

Supplementary Information

Cas9 targeted enrichment of mobile elements using nanopore sequencing

Torrin L. McDonald^{*1}, Weichen Zhou^{*2}, Christopher Castro², Camille Mumm¹, Jessica A. Switzenberg², Ryan E. Mills^{1,2,†}, Alan P. Boyle^{1,2,†}

¹Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

²Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

^{*}These authors contributed equally

[†]To whom correspondence should be addressed. Email: remills@umich.edu (R.E.M.) and apboyle@umich.edu (A.P.B.)

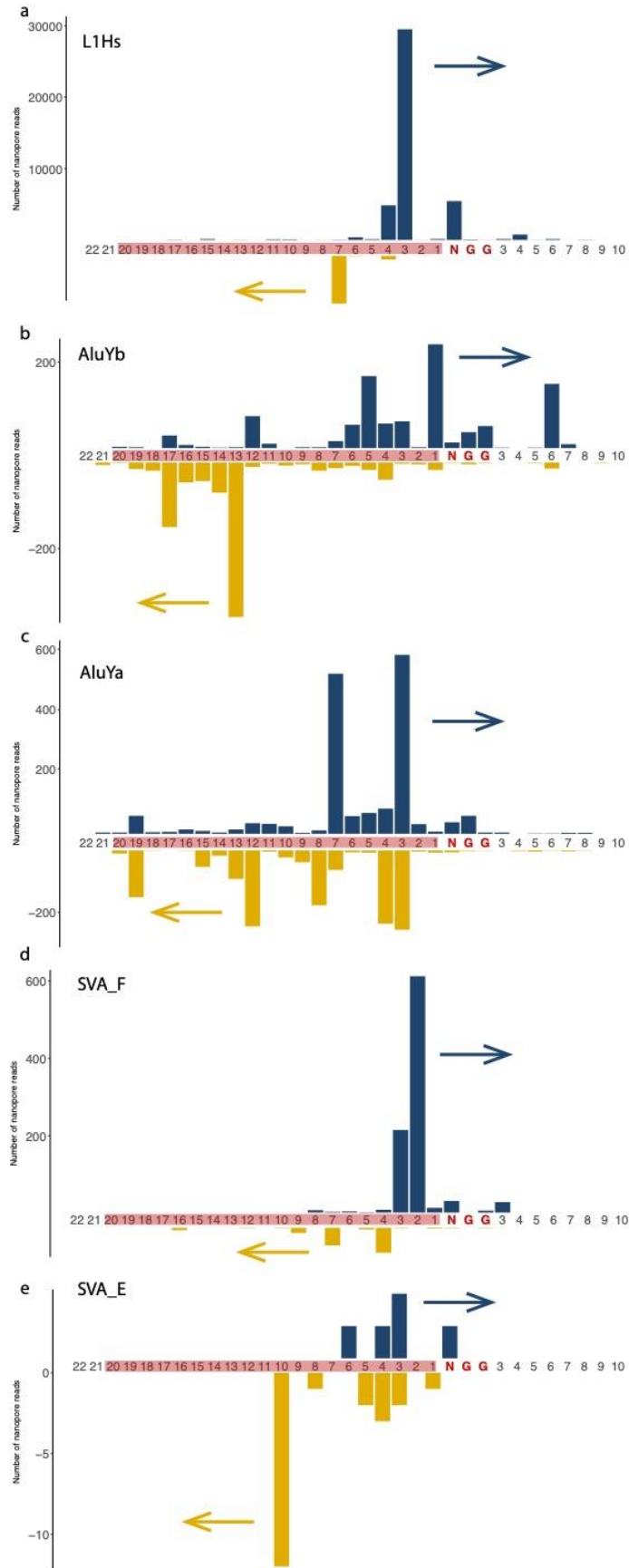
AluYb8 for AluYb subfamily



AluYa5 for AluYa subfamily

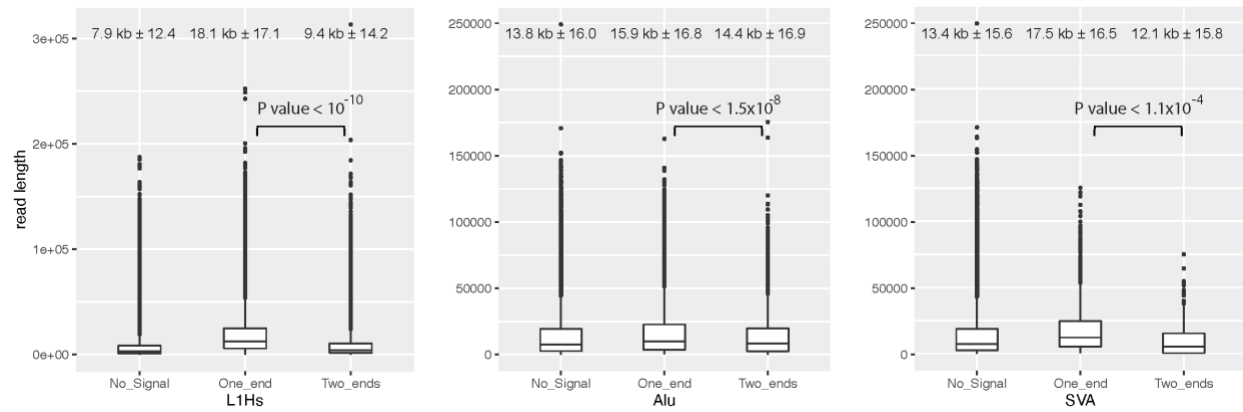
Supplementary Fig.1: Guide RNA design for *AluYb* and *AluYa* element. Consensus sequences are aligned and showed for *AluYb*, *AluYa*, and *AluY* in the middle. The candidate guide RNAs are distributed based on different tiers (see **Methods**). The final list is highlighted by yellow.

Key:
CGCCACTGCAGTCCGCGAGTCCGG 1863+148
Guide RNA sequence + pam site red highlighted
+ number counted in the reference genome for exact sequence
+ number counted in the reference genome for other three nucleotides in NGG site
* means redundant guide RNA sequence with different NGG/CCN site on the other end



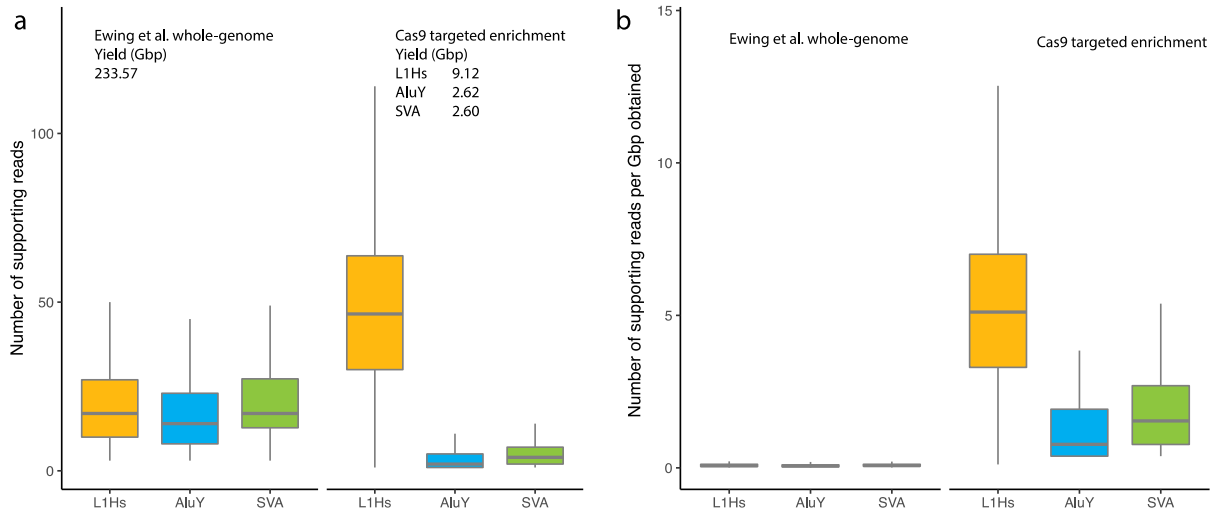
Supplementary Fig.3: Distributions of guide RNA cleavage-site for five MEI subfamilies

a, Cleavage-site distribution of L1Hs guide RNA. X-axis shows the position where the read ends or begins with the number indicating the distance from the 'N' of the PAM site (NGG). The PAM site (NGG) was colored red and guide RNA bases were highlighted by red background. Y-axis is the number of nanopore reads counted. The upper blue bar represents the reads with forward strand sequencing outward the 3' end of guide RNA and the lower yellow bar represents the reads with reverse strand sequencing outward the 5' end of guide RNA. **b**, Cleavage-site distribution of *AluYb* guide RNA. **c**, Cleavage-site distribution of *AluYa* guide RNA. **d**, Cleavage-site distribution of SVA_F guide RNA. **e**, Cut-site distribution of SVA_E guide RNA.



Supplementary Fig.4: Read length distributions for MEI categories (L1Hs, *AluY*, and SVA).

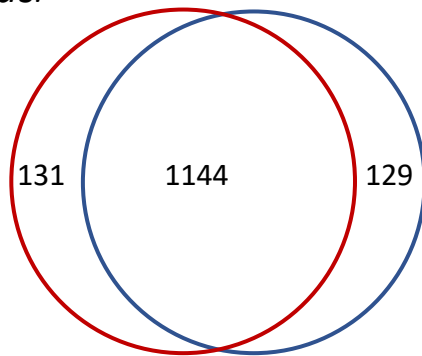
All reads were identified into three categories: reads with on-target reference MEI signals, ones with on-target non-reference MEI signals, and ones with no signals (off-target). The read length (mean $\pm$ standard deviation) are shown above the boxplots. For L1Hs experiments, $n=442,674$ reads have no signal, $n=244,643$ reads have signal on one end, and $n=44,059$ reads have signals on both ends. For *AluY* experiments, $n=84,970$ reads have no signal, $n=30,778$ reads have signal on one end, and $n=4,460$ reads have signals on both ends. For SVA experiments, $n=122,585$ reads have no signal, $n=4,083$ reads have signal on one end, and $n=148$ reads have signals on both ends. And the P-values (student's T-test, two-tailed), $<10^{-10}$, $<1.5 \times 10^{-8}$, and $<1.1 \times 10^{-4}$, are shown between one-end reads and two-end reads in the L1Hs, *AluY*, and SVA experiments, respectively. Error bars range from $Q1-1.5IQR$ to $Q3+1.5IQR$ (IQR, interquartile range).



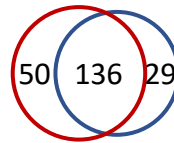
Supplementary Fig.5: Number of supporting reads for three categories of non-reference MEI from the Cas9 targeted nanopore sequencing and the whole-genome nanopore sequencing by Ewing et al. 2020.

a, Number of supporting reads for non-reference L1Hs, *AluY*, and SVA in the whole-genome nanopore sequencing (from five PromethION flow cells) and the Cas9 targeted nanopore sequencing (L1Hs from 11 MinION/Flongle flow cells, *AluY* from four from 4 MinION/Flongle flow cells, and SVA from 4 MinION/Flongle flow cells). **b**, Number of supporting reads of non-reference MEIs normalized by the total yield base pairs from flow cells in two studies. For whole-genome Ewing et al., there are 20.35 ± 13.26 (mean $\pm$ standard deviation) supporting reads and 0.087 ± 0.057 supporting reads per Gbp obtained for non-reference L1Hs ($n=311$), 17.52 ± 12.01 supporting reads and 0.075 ± 0.051 supporting reads per Gbp obtained for non-reference *AluY* ($n=2359$), and 21.43 ± 11.85 supporting reads and 0.092 ± 0.051 supporting reads per Gbp obtained for non-reference SVA ($n=108$). For Cas9 targeted enrichment experiments, there are 49.24 ± 31.25 supporting reads and 5.41 ± 3.43 supporting reads per Gbp obtained for non-reference L1Hs ($n=194$), 3.59 ± 3.29 supporting reads and 1.38 ± 1.26 supporting reads per Gbp obtained for non-reference *AluY* ($n=606$), and 4.88 ± 3.79 supporting reads and 1.88 ± 1.46 supporting reads per Gbp obtained for non-reference SVA ($n=49$). The error bars of boxplot range from Q1-1.5IQR to Q3+1.5IQR (IQR, interquartile range).

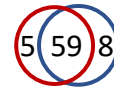
Alus:



L1Hs:

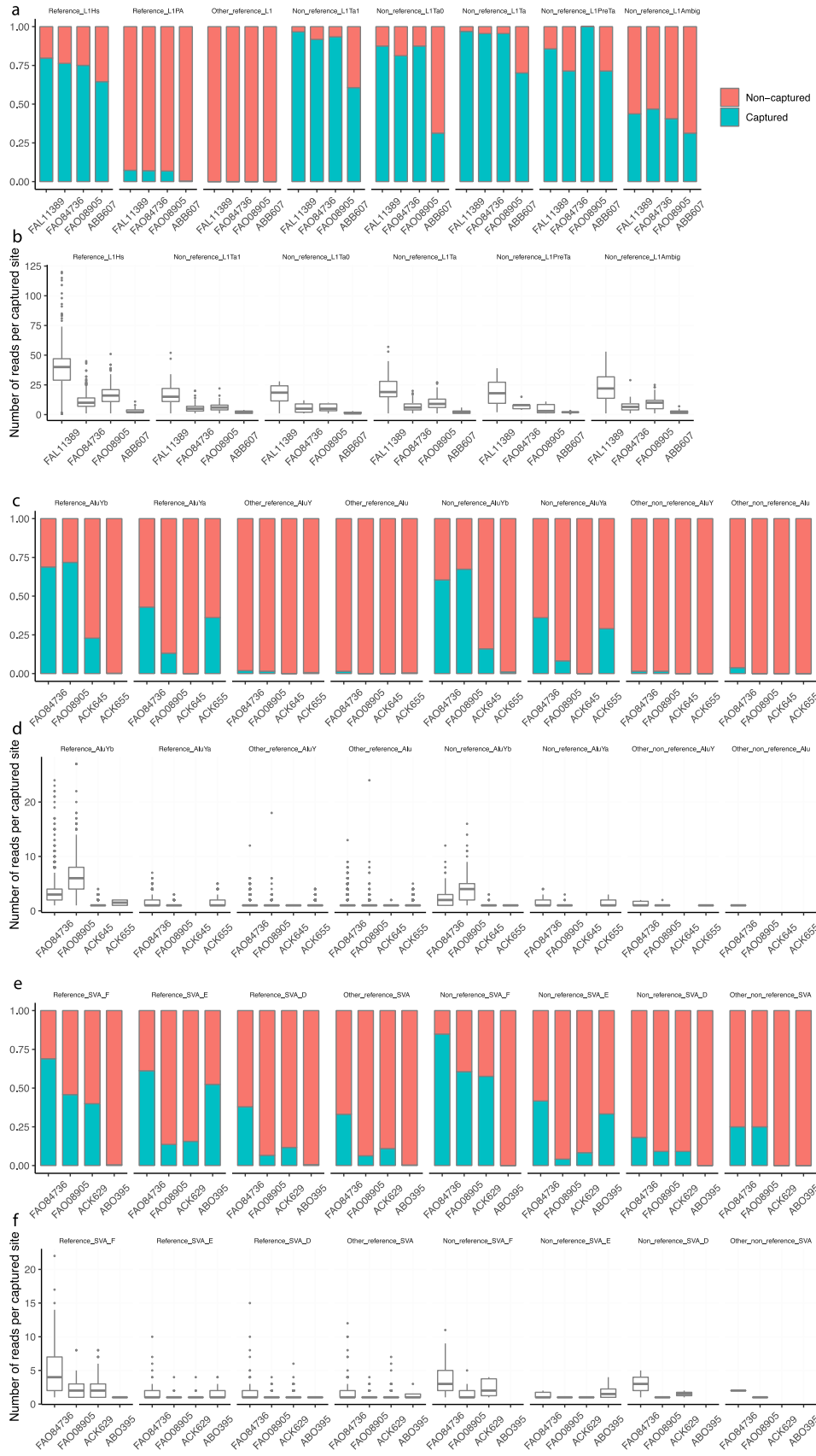


SVA:



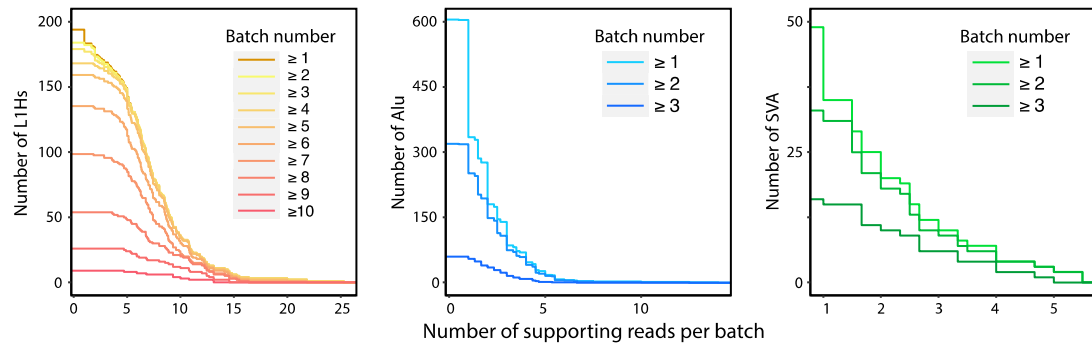
Supplementary Fig.6: Venn diagram of the PALMER callset and the PAV callset for non-reference MEIs in NA12878 genomes.

PALMER callset (red circle) is from PacBio raw sub-reads, and PAV (the Phased Assembly Variant caller, <https://github.com/EichlerLab/pav>, see **Methods**) callset (blue circle) is from PacBio assembly-based pipeline. The circles are depicted by the scale of the numbers showed inside. The union of two sets is generated as 'PacBio-MEI' to be the gold standard set to compare with the calls from nanopore data.



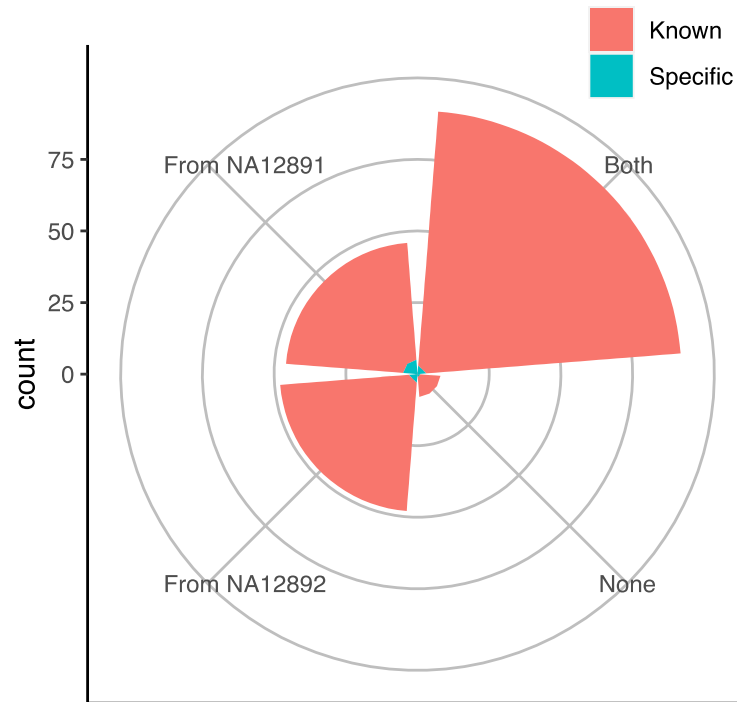
Supplementary Fig.7: Summary of recovered known reference and non-reference MEIs.

a, Known L1Hs in GM12878 recovered by Cas9 targeted enrichment from the individual MinION flow cell (FAL11389), pooled-MEI MinION flow cell (FAO84736), and individual Flongle flow cell (ABB607), displayed in a way of proportion of the upper-bound known reference L1Hs, L1Pa, and other L1 as well as non-reference (non-ref.) subfamilies (L1Ta1, L1Ta0, L1Ta, L1PreTa, and L1Hs with ambiguous subfamilies) of L1Hs from PacBio-MEI set. **b**, The number of supporting reads in each captured L1 in the context of **a**. **c**, Known *AluY* elements in GM12878 recovered by Cas9 enrichment in two pooled MinION flow cells (FAO84736 and FAO08905), one individual *AluYb* Flongle flow cell (ACK645), and one individual *AluYa* Flongle flow cell (ACK655). **d**, The number of supporting reads in each captured *Alu* element in the context of **c**. **e**, Known SVA elements in GM12878 recovered by Cas9 enrichment in two pooled MinION flow cells (FAO84736 and FAO08905), one individual SVA_F Flongle flow cell (ACK629), and one individual SVA_E Flongle flow cell (ACK395). **f**, The number of supporting reads in each captured *Alu* element in the context of **e**. In **b**, **d**, **f**, the numbers of captured MEI subfamily in the boxplots can be found in **Supplementary Data 6** with information of mean and standard deviation; The error bars of boxplot range from $Q1-1.5IQR$ to $Q3+1.5IQR$ (IQR, interquartile range).



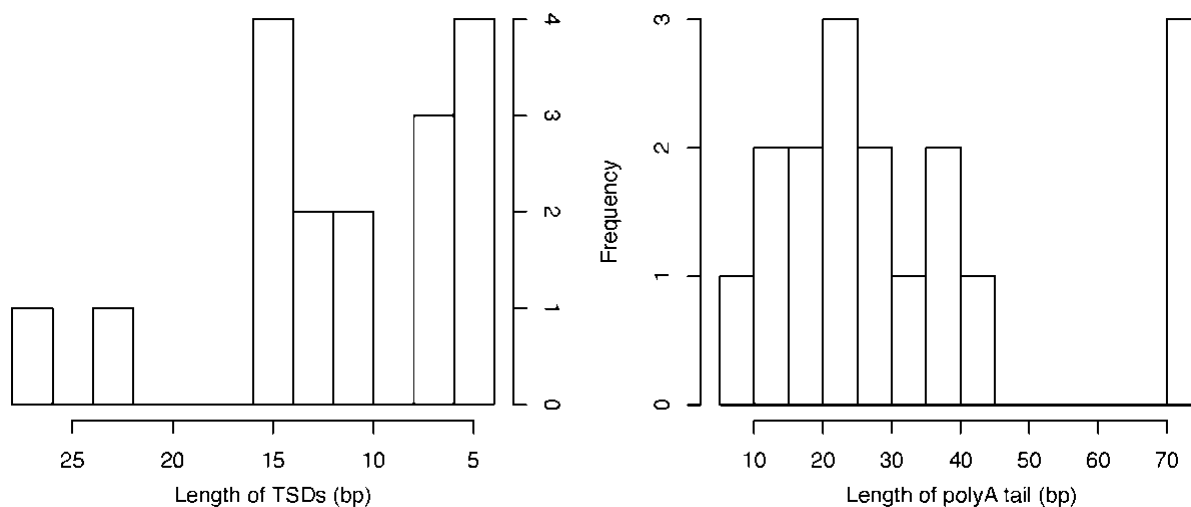
Supplementary Fig.8: MEI distributions in various number of flow cells.

The number of MEIs (L1Hs, yellow; *AluY*, blue; SVA, green) can be captured by nanopore Cas9 enrichment regarding different numbers of flow cells and cutoffs of supporting reads.



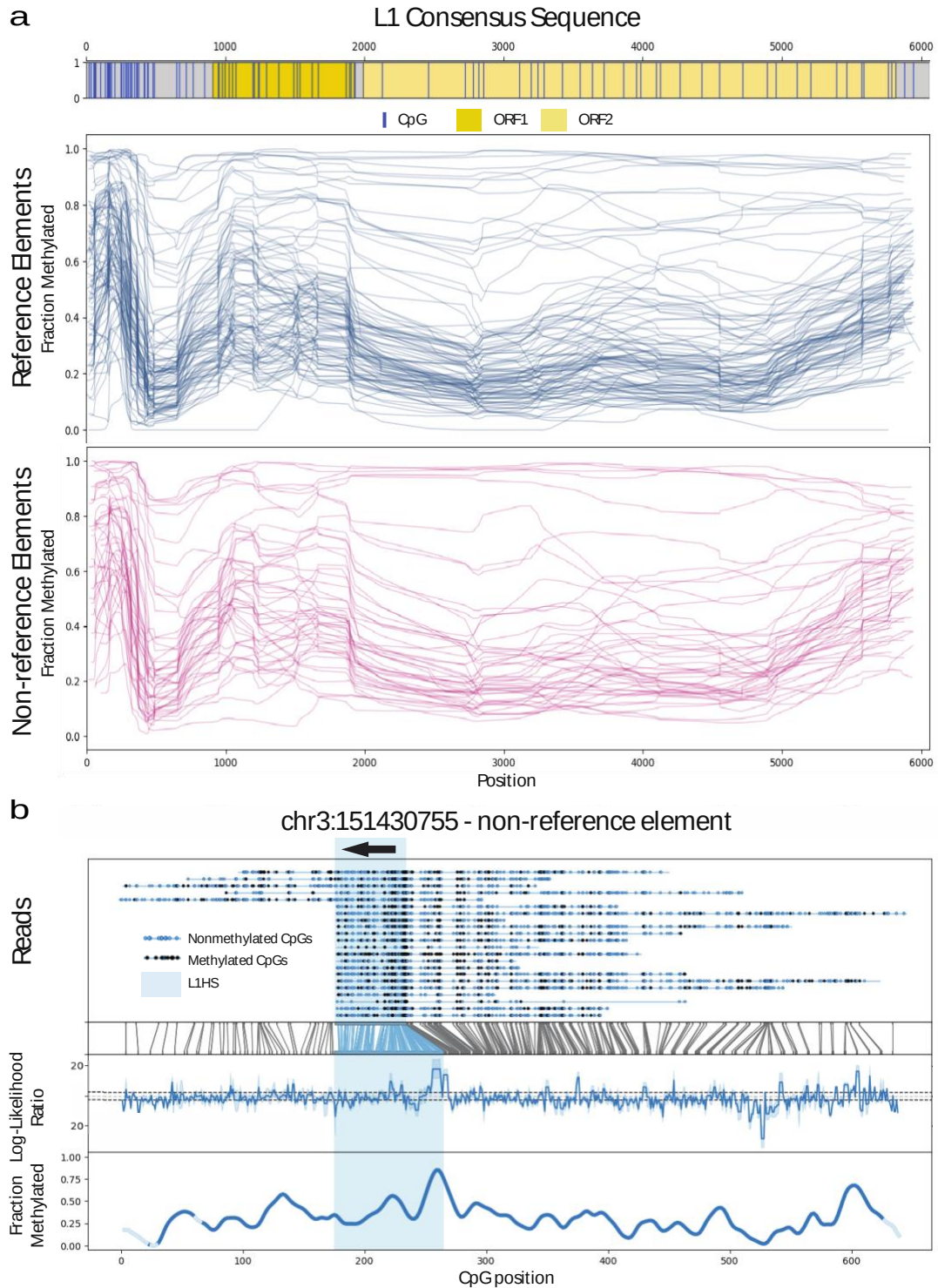
Supplementary Fig.9: Trio transmission of 194 non-reference L1Hs captured by nanopore in GM12878 sample.

The intersections with 'PacBio-MEI' were shown by red and the nanopore-specific non-reference L1Hs that were missed by 'PacBio-MEI' were shown by green.



Supplementary Fig.10: Seventeen nanopore-specific MEIs have hallmarks of retrotransposition consistent with bona fide insertions.

Left, length distribution of TSD motifs. Right, length distribution of poly(A) tails.



Supplementary Fig.11: Examining CpG methylation of captured L1HS reads.

a, L1HS methylation profile over consensus L1 sequence in reference (blue) and non-reference elements (magenta). **b**, An example of methylation profile at chr3:151430755 non-reference L1HS (black arrow 3'→5' orientation). Individual reads with methylation profiles are shown (top) along with their aggregate methylation profile (bottom).