=1)
1465 MDA single–neurons (n=16)
YH bulk DNA (n=1)
YH MDA single-cell: YH-1 (n=1)
SW480 bulk DNA (n=1)
SW480 MALBAC single–cells: SRX202787, SRX202978 (n=2)
≥1x read depth
≥10x read depth
Subsampled read depth (genome-wide average)

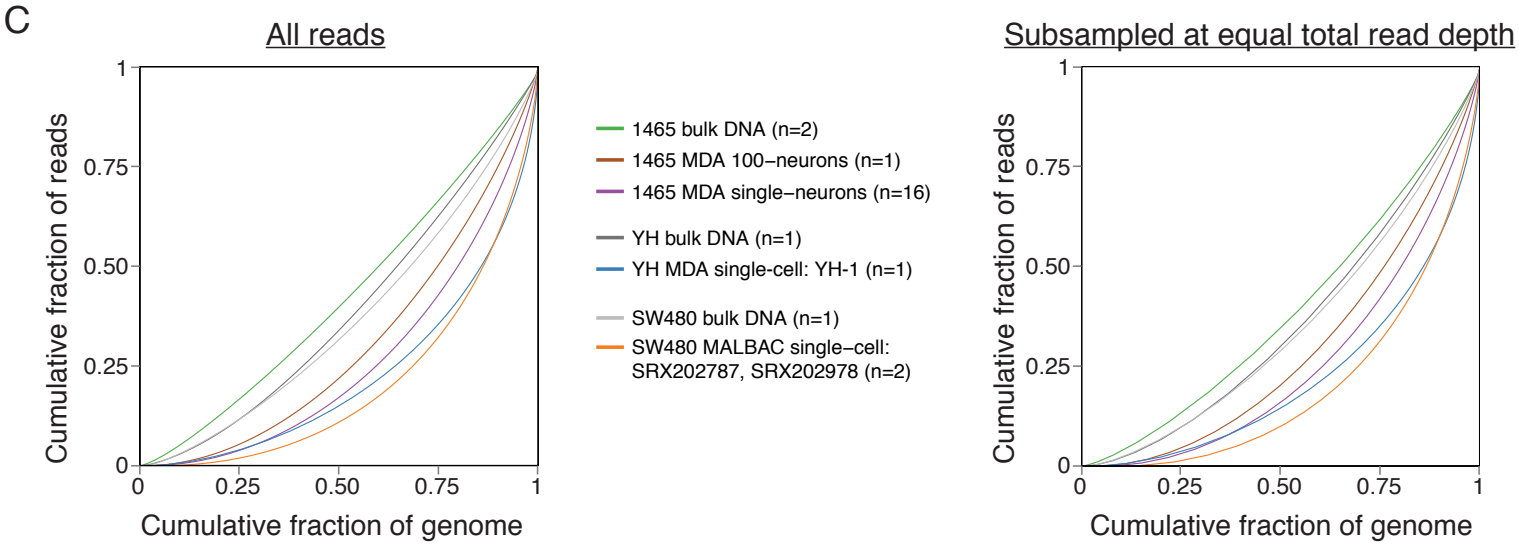
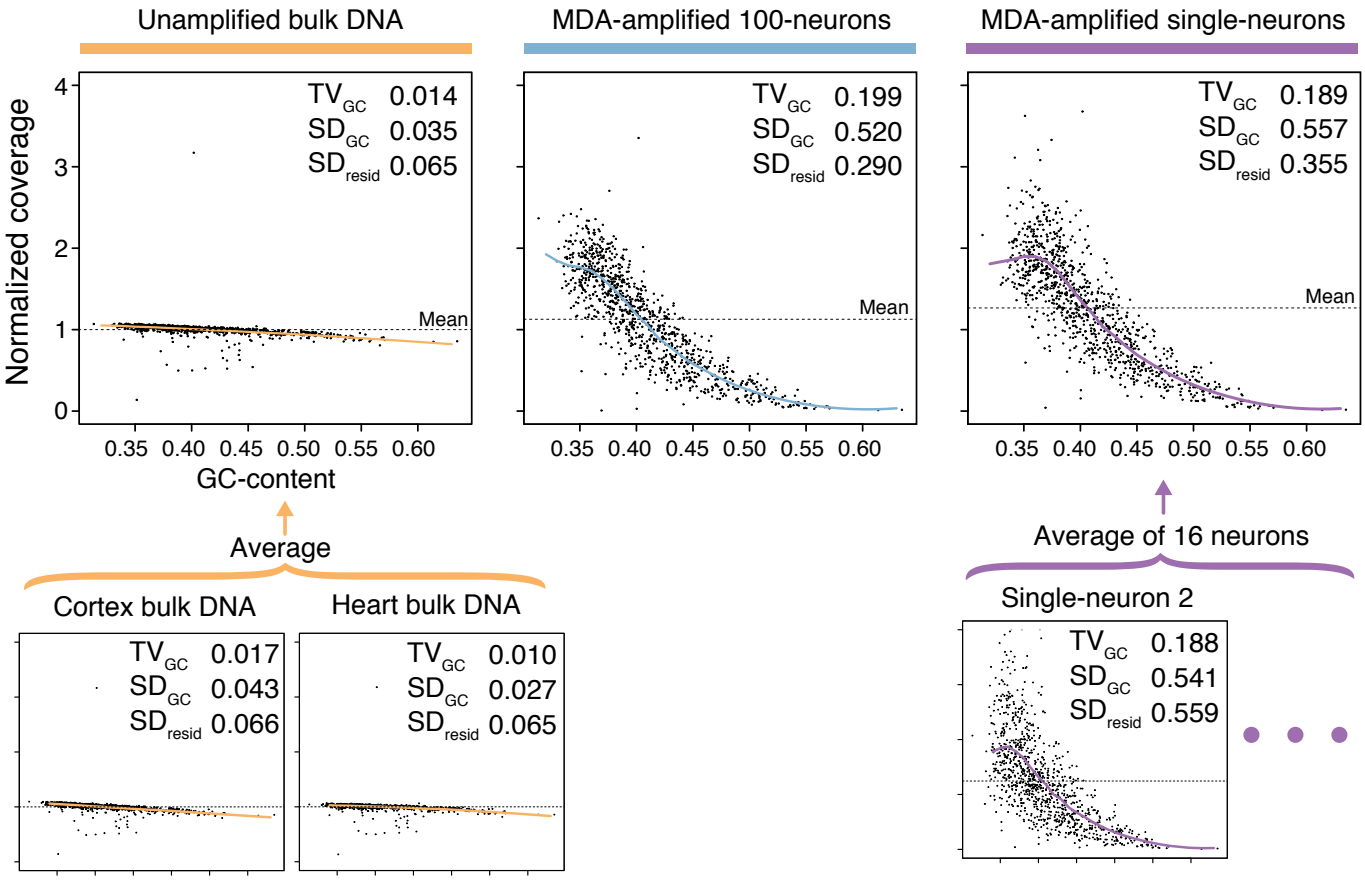

C
All reads
Subsampled at equal total read depth
Cumulative fraction of reads
1465 bulk DNA (n=2)
1465 MDA 100–neurons (n=1)
1465 MDA single–neurons (n=16)
YH bulk DNA (n=1)
YH MDA single-cell: YH-1 (n=1)
SW480 bulk DNA (n=1)
SW480 MALBAC single–cell: SRX202787, SRX202978 (n=2)
Cumulative fraction of genome

Figure S2

A



B

Samples evaluated	Reference used for correction	MAPD mean (SD)	MDAD mean (SD)
Unamplified bulk DNA	None (uncorrected)	0.03 (0.00)	0.04 (0.01)
Unamplified bulk DNA	GC/mappability modeling[1]	0.02 (0.00)	0.02 (0.00)
MDA 100-neuron	None (uncorrected)	0.32	0.57
MDA 100-neuron	GC/mappability modeling[1]	0.26	0.23
MDA 100-neuron	Pooled 16 MDA single neurons[2]	0.17	0.22
MDA single neurons	None (uncorrected)	0.55 (0.03)	0.73 (0.05)
MDA single neurons	GC/mappability modeling[1]	0.50 (0.03)	0.44 (0.04)
MDA single neurons	MDA 100-neuron[2]	0.43 (0.03)	0.40 (0.05)
MDA single neurons	Pooled 15 MDA single neurons[2]	0.41 (0.03)	0.37 (0.05)

[1]Non-paired correction
[2]Paired correction

C

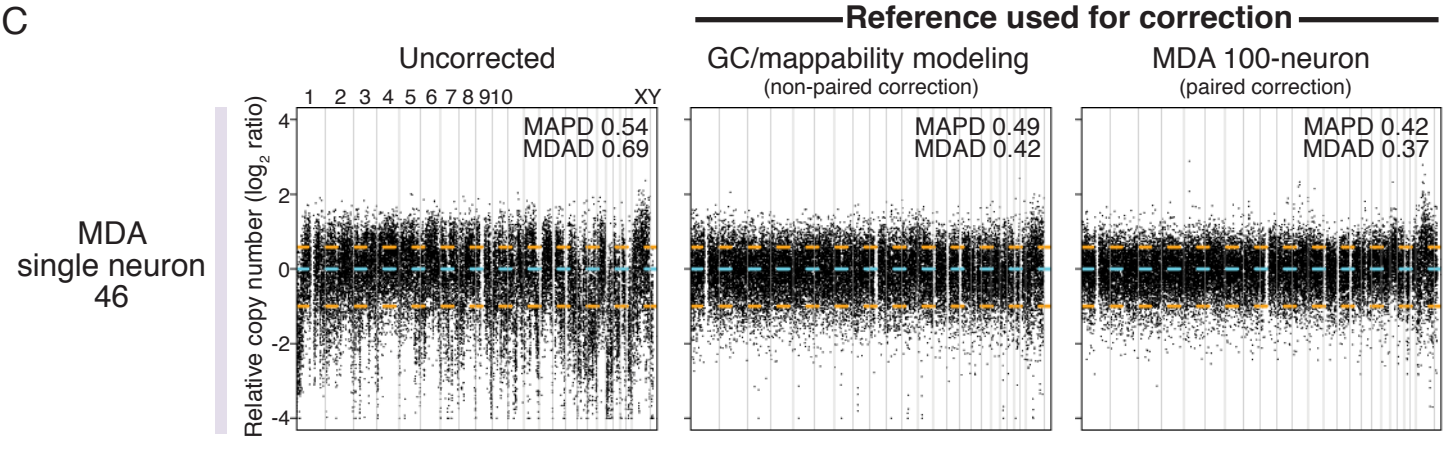


Figure S3

A

Samples evaluated	Sample type	Reference	MAPD mean (SD)	MDAD mean (SD)
1465-cortex bulk DNA; 1465-heart bulk DNA	Unamplified bulk DNA	Simulated reads	0.02 (0.00)	0.07 (0.02)
1465-cortex bulk DNA; 1465-heart bulk DNA	Unamplified bulk DNA	Unamplified bulk DNA (excluding test sample)	0.02 (0.00)	0.02 (0.00)
1465-caudate 100-neurons	MDA 100-cell	Simulated reads	0.34	0.58
1465-caudate 100-neurons	MDA 100-cell	Unamplified bulk DNA	0.33	0.51
1465-caudate 100-neurons	MDA 100-cell	Pooled 8 MDA single neurons	0.20	0.20
Each of 16 1465-cortex single neurons	MDA single-cell	Simulated reads	0.60 (0.03)	0.75 (0.04)
Each of 16 1465-cortex single neurons	MDA single-cell	Unamplified bulk DNA	0.59 (0.03)	0.70 (0.04)
Each of 16 1465-cortex single neurons	MDA single-cell	MDA 100-neuron	0.43 (0.03)	0.42 (0.05)
Each of 16 1465-cortex single neurons	MDA single-cell	Pooled 15 MDA single neurons (each time excluding test sample)	0.40 (0.03)	0.40 (0.05)

B

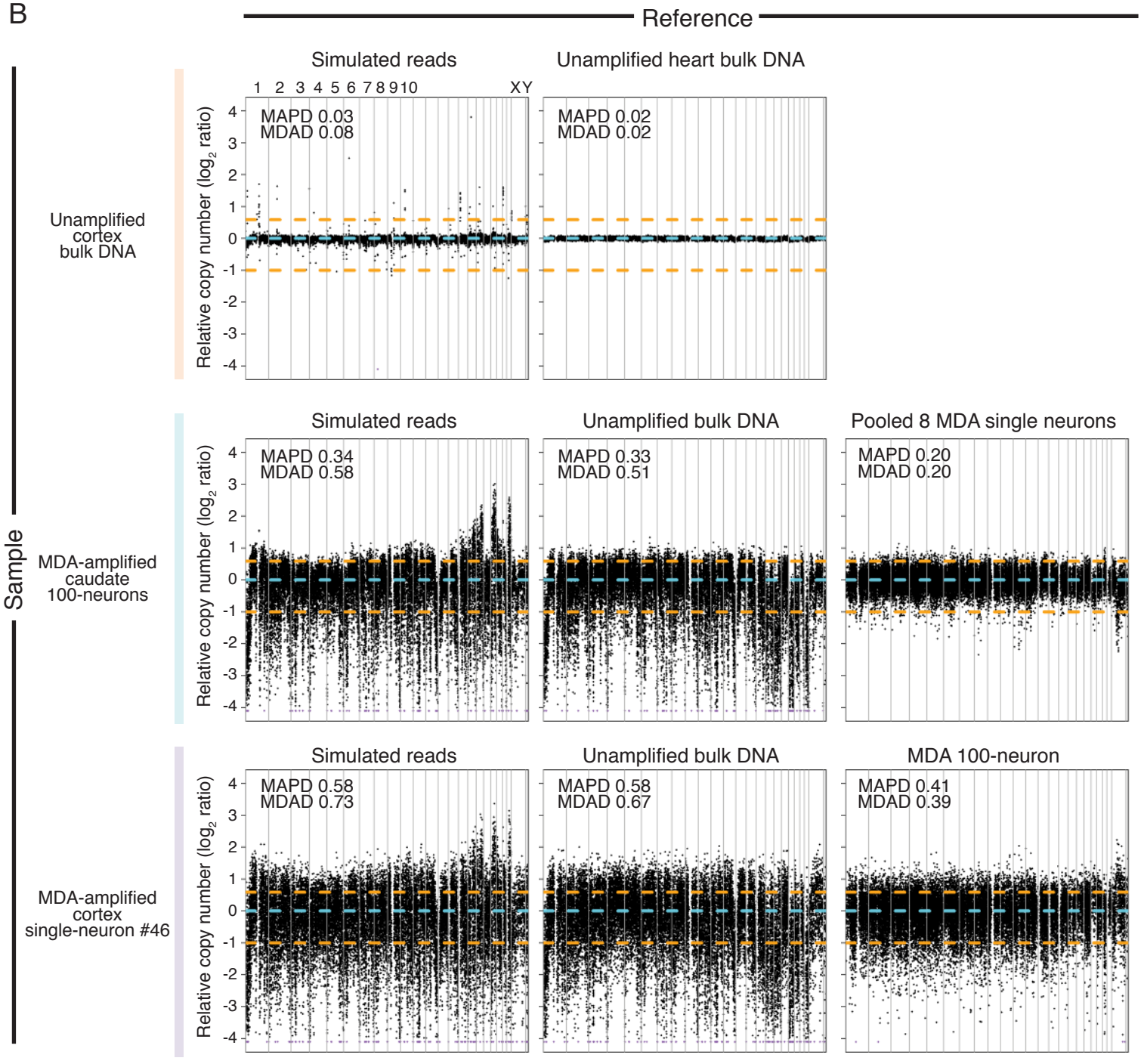
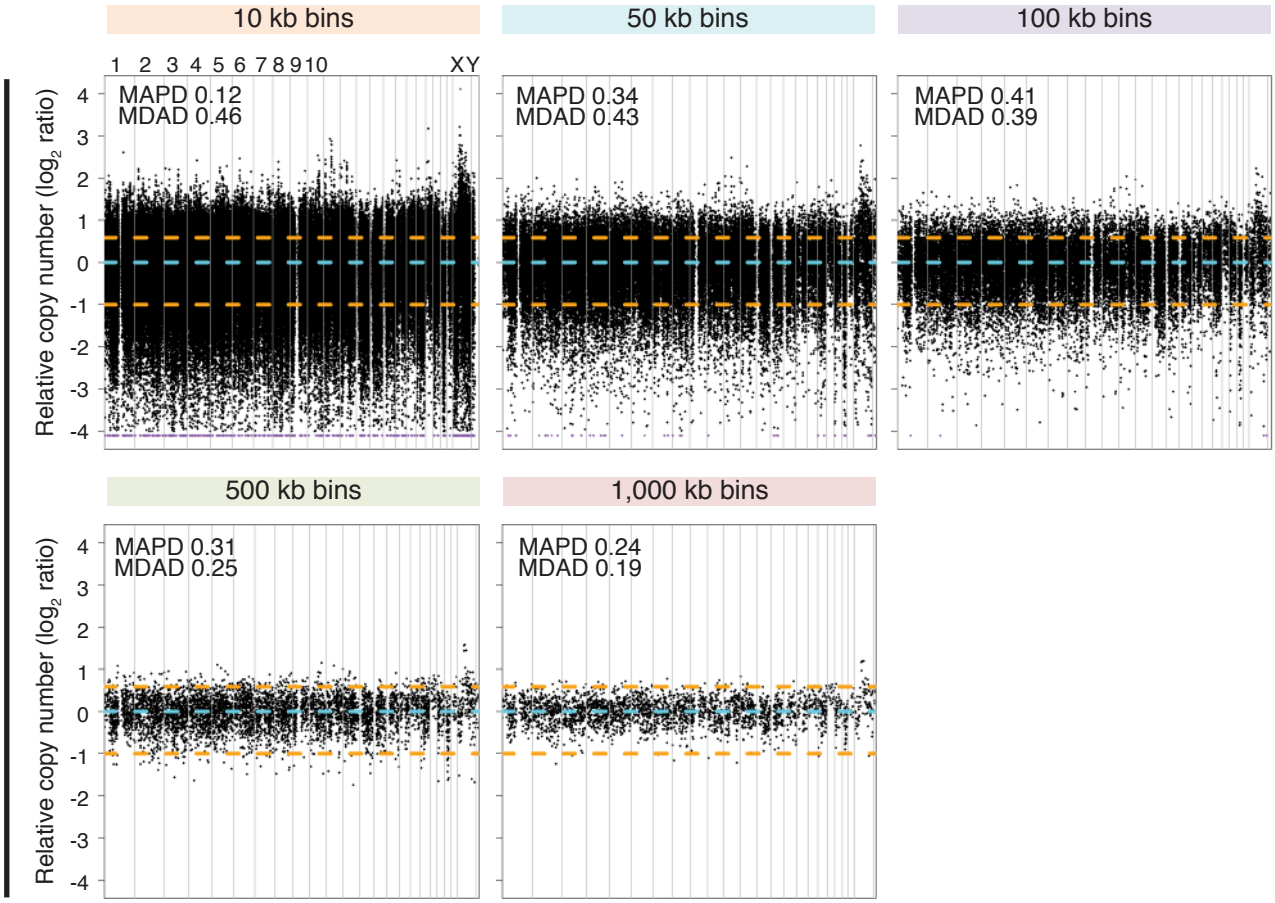


Figure S4

A

Samples evaluated	Reference	Bin size (kb)	MAPD mean (SD)	MDAD mean (SD)
Each of 16 1465-cortex single neurons	MDA 100-cell	10	0.13 (0.01)	0.49 (0.05)
Each of 16 1465-cortex single neurons	MDA 100-cell	50	0.35 (0.02)	0.46 (0.05)
Each of 16 1465-cortex single neurons	MDA 100-cell	100	0.43 (0.03)	0.42 (0.05)
Each of 16 1465-cortex single neurons	MDA 100-cell	500	0.33 (0.04)	0.27 (0.04)
Each of 16 1465-cortex single neurons	MDA 100-cell	1,000	0.27 (0.04)	0.21 (0.03)

B



C

Samples evaluated	Subsampled read depth	MAPD mean (SD)		MDAD mean (SD)	
		50kb bins	500kb bins	50kb bins	500kb bins
Each of 16 1465-cortex single neurons	0.1x	0.51 (0.02)	0.35 (0.04)	0.52 (0.05)	0.28 (0.04)
Each of 16 1465-cortex single neurons	0.5x	0.39 (0.02)	0.33 (0.04)	0.47 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	1x	0.37 (0.02)	0.33 (0.04)	0.47 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	5x	0.36 (0.02)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	10x	0.35 (0.02)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	15x	0.35 (0.02)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	20x	0.35 (0.02)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	25x	0.35 (0.03)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)
Each of 16 1465-cortex single neurons	30x	0.35 (0.03)	0.33 (0.04)	0.46 (0.05)	0.27 (0.04)

D

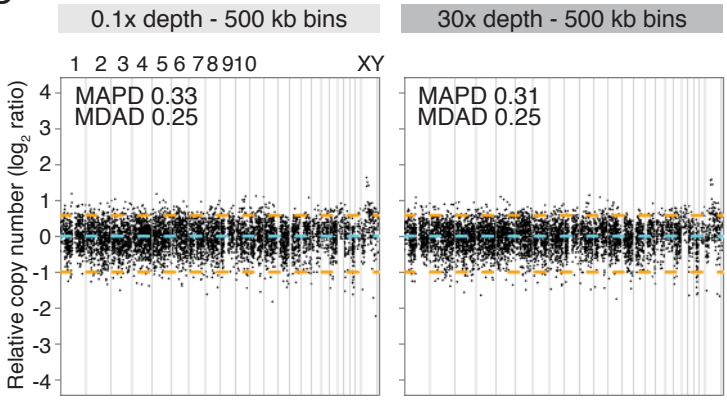


Figure S5

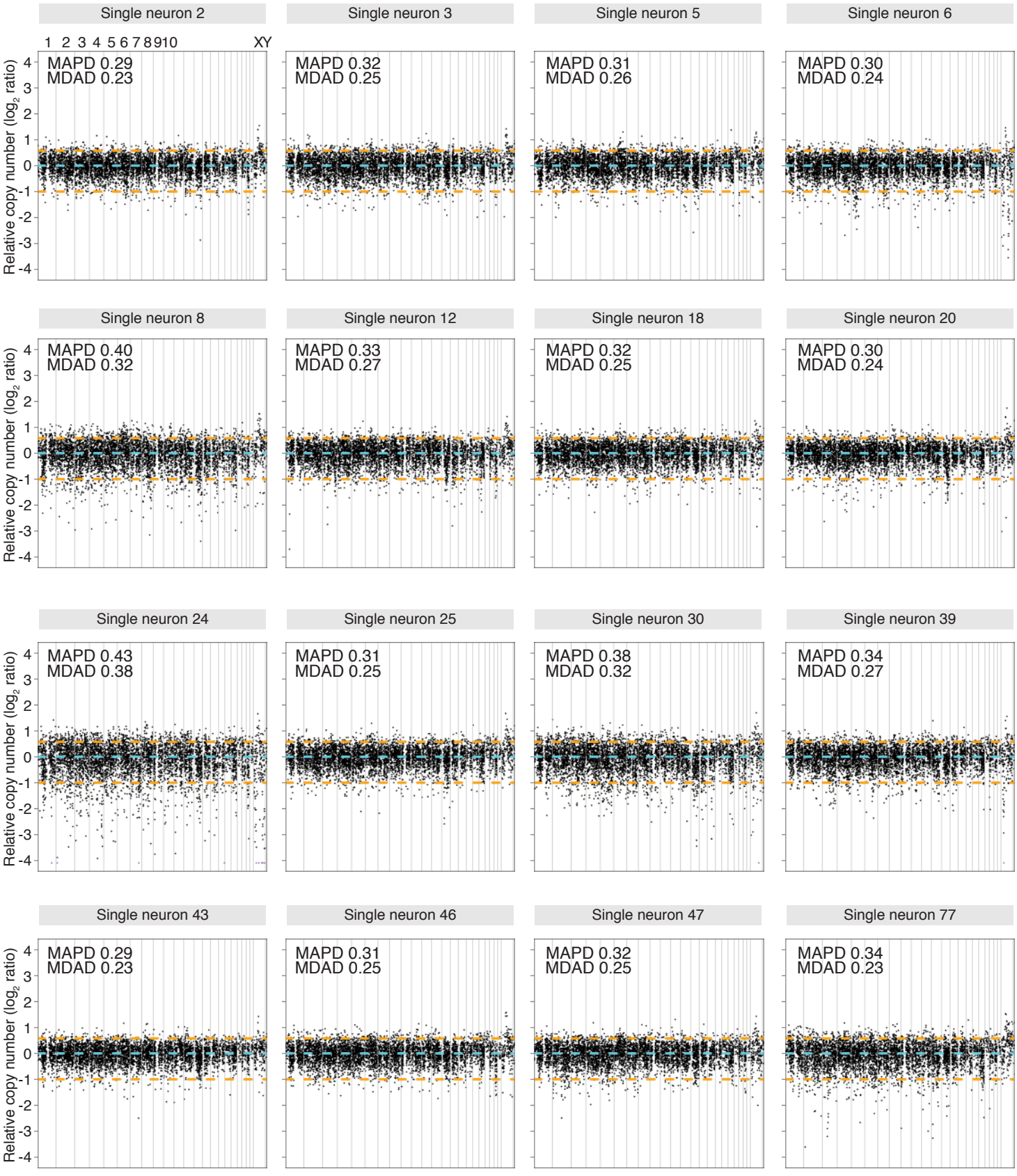
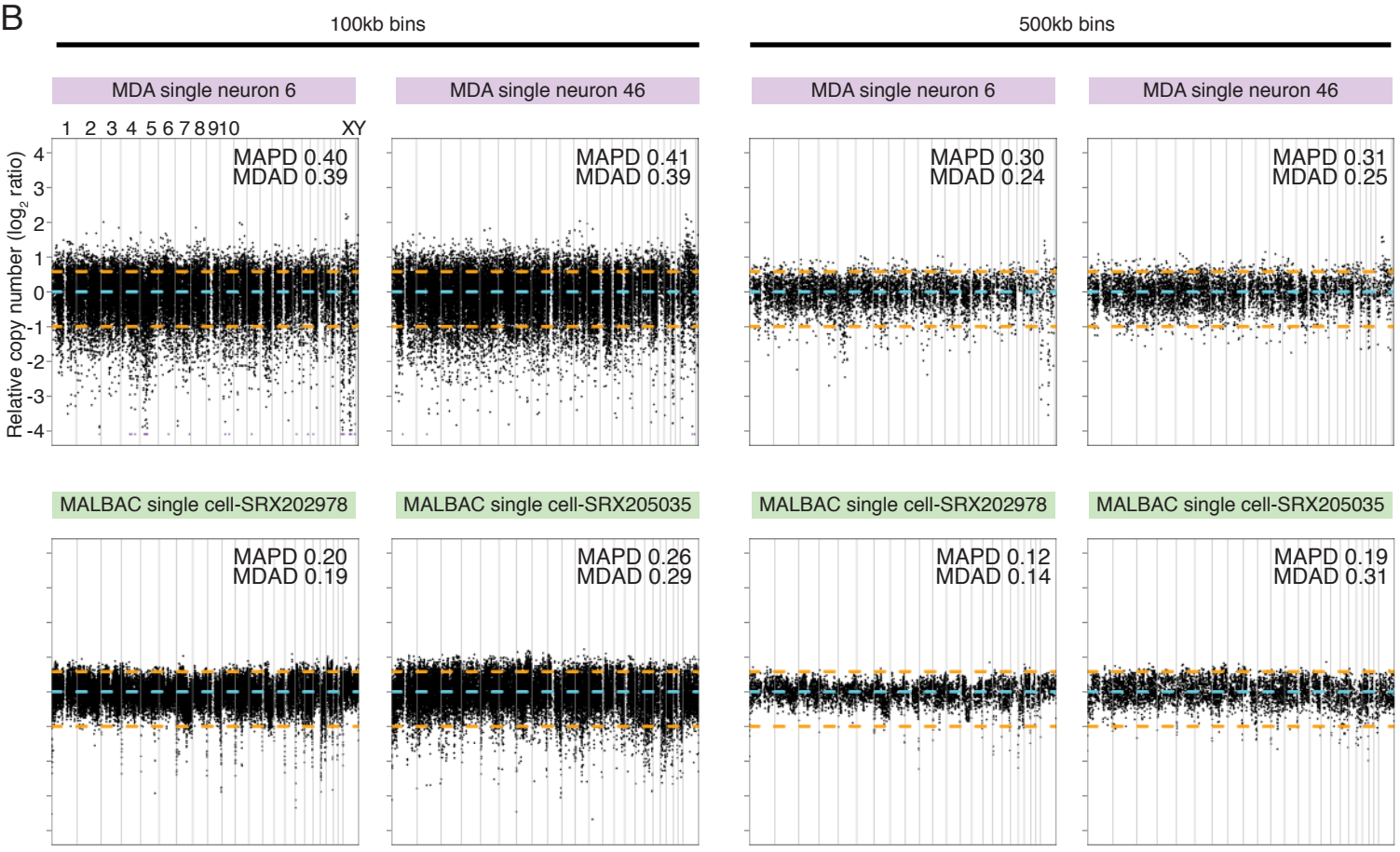

Single neuron 2
MAPD 0.29
MDAD 0.23
Single neuron 3
MAPD 0.32
MDAD 0.25
Single neuron 5
MAPD 0.31
MDAD 0.26
Single neuron 6
MAPD 0.30
MDAD 0.24
Single neuron 8
MAPD 0.40
MDAD 0.32
Single neuron 12
MAPD 0.33
MDAD 0.27
Single neuron 18
MAPD 0.32
MDAD 0.25
Single neuron 20
MAPD 0.30
MDAD 0.24
Single neuron 24
MAPD 0.43
MDAD 0.38
Single neuron 25
MAPD 0.31
MDAD 0.25
Single neuron 30
MAPD 0.38
MDAD 0.32
Single neuron 39
MAPD 0.34
MDAD 0.27
Single neuron 43
MAPD 0.29
MDAD 0.23
Single neuron 46
MAPD 0.31
MDAD 0.25
Single neuron 47
MAPD 0.32
MDAD 0.25
Single neuron 77
MAPD 0.34
MDAD 0.23
1 2 3 4 5 6 7 8 9 10 XY
Relative copy number (log2 ratio)
4
3
2
1
0
-1
-2
-3
-4

Figure S6

A

Samples evaluated	Reference	Bin size (kb)	MAPD mean (SD)	MDAD mean (SD)
16 MDA cortex single neurons	MDA 100-cell	100	0.43 (0.03)	0.42 (0.05)
3 MALBAC single cells	MALBAC single-cell	100	0.27 (0.08)	0.29 (0.10)
16 MDA cortex single neurons	MDA 100-cell	500	0.33 (0.04)	0.27 (0.04)
3 MALBAC single cells	MALBAC single-cell	500	0.18 (0.06)	0.23 (0.10)

B



C

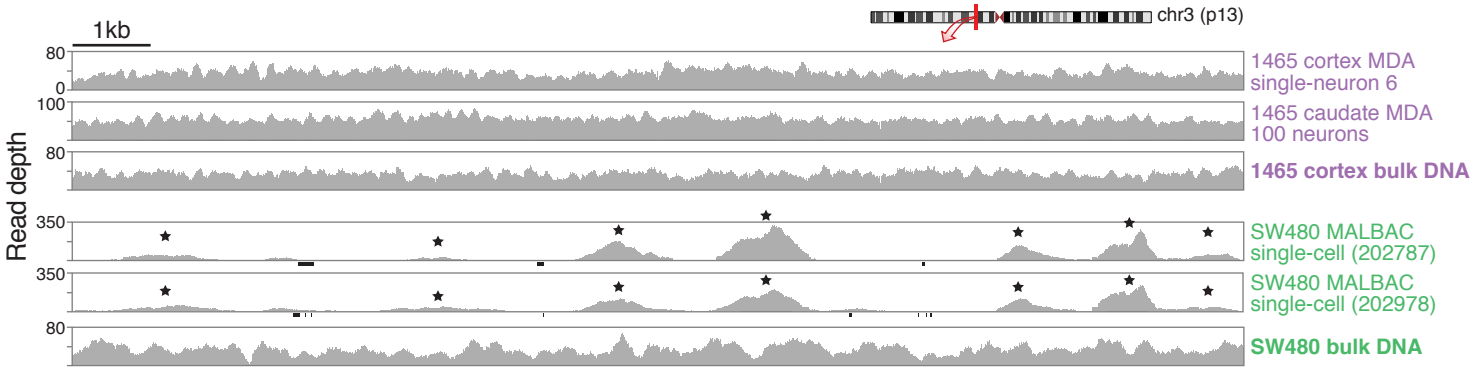


Figure S7

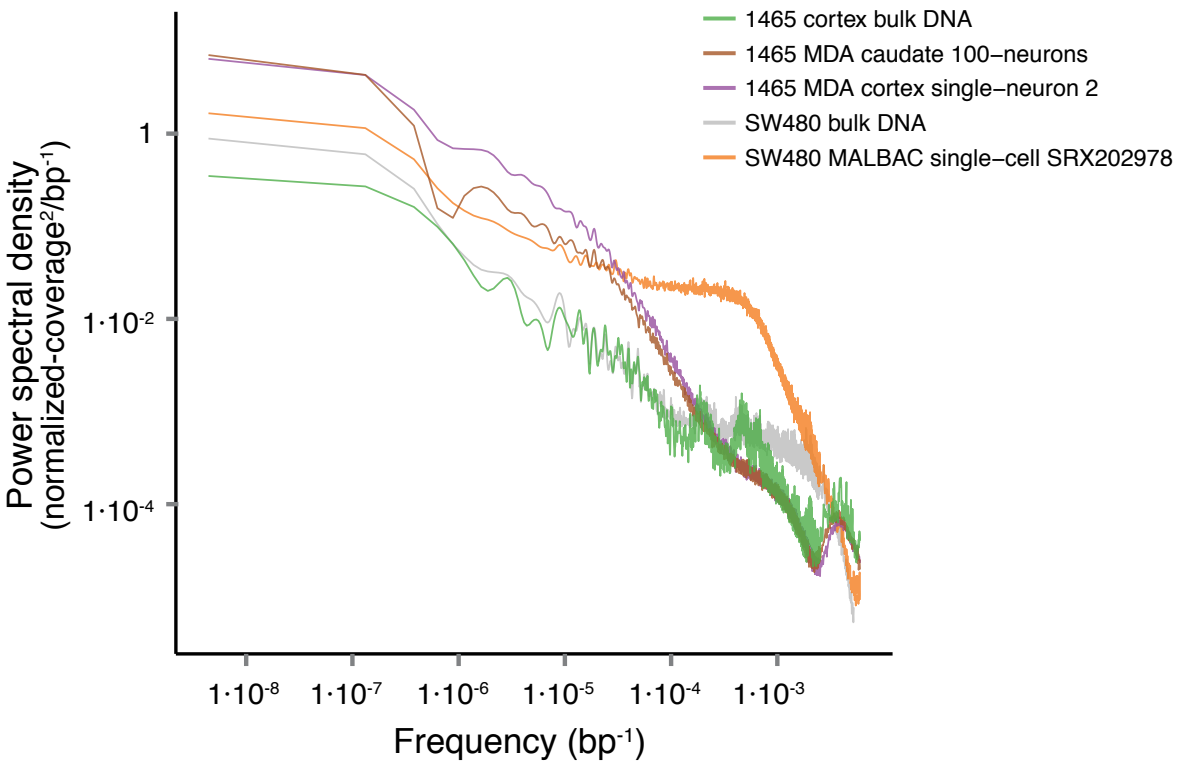

1465 cortex bulk DNA
1465 MDA caudate 100−neurons
1465 MDA cortex single−neuron 2
SW480 bulk DNA
SW480 MALBAC single−cell SRX202978
Power spectral density (normalized-coverage²/bp⁻¹)
1
1·10⁻²
1·10⁻⁴
1·10⁻⁸
1·10⁻⁷
1·10⁻⁶
1·10⁻⁵
1·10⁻⁴
1·10⁻³
Frequency (bp⁻¹)

Figure S8

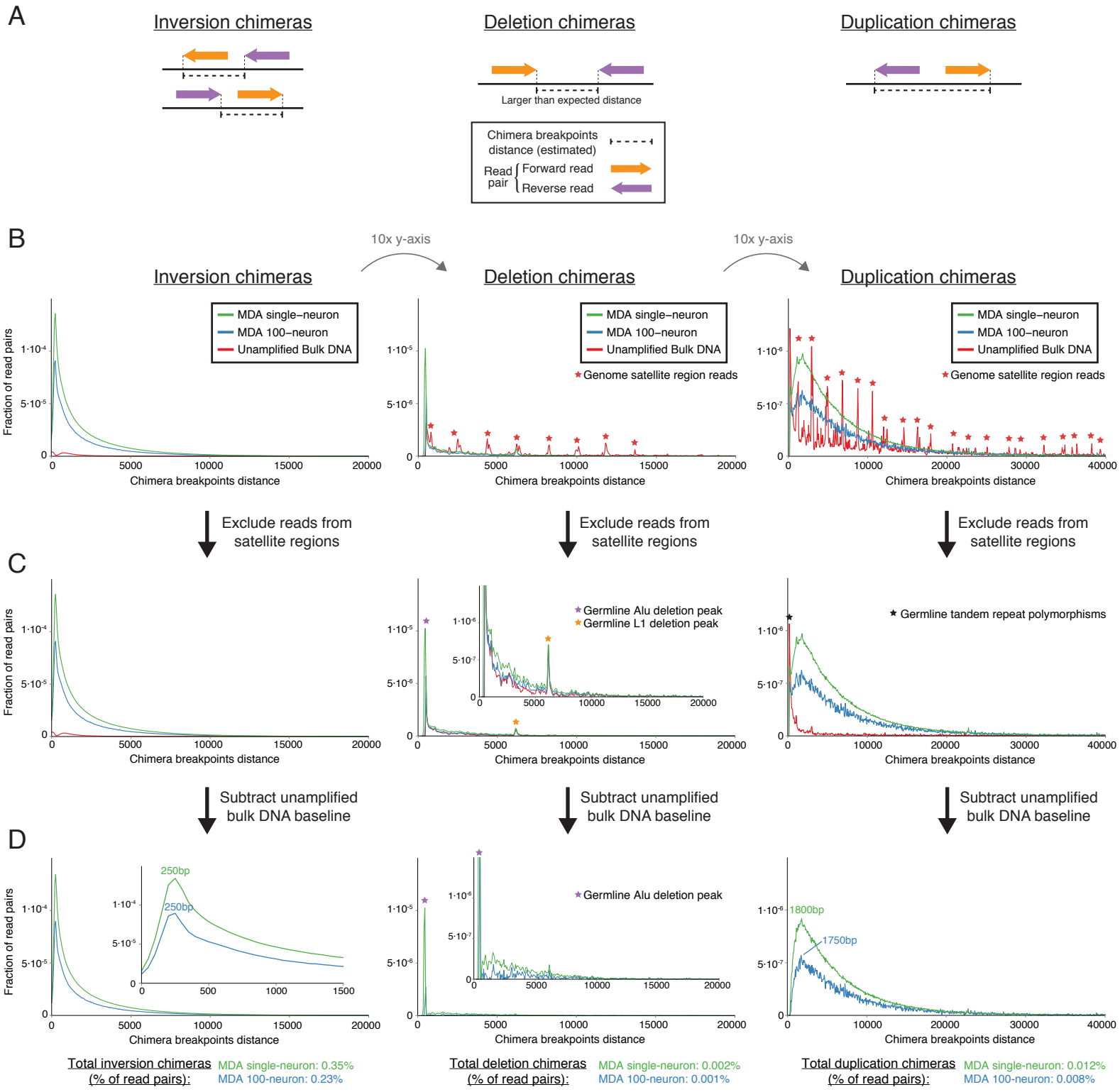

A
Inversion chimeras
Deletion chimeras
Larger than expected distance
Duplication chimeras
Chimera breakpoints distance (estimated)
Read pair
Forward read
Reverse read
B
10x y-axis
10x y-axis
Inversion chimeras
Deletion chimeras
Duplication chimeras
MDA single−neuron
MDA 100−neuron
Unamplified Bulk DNA
Genome satellite region reads
Fraction of read pairs
Chimera breakpoints distance
Exclude reads from satellite regions
C
Germline Alu deletion peak
Germline L1 deletion peak
Germline tandem repeat polymorphisms
Subtract unamplified bulk DNA baseline
D
250bp
250bp
1800bp
1750bp
Total inversion chimeras (% of read pairs):
MDA single-neuron: 0.35%
MDA 100-neuron: 0.23%
Total deletion chimeras (% of read pairs):
MDA single-neuron: 0.002%
MDA 100-neuron: 0.001%
Total duplication chimeras (% of read pairs):
MDA single-neuron: 0.012%
MDA 100-neuron: 0.008%

Figure S9

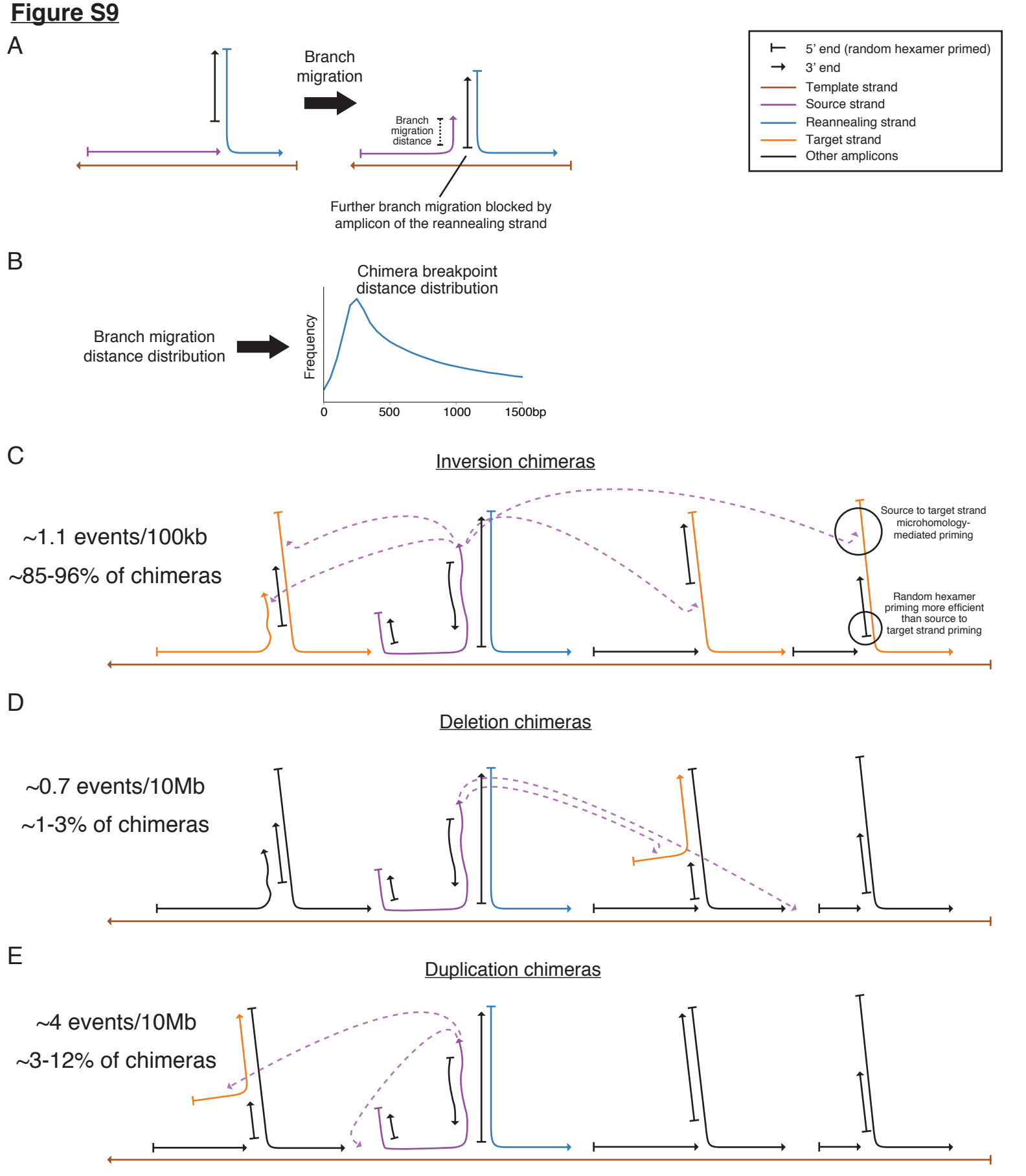

A
Branch migration
Branch migration distance
Further branch migration blocked by amplicon of the reannealing strand
5' end (random hexamer primed)
3' end
Template strand
Source strand
Reannealing strand
Target strand
Other amplicons
B
Branch migration distance distribution
Chimera breakpoint distance distribution
Frequency
0
500
1000
1500bp
C
Inversion chimeras
~1.1 events/100kb
~85-96% of chimeras
Source to target strand microhomology-mediated priming
Random hexamer priming more efficient than source to target strand priming
D
Deletion chimeras
~0.7 events/10Mb
~1-3% of chimeras
E
Duplication chimeras
~4 events/10Mb
~3-12% of chimeras

Figure S10

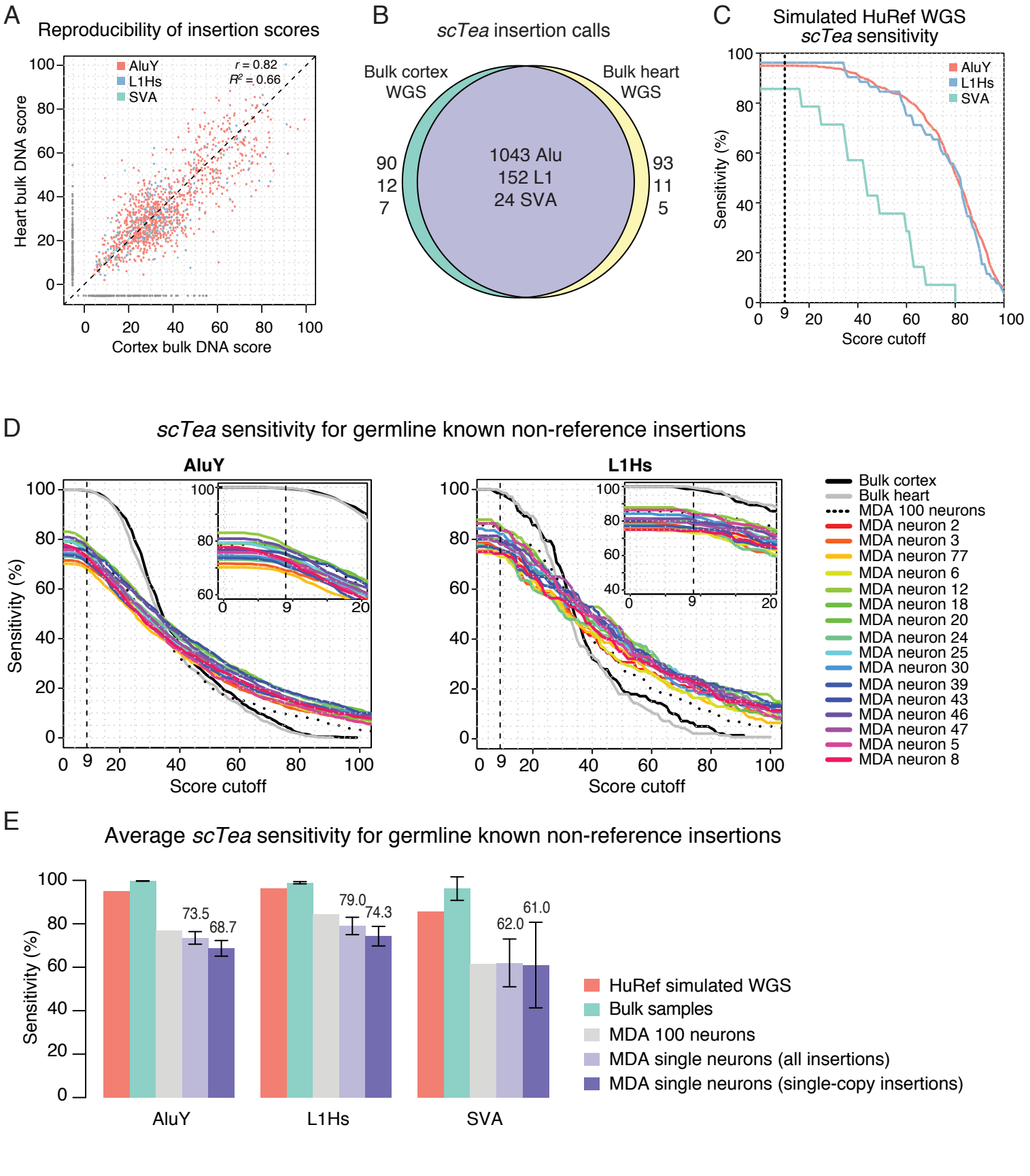

A
Reproducibility of insertion scores
AluY
L1Hs
SVA
r = 0.82
R² = 0.66
Heart bulk DNA score
Cortex bulk DNA score
B
scTea insertion calls
Bulk cortex WGS
Bulk heart WGS
90
12
7
1043 Alu
152 L1
24 SVA
93
11
5
C
Simulated HuRef WGS
scTea sensitivity
AluY
L1Hs
SVA
Sensitivity (%)
Score cutoff
D
scTea sensitivity for germline known non-reference insertions
AluY
L1Hs
Sensitivity (%)
Score cutoff
Bulk cortex
Bulk heart
MDA 100 neurons
MDA neuron 2
MDA neuron 3
MDA neuron 77
MDA neuron 6
MDA neuron 12
MDA neuron 18
MDA neuron 20
MDA neuron 24
MDA neuron 25
MDA neuron 30
MDA neuron 39
MDA neuron 43
MDA neuron 46
MDA neuron 47
MDA neuron 5
MDA neuron 8
E
Average scTea sensitivity for germline known non-reference insertions
Sensitivity (%)
73.5
68.7
79.0
74.3
62.0
61.0
AluY
L1Hs
SVA
HuRef simulated WGS
Bulk samples
MDA 100 neurons
MDA single neurons (all insertions)
MDA single neurons (single-copy insertions)

Figure S11

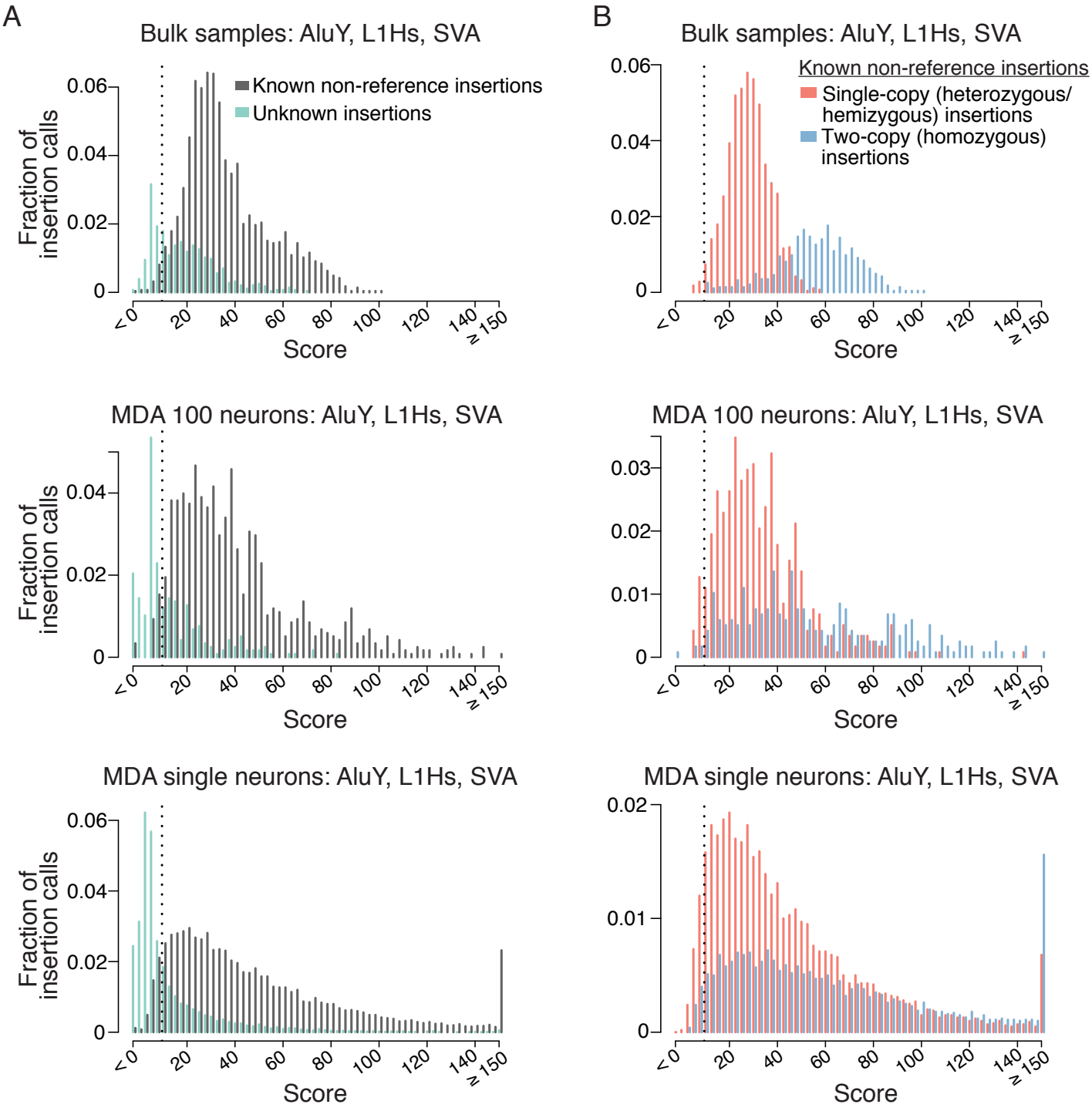
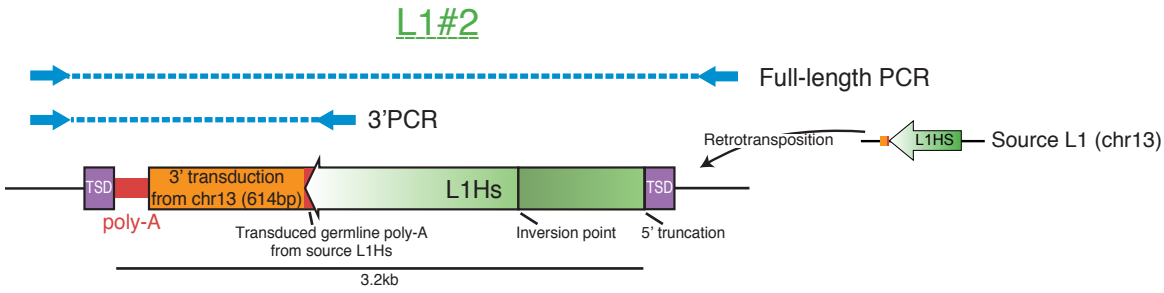
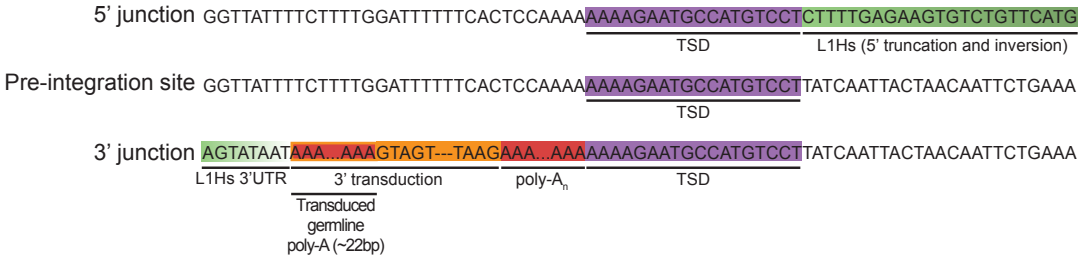

A
Bulk samples: AluY, L1Hs, SVA
Known non-reference insertions
Unknown insertions
Fraction of insertion calls
0.06
0.04
0.02
0
<0
20
40
60
80
100
120
140
≥150
Score
B
Bulk samples: AluY, L1Hs, SVA
Known non-reference insertions
Single-copy (heterozygous/ hemizygous) insertions
Two-copy (homozygous) insertions
MDA 100 neurons: AluY, L1Hs, SVA
0.03
0.01
MDA single neurons: AluY, L1Hs, SVA

Figure S12

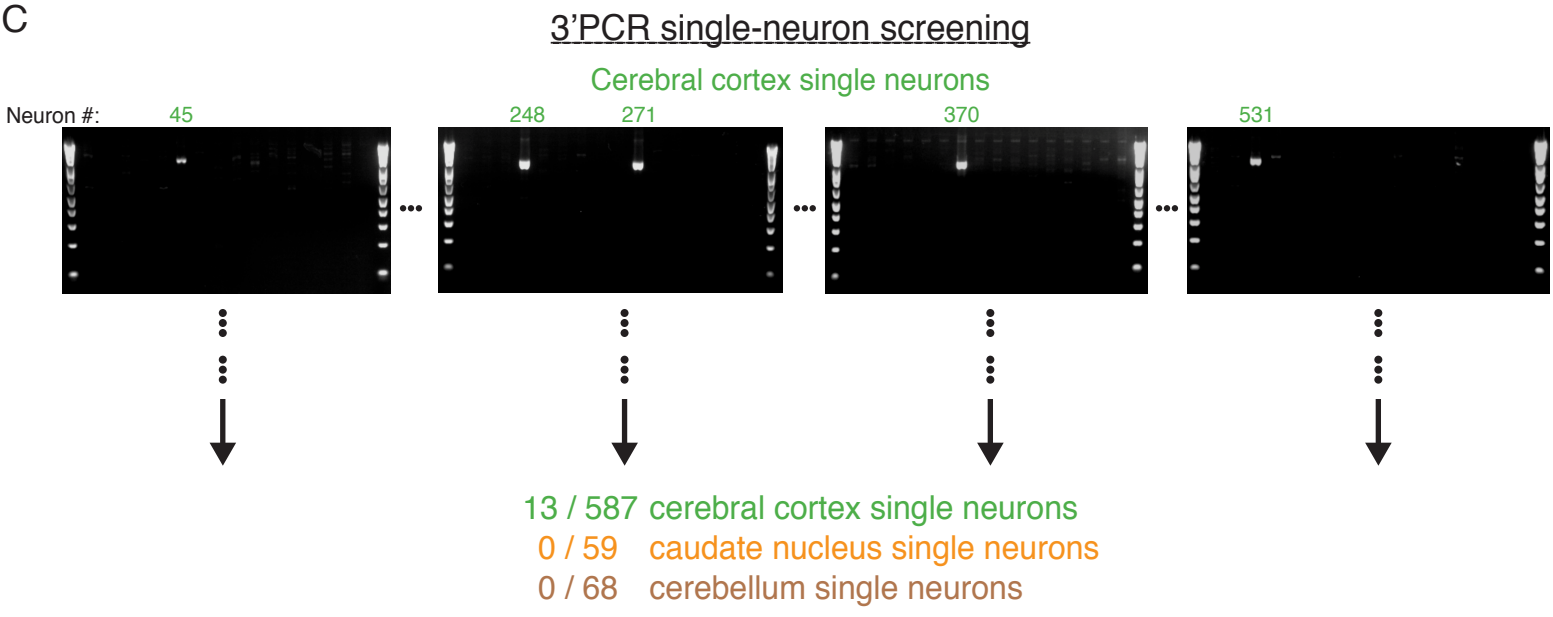
A



B



C



D

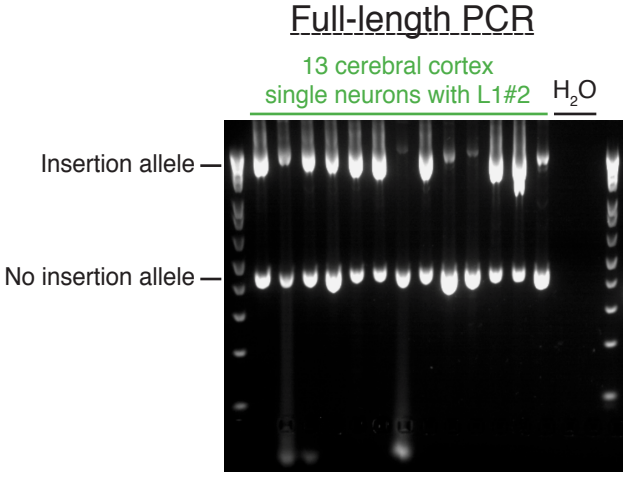
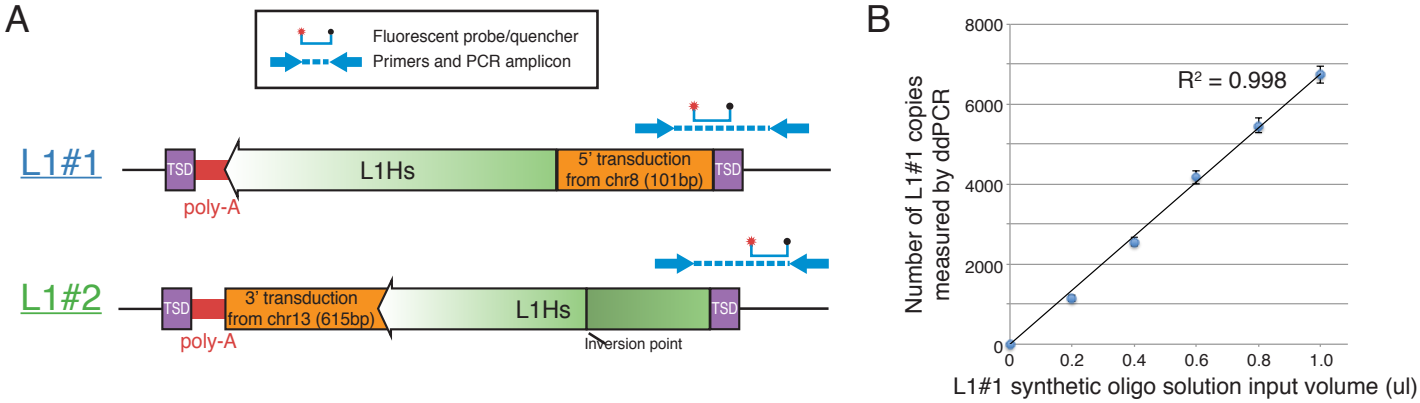


Figure S13


A
Fluorescent probe/quencher
Primers and PCR amplicon
L1#1
TSD
L1Hs
5' transduction
from chr8 (101bp)
TSD
poly-A
L1#2
TSD
3' transduction
from chr13 (615bp)
L1Hs
TSD
poly-A
Inversion point
B
Number of L1#1 copies
measured by ddPCR
8000
6000
4000
2000
0
R² = 0.998
0
0.2
0.4
0.6
0.8
1.0
L1#1 synthetic oligo solution input volume (ul)

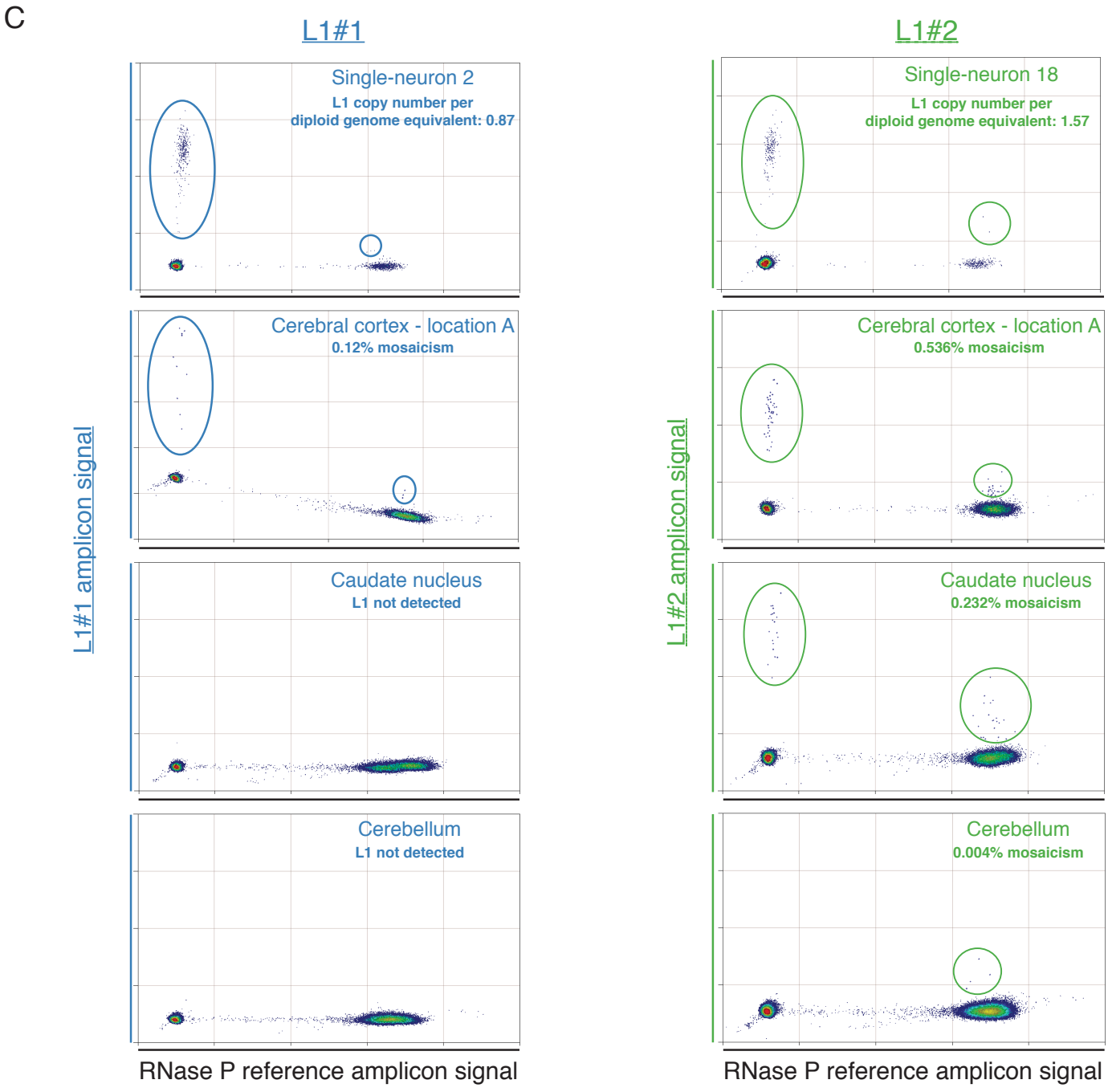
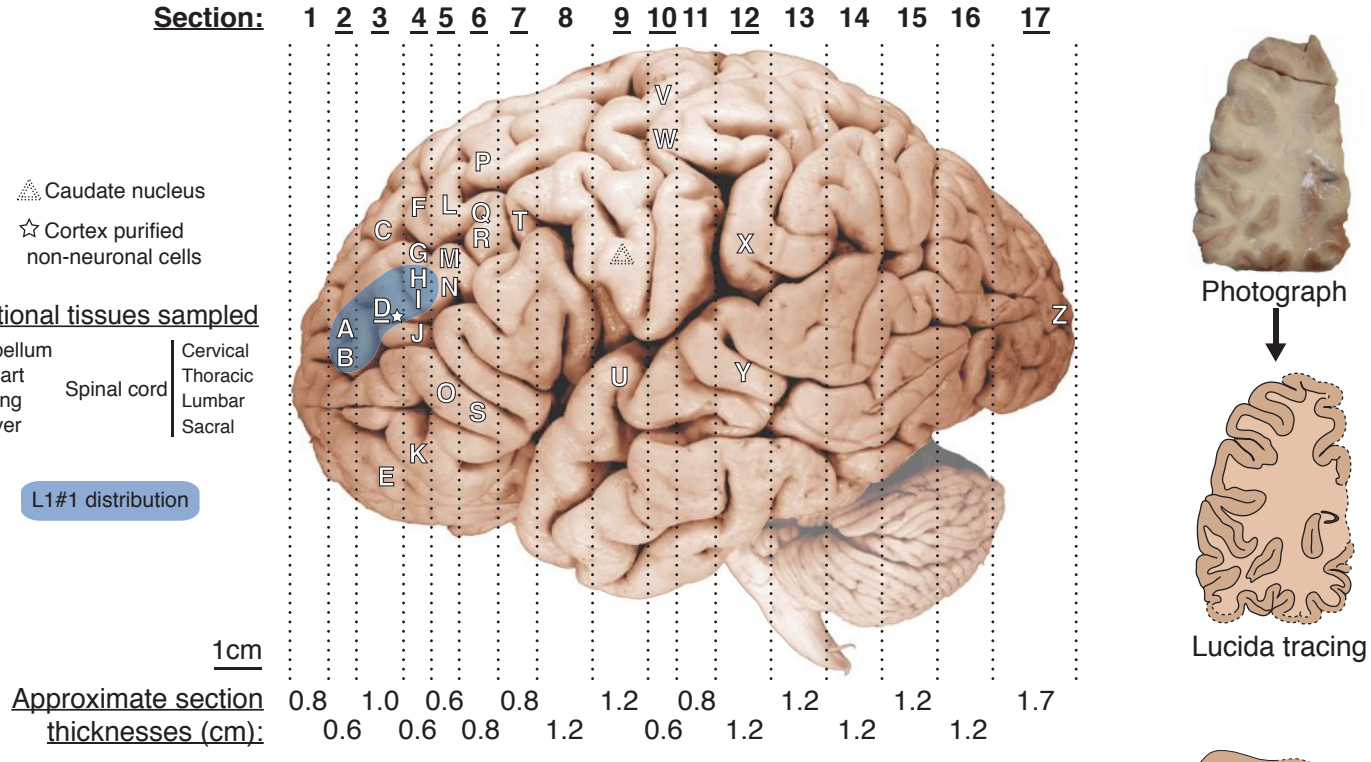

C
L1#1
L1#2
Single-neuron 2
L1 copy number per
diploid genome equivalent: 0.87
Single-neuron 18
L1 copy number per
diploid genome equivalent: 1.57
Cerebral cortex - location A
0.12% mosaicism
Cerebral cortex - location A
0.536% mosaicism
Caudate nucleus
L1 not detected
Caudate nucleus
0.232% mosaicism
Cerebellum
L1 not detected
Cerebellum
0.004% mosaicism
L1#1 amplicon signal
L1#2 amplicon signal
RNase P reference amplicon signal
RNase P reference amplicon signal

Figure S14

A

Brain sections and locations sampled from individual 1465



B

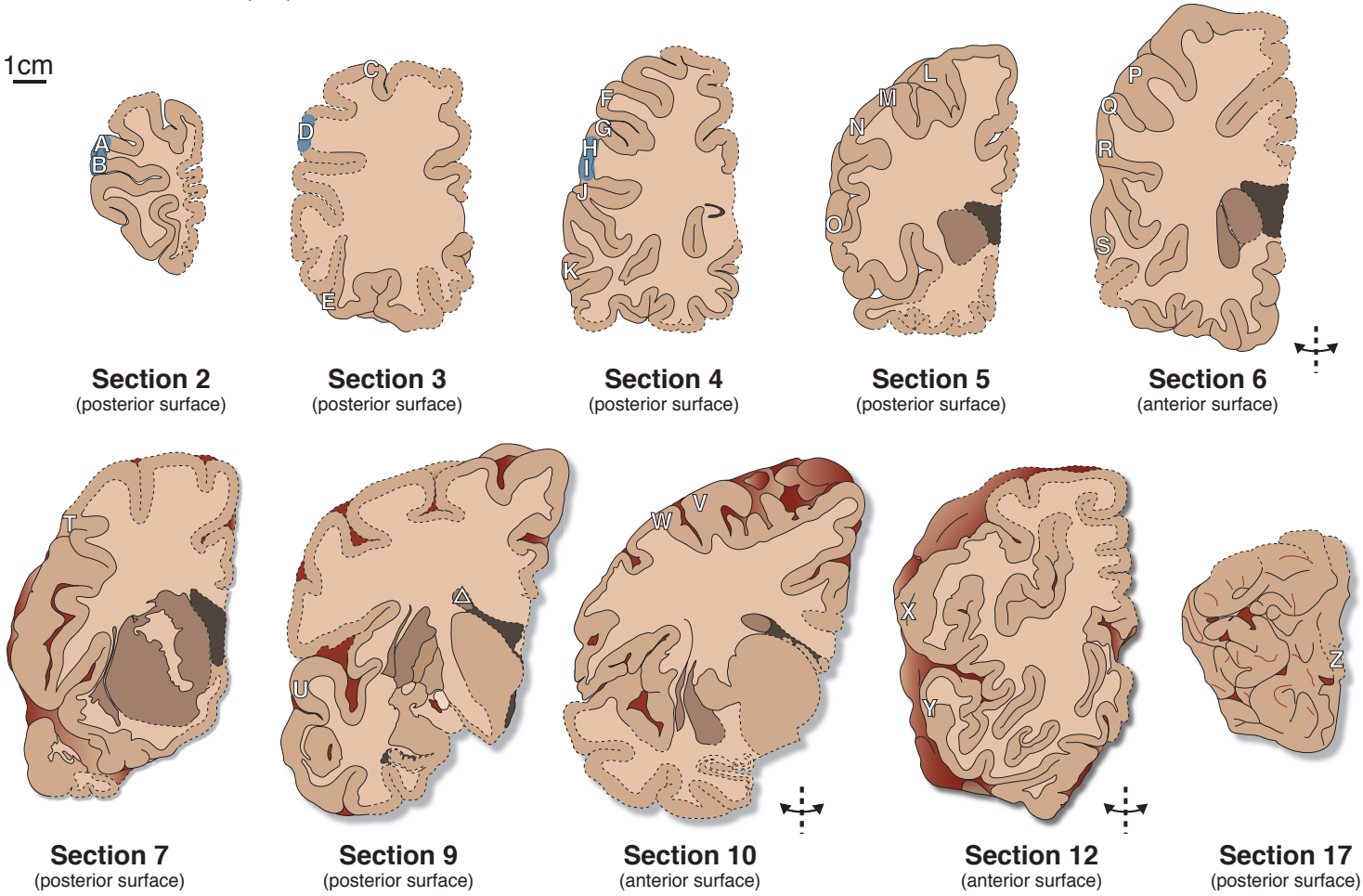


Figure S15

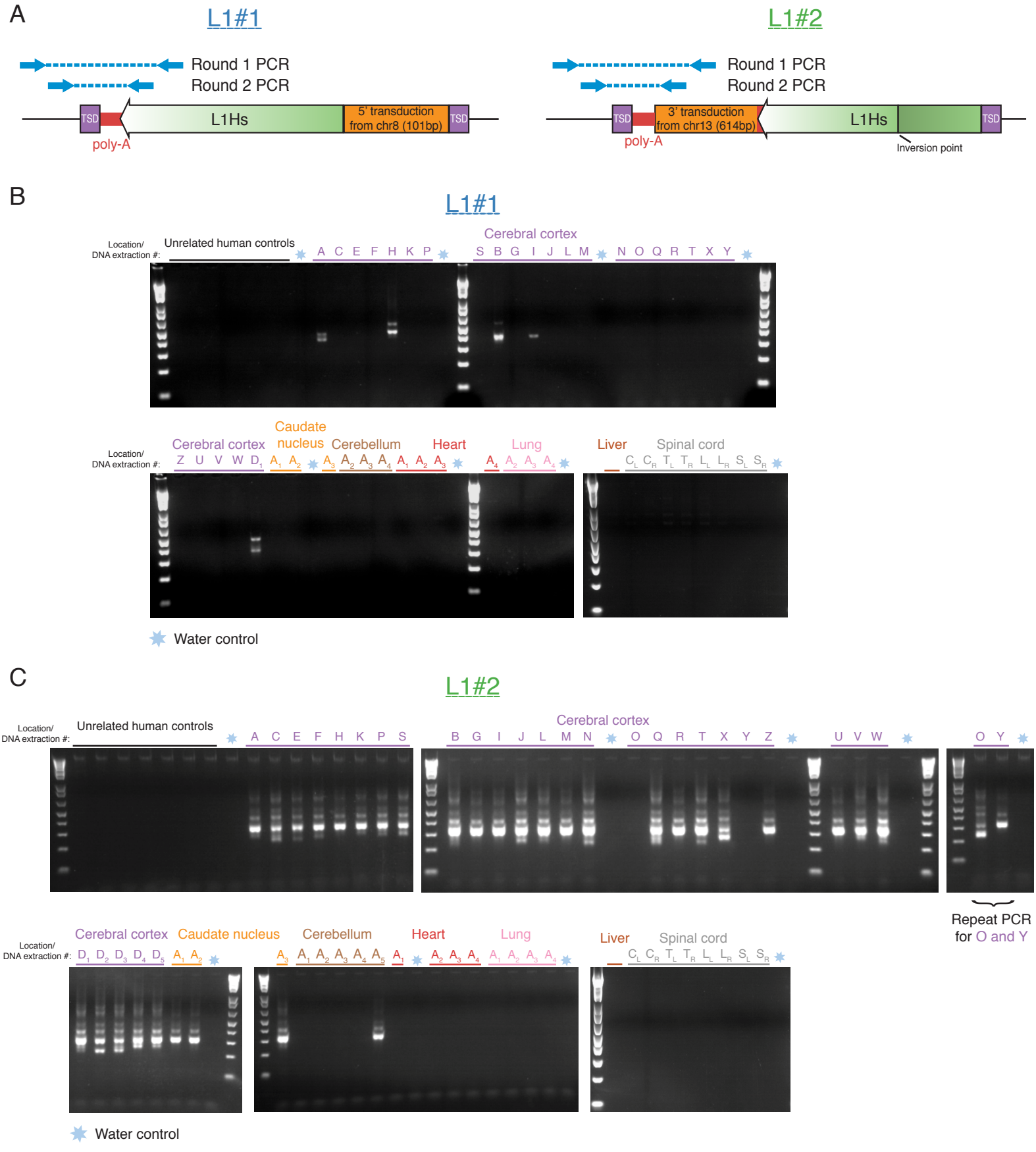

A
L1#1
Round 1 PCR
Round 2 PCR
TSD
L1Hs
5' transduction from chr8 (101bp)
TSD
poly-A
L1#2
Round 1 PCR
Round 2 PCR
TSD
3' transduction from chr13 (614bp)
L1Hs
TSD
poly-A
Inversion point
B
L1#1
Location/ DNA extraction #:
Unrelated human controls
Cerebral cortex
A C E F H K P
S B G I J L M
N O Q R T X Y
Location/ DNA extraction #:
Cerebral cortex
Caudate nucleus
Cerebellum
Heart
Lung
Liver
Spinal cord
Z U V W D1 A1 A2 A3 A2 A3 A4 A1 A2 A3 A4 A2 A3 A4 CL CR TL TR LL LR SL SR
Water control
C
L1#2
Location/ DNA extraction #:
Unrelated human controls
Cerebral cortex
A C E F H K P S
B G I J L M N
O Q R T X Y Z
U V W
O Y
Repeat PCR for O and Y
Location/ DNA extraction #:
Cerebral cortex
Caudate nucleus
Cerebellum
Heart
Lung
Liver
Spinal cord
D1 D2 D3 D4 D5 A1 A2
A3 A1 A2 A3 A4 A5 A1
A2 A3 A4
A1 A2 A3 A4
CL CR TL TR LL LR SL SR
Water control

Figure S16

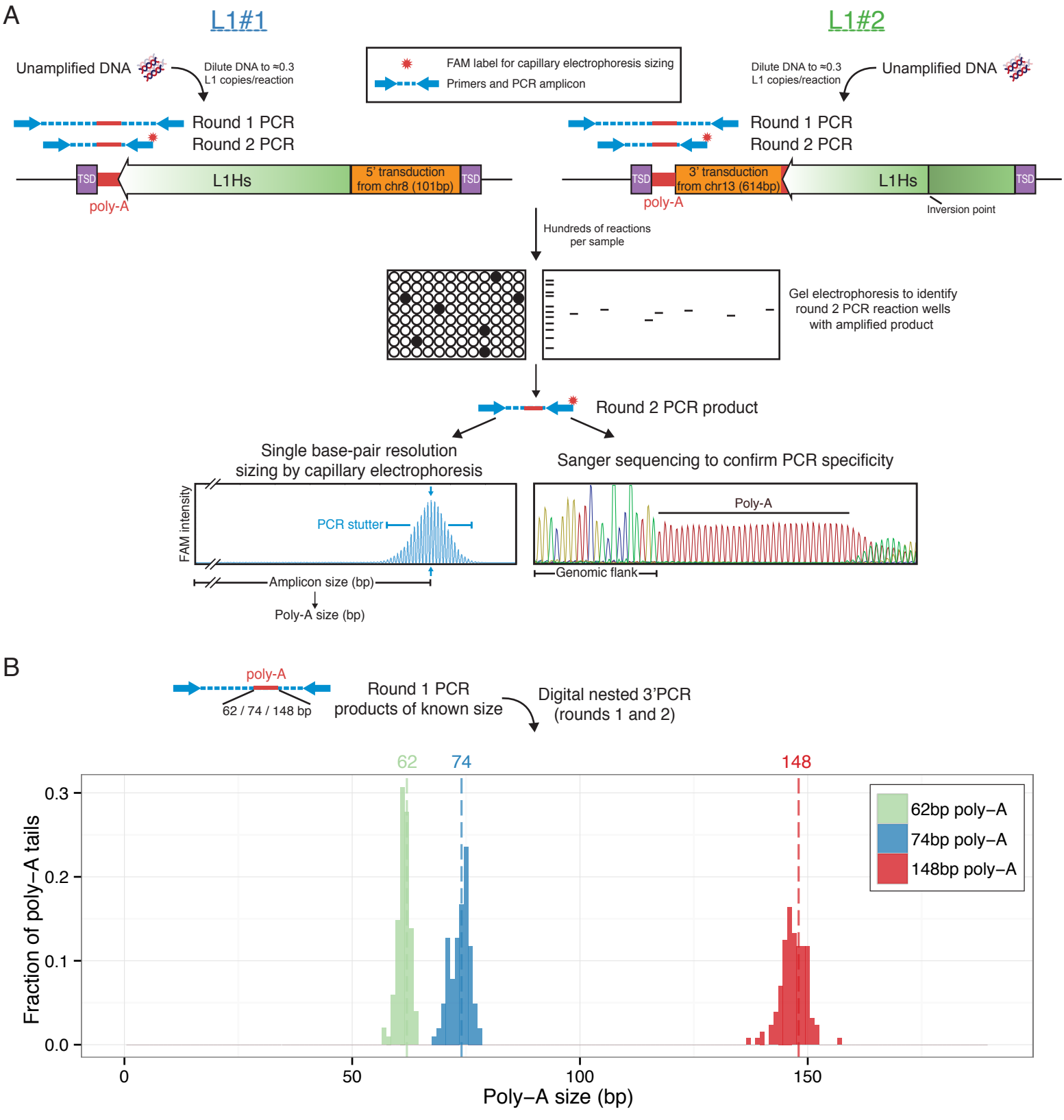
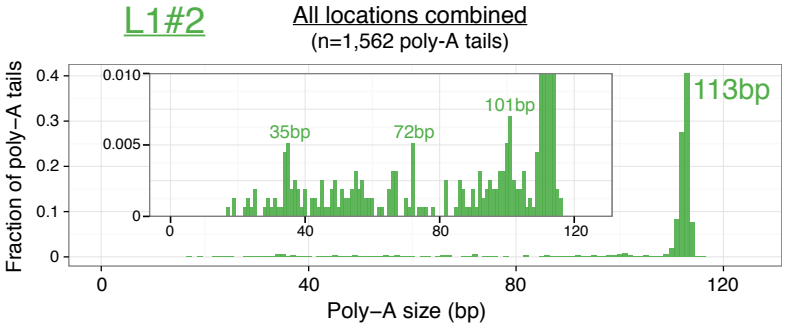

A
L1#1
L1#2
Unamplified DNA
Dilute DNA to ≈0.3 L1 copies/reaction
FAM label for capillary electrophoresis sizing
Primers and PCR amplicon
Round 1 PCR
Round 2 PCR
TSD
L1Hs
5' transduction from chr8 (101bp)
3' transduction from chr13 (614bp)
poly-A
Inversion point
Hundreds of reactions per sample
Gel electrophoresis to identify round 2 PCR reaction wells with amplified product
Round 2 PCR product
Single base-pair resolution sizing by capillary electrophoresis
Sanger sequencing to confirm PCR specificity
FAM intensity
PCR stutter
Amplicon size (bp)
Poly-A size (bp)
Poly-A
Genomic flank
B
poly-A
62 / 74 / 148 bp
Round 1 PCR products of known size
Digital nested 3'PCR (rounds 1 and 2)
62
74
148
Fraction of poly−A tails
0.0
0.1
0.2
0.3
0
50
100
150
Poly−A size (bp)
62bp poly−A
74bp poly−A
148bp poly−A

Figure S17

A



B

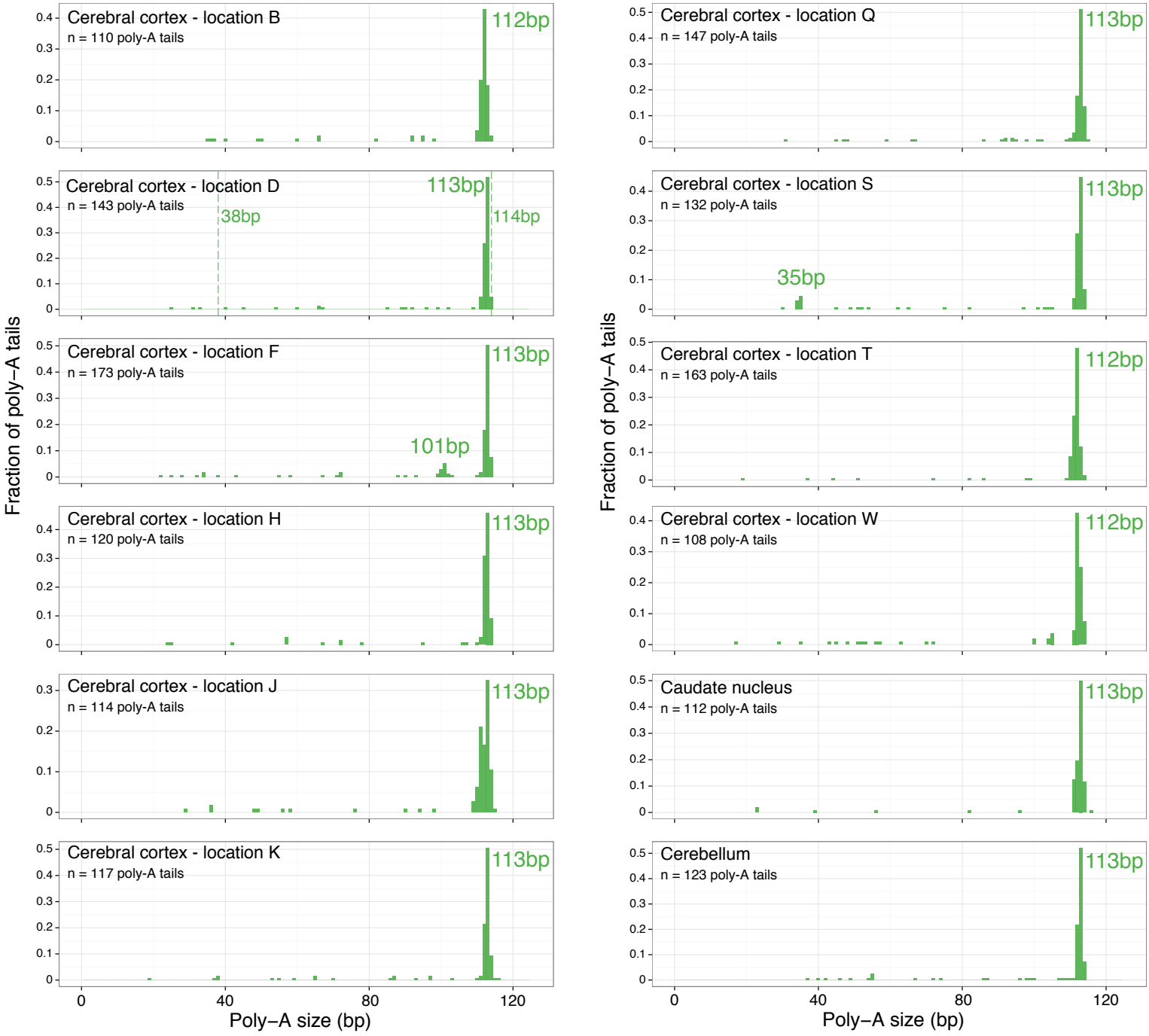


Figure S18

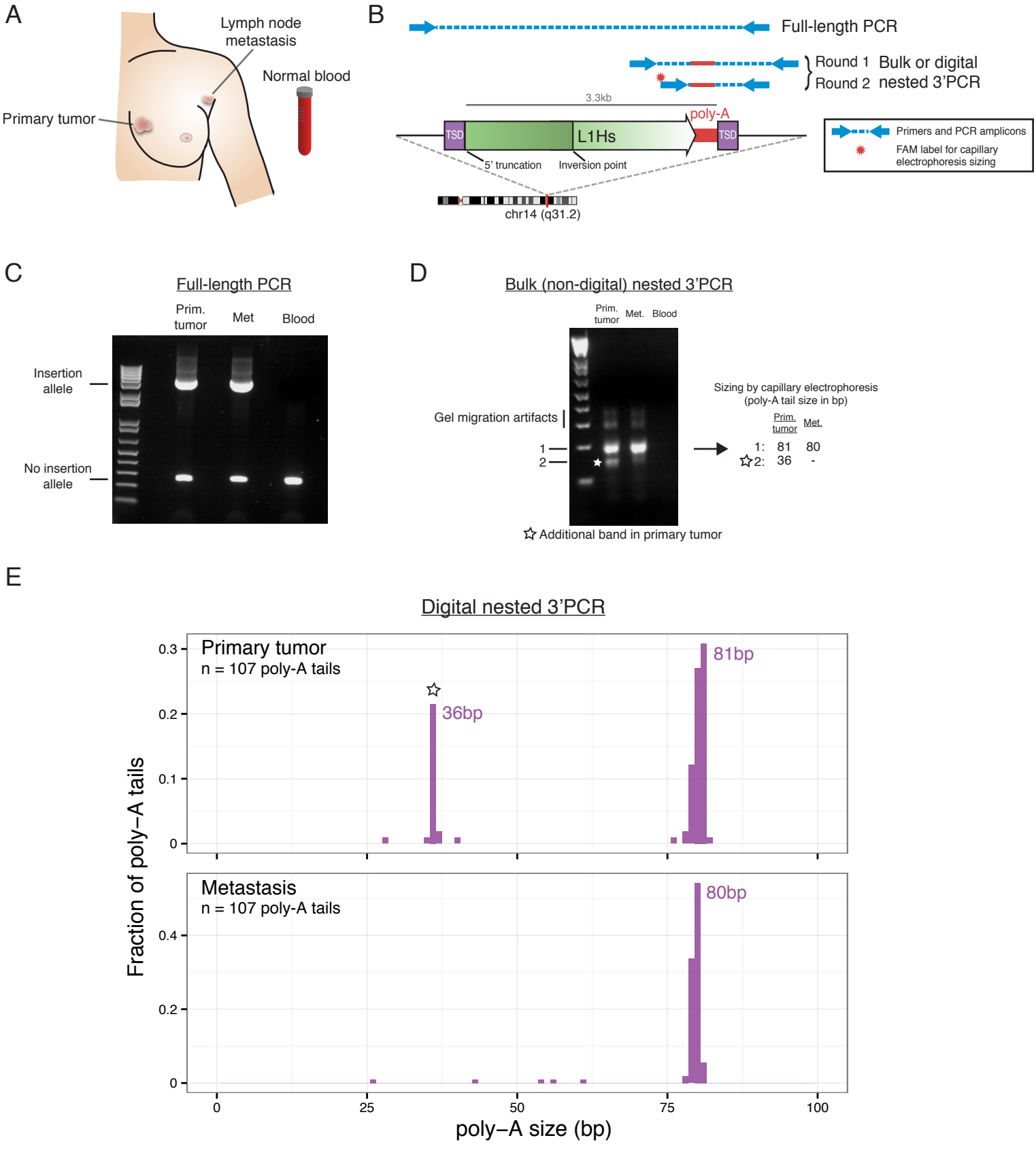
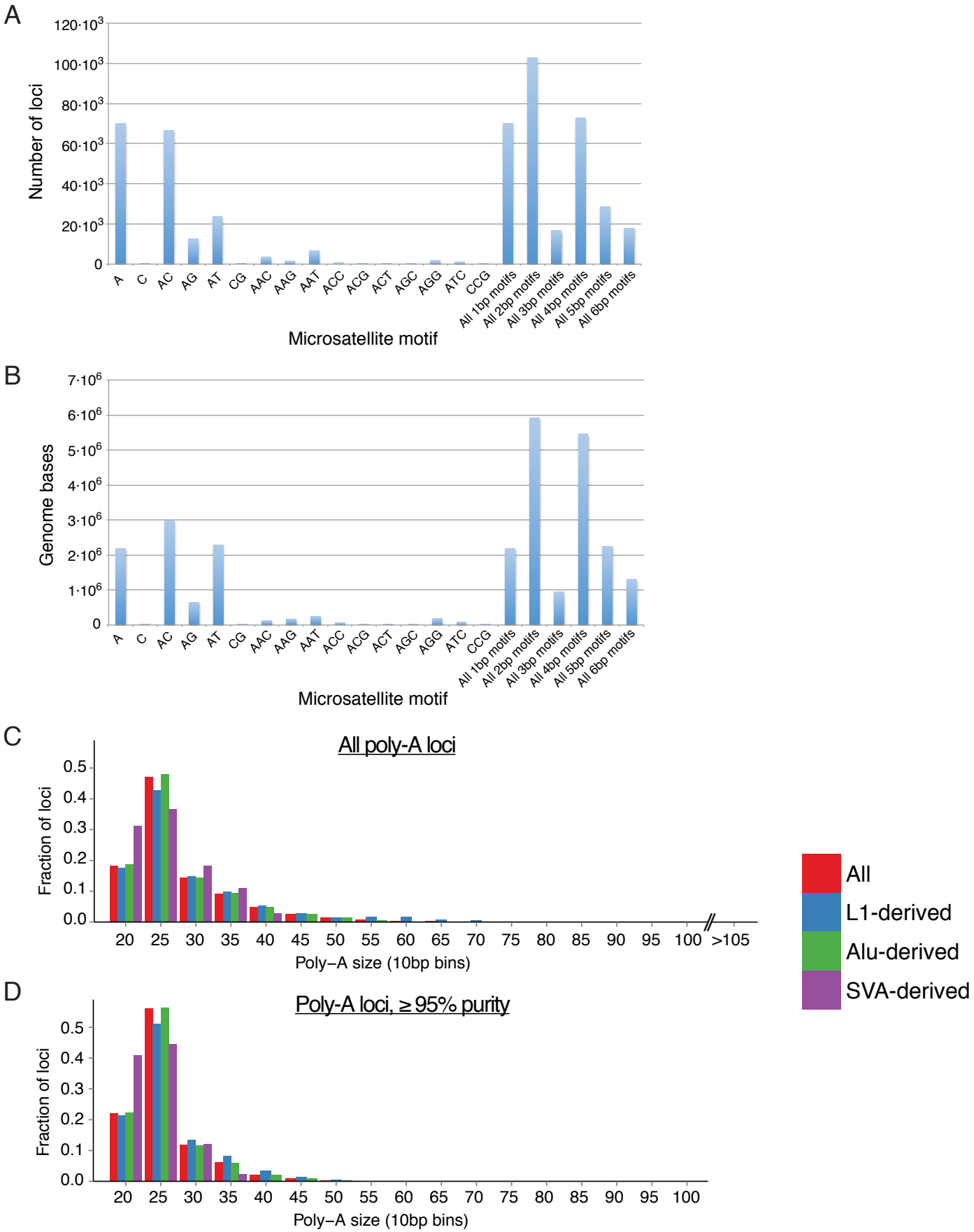

A
Lymph node metastasis
Normal blood
Primary tumor
B
Full-length PCR
Round 1 Bulk or digital
Round 2 nested 3'PCR
3.3kb
poly-A
TSD
L1Hs
TSD
5' truncation
Inversion point
chr14 (q31.2)
Primers and PCR amplicons
FAM label for capillary electrophoresis sizing
C
Full-length PCR
Prim. tumor
Met
Blood
Insertion allele
No insertion allele
D
Bulk (non-digital) nested 3'PCR
Prim. tumor
Met.
Blood
Gel migration artifacts
1
2
Sizing by capillary electrophoresis (poly-A tail size in bp)
Prim. tumor
Met.
1: 81 80
☆2: 36 -
☆ Additional band in primary tumor
E
Digital nested 3'PCR
Primary tumor
n = 107 poly-A tails
36bp
81bp
Metastasis
n = 107 poly-A tails
80bp
Fraction of poly–A tails
0.3
0.2
0.1
0
0.4
0.2
0
0
25
50
75
100
poly–A size (bp)

Figure S19


A
Number of loci
120·10^3
100·10^3
80·10^3
60·10^3
40·10^3
20·10^3
0
A C AC AG AT CG AAC AAG AAT ACC ACG ACT AGC AGG ATC CCG All 1bp motifs All 2bp motifs All 3bp motifs All 4bp motifs All 5bp motifs All 6bp motifs
Microsatellite motif
B
Genome bases
7·10^6
6·10^6
5·10^6
4·10^6
3·10^6
2·10^6
1·10^6
0
A C AC AG AT CG AAC AAG AAT ACC ACG ACT AGC AGG ATC CCG All 1bp motifs All 2bp motifs All 3bp motifs All 4bp motifs All 5bp motifs All 6bp motifs
Microsatellite motif
C
All poly-A loci
Fraction of loci
0.5
0.4
0.3
0.2
0.1
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 >105
Poly–A size (10bp bins)
All
L1-derived
Alu-derived
SVA-derived
D
Poly-A loci, ≥ 95% purity
Fraction of loci
0.5
0.4
0.3
0.2
0.1
0.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Poly–A size (10bp bins)