Supplementary Information

Title

Mobile element insertion detection in 89,874 clinical exomes

Authors

Rebecca I. Torene, PhD, MMSc[1]*; Kevin Galens, MS[1]; Shuxi Liu, PhD[1]; Kevin Arvai, MS[1]; Carlos Borroto, MS[1]; Julie Scuffins, MS[1]; Zhancheng Zhang, PhD[1]; Bethany Friedman, MS, CGC[1]; Hana Sroka, MS, CGC[1]; Jennifer Heeley, MD[2]; Erin Beaver, MS, CGC[2]; Lorne Clarke, MD[3]; Sarah Neil, MSc[3]; Jagdeep Walia, MD, FRCPC, FCCMG[4]; Danna Hull, MS, CGC[4]; Jane Juusola, PhD, FACMG[1]; Kyle Retterer, MS[1]

Table of Contents

Supplemental Methods

Details on the SCRAMble method of MEI detection

SCRAMble is a software package written in C and R that can be found on GitHub (https://github.com/GeneDx/scramble). In brief, SCRAMble identifies clusters of soft-clipped reads in a BAM file, builds consensus sequences, aligns to representative L1Ta, AluYa5, and SVA-E sequences, and outputs a tab-delimited file with MEI calls (Figure S3, Table S3). We note that the current implementation of SCRAMble provides MEI calls, but does not distinguish between heterozygous or homozygous genotypes.

Using default settings, SCRAMble extracts reads with at least 10 nucleotides of clipped sequence on one side of the read from an alignment. It then identifies clusters of at least 5 reads that are clipped on the same side at the same genomic position. SCRAMble then builds a majority-rules consensus extended to the length of the longest clipped read. The clipped consensus sequences and their reverse complements are then pairwise-aligned using Smith-Waterman local alignment (as implemented in the pairwiseAlignment function in the R Biostrings package) to a consensus L1, Alu, and SVA sequence. By default, SCRAMble requires the alignment to be at least 70% the length of the clipped consensus sequence, an alignment score ≥50, and a percent identity ≥75%. To increase sensitivity in our clinical cohort, we removed the default 70% alignment length requirement. By allowing for the clipped consensus sequence to contain additional sequence that is not part of an MEI reference sequence, SCRAMble has the potential to detect MEIs that may have short 5' or 3' transductions. Additionally, allowing for a partial-read alignment may detect the 3' end of MEIs after traversing the poly(A) tail. An alignment score of 50 effectively requires the clipped consensus sequence to be >25 nt long. For clipped clusters passing these filters, SCRAMble then scans the region for additional clipped clusters that could represent the second MEI breakpoint. Specifically, it looks for clipped consensus sequences that are ≥75% A (or ≥75% T for minus strand MEIs) within ±200 bp of the originating cluster. If a second cluster is identified, SCRAMble attempts to infer the target site duplication of the MEI. An example output can be found in Table S3.

Sensitivity Analysis

We note that our MEI diagnostic rates are similar to those previously reported[1,2]. Thus, it is likely that our clinical sensitivity is comparable to other methods.

For technical sensitivity, we examined recall on a high confidence MEI call set. The Genome in a Bottle (GIAB) consortium recently made their benchmark structural variant call set available[3]. We, therefore, ran SCRAMble on 2x250 Illumina alignments for the GIAB sample to assess recall of known MEIs. Running SCRAMble for high sensitivity (minimum number of 2 supporting reads and no requirement for alignment length), recalls 1,343 out of 1,467 (91.5%) of PALMER annotated MEIs that were FILTER=PASS in the v0.6 Tier 1 GIAB SV call set (https://github.com/WeichenZhou/PALMER). Using the more conservative default settings, recall is 85.0% (Table S4). We note that the GIAB SV call set was made using a combination of technologies including long-read and optical mapping which may be better able to resolve some more complex MEI events than short-read technologies. We also note that MEIs with long 5’ or 3’ transductions might be under-called by SCRAMble if the clipped consensus cannot span the transduction into the MEI sequence.

We examined the effect of sequencing coverage, allelic fraction, and minimum number of reads on SCRAMble’s sensitivity (Figure S4). We anticipate that some MEIs will be captured less efficiently on a given targeted sequencing platform and will have varying allelic fractions. Using

50,000 iterations of simulated data, we estimated the number of times at each sequencing coverage and at each allelic fraction we would have enough clipped reads to meet a given calling threshold. As expected, lower allelic fractions, lower sequencing coverage, and a higher number of required reads leads to lower sensitivity. Given an allelic fraction of 0.2 and minimum required reads of 5, we estimate a sequencing coverage at the MEI site of at least 38 should allow for >90% sensitivity.

We also examined how the length of the clipped consensus sequence and the level of 5' truncation of an MEI might affect SCRAMble's sensitivity, specifically for Alus. We permuted every possible clipped sequence of the AluYa5 consensus for lengths 25-40 nt. For reads >25 nt, neither clipped read length nor 5' truncation had an effect on SCRAMble's sensitivity (100% of the permuted reads led to an MEI call).

Since the alignment score is a product of alignment length, we wanted to test SCRAMble's performance on 100 bp, instead of 150 bp, reads. We trimmed the reads for the 14 positive samples in Table 1 from 150 bp to 100 bp and applied SCRAMble using a minimum of 2 supporting reads. Of the positive MEIs, 13 of 14 could still be detected. For the sample that could not be detected with 100 bp reads, the original call was made on right-clipped reads where the longest clipped length was 56 bp. By trimming off the last 50 bp, there was not enough clipped sequence remaining for MEI detection.

If higher sensitivity is preferred, we recommend lowering the required minimum number of reads and also lowering the alignment score threshold which will allow for shorter clipped read consensus sequences to be called as MEIs. If desired, additional MEI sequences can be included other than the defaults to examine specific subfamilies or to identify MEIs in other species.

Precision Analysis

When considering precision, we distinguish between clinical precision and technical precision. Clinical precision is based on which MEIs were selected for confirmation by PCR and Sanger because they might have clinical relevance to a case. Technical precision considers MEIs outside clinically relevant regions which have not had PCR confirmation, however, we can estimate precision based on the overlap of MEIs with known polymorphisms and from visual inspection of alignments.

All clinically-reportable MEIs, defined as any *de novo* MEIs in known Mendelian disease genes or inherited MEIs in recessive disorders overlapping the patient phenotype, were sent for Sanger confirmation and 18/30 did confirm, giving a clinical diagnostic positive predictive value (PPV) of 60.0% (95% confidence interval by binomial test of 40.6-77.3%). If the same samples had been run using SCRAMble's default settings (including the minimum 70% read length requirement), then 64.3% (95% confidence interval by binomial test of 44.1-81.4%) of MEIs attempted for confirmation would have confirmed. We note that MEIs sent for confirmation are most often rare or only observed once and will be enriched for false positives, therefore clinical precision does not necessarily represent technical precision of SCRAMble.

For technical precision, we note that 76.4% of the MEI calls in this study overlap with MEI variants from previous studies (were within +- 500 nt of an MEI of the same family in a prior publication)[4–8]. We, therefore, assume that 76.4% is the floor for SCRAMble's precision. Next, we evaluated a random subset of the remaining, novel MEI calls for whether they are likely real. We visually inspected alignments for each of 122 unique MEI variants (representing 20,546 MEI calls across all samples) that were not observed in previous studies. If the variant was present in more than 1 sample, we randomly selected 1 sample for visual inspection. We found that 47/122

(38.5%) of variants specific to this study appeared to be real. While this number seems low, we note that 64/75 (85.3%) of the apparent false positive variants were singletons (i.e., observed in only 1 sample) and the most common apparent false positive variant was only observed in 16/89,874 (0.018%) samples. In fact, when comparing the number of calls between the 47 likely real variants and the 75 apparent false positives, the likely real variants were more likely to be more common in the cohort (two-sided ks-test p=0.043). If we assume that when a variant looks real in 1 sample, that it also likely real when found in other samples, then 20,417/20,546 (99.4%) of the calls represented by the 122 variants appear likely to be real MEIs based on visual inspection of alignments. One of these likely real variants was found in >20k samples and might be inflating the precision estimate. If we remove the variant with the highest number of calls from this analysis, then 359/488 (73.6%) of the remaining calls represented by the 121 variants evaluated appear real. We then estimated an overall technical precision starting with the floor (76.4% of variants which were observed in prior studies) and extrapolating the 73.6-99.4% rate for the remaining 23.6% of MEI calls which were not observed in previous studies. Therefore, the lower estimate for precision would by 76.4 + (23.6*0.736) = 93.8 and the higher estimate would be 76.4 + (23.6*0.994) = 99.9. We, thus, estimate that SCRAMble has a technical precision of 93.8-99.9%.

	MEI variants	MEI calls
Real – with common variant	47/122 (38.5%)	20,417/20,546 (99.4%)
Real – without common variant	46/121 (38.0%)	359/488 (73.6%)

SCRAMble's precision can be modulated by raising the minimum required alignment score from the default of 50 and/or by increasing the percent length alignment required from the default of 70% (a default percent length of alignment of less than 100% was used to allow for some 5' or 3' transduction of sequence that would not align to an MEI consensus). Other features in SCRAMble's output can be used to filter these calls such as the percent identity and the start position in the MEI. Most true Alu insertions are full length while other Alu-mediated structural variants may create clipped reads aligning to the middle of the Alu and be of lower percent identity relative to an AluYa5 consensus.

Comparison to other callers

We ran SCRAMble, Mobster[9], and MELT[6] on the same 1,084 Exome samples with default settings for Mobster and SCRAMble. We also ran all 3 callers on the pathogenic MEI positive samples from Table 1. For 12 of the samples, we had enough DNA to re-sequence the sample to test whether SCRAMble is sensitive to slight differences in capture and coverage. It should be noted that while Mobster is open source, MELT is not open source.

For MELT, we used the following parameters with joint genotyping then FILTER=PASS or rSD in the final VCFs:

```
-exome
-r 150
-c 150
-cov 150
-e 500
```

For Mobster, we used default parameters. Outlier samples (with >100 calls by any 1 method) were removed from analysis leading to a final comparison set of 1,075 samples. Overall, MELT, Mobster, and SCRAMble detected a median number of 5, 10, and 18 MEI calls per person respectively (Table S5). When considering only targeted regions, defined as being within a 250 bp region around the center of a target probe, SCRAMble still detects more MEIs (Figure S5). To estimate whether SCRAMble's additional calls are due to increased sensitivity or an increased false positive rate, we evaluated known MEI polymorphisms from the 1000 Genomes project. In targeted regions, SCRAMble identified more MEI calls for variants previously identified by the 1000 Genomes Project[6] than either MELT or Mobster, even though MELT was the tool used to detect MEIs in the 1000 Genomes cohort (Table S5). We suspect that the low rates of discordant read pairs in our clinical exomes leads to reduced sensitivity for both MELT and Mobster. Of note, SCRAMble provides precise insertion sites whereas 29% of the Mobster calls did not report split read evidence and therefore provide only an approximate insertion site.

We re-sequenced 12 of the 14 MEI positive samples, where sufficient material was still available, from Table 1 to evaluate the reproducibility of SCRAMble calls and to compare recall rates to MELT and Mobster. We were unable to use the original data for these samples because discordant read pairs, which are used by MELT and Mobster, were not archived in all samples. Two of the samples did not have enough DNA for re-sequencing. SCRAMble was able to recall all 12 pathogenic MEIs while Mobster recalled 8 and MELT recalled 6 (Table S5).

ETFB Run of Homozygosity Analysis

Kinship analysis was performed for all samples submitted for a given case as previously described[10]. PLINK was used with the following settings to identify regions of homozygosity (ROH) around the homozygous *ETFB* Alu insertion (Figure S7).

```
--homozyg-snp 10

--homozyg-kb 10

--homozyg-gap 10000
```

```
      --homozyg-window-het 3
```

To estimate the age of the founder event, the boundaries of the ROH surrounding the MEI were used to assess genetic distance based on linkage maps (http://compgen.rutgers.edu/map_interpolator.shtml). The genetic distance was then used to estimate the generational age based on methods described previously[11].

Supplemental Tables

Table S1. Demographic summary of exome referral cohort

	Affected	Unaffected
n	43,118	46,756
Age at testing (years), mean (sd)	16.4 (18.2)	40.1 (11.4)
Female, n (%)	19,525 (45.3)	24,842 (53.1)
Predicted Ancestry:		
African, n (%)	3,620 (8.4)	3,109 (6.6)
American, n (%)	7,144 (16.6)	6,928 (14.8)
Caucasian, n (%)	25,631 (59.4)	28,419 (60.8)
East Asian, n (%)	1,058 (2.5)	1,431 (3.1)
Middle Eastern, n (%)	4,072 (9.4)	4,501 (9.6)
South Asian, n (%)	1,586 (3.7)	2,357 (5.0)
Target coverage, mean (sd)	109.7 (32.6)	110.6 (32.6)

Predicted ancestry was determined by using 20,000 polymorphic SNPs identified in the 1000 Genomes Project and determining nearest distance in PCA space to 6 ancestral groups listed as previously described[10].

Table S2. Filtering strategy for identifying possibly pathogenic MEIs

	Alu	L1	SVA
All Calls	827857	157875	116058
In Targeted Regions	372563	57894	50309
In Coding +/- 5 nt	56127	15367	42523
In Known Disease Gene	398	558	39
Rare in Cohort (<100 samples)	398	318	39
Sent for Confirmation	14	9	1
Confirmed and Reported	11	2	1

Note that 6 additional MEIs were sent for confirmation that were not in known disease genes or turned out later to be common in our cohort. The filtering strategy above is a rough guideline, however, each case is given unique consideration. Several factors are considered before sending an MEI for confirmation, including: relevance to phenotype, observed inheritance pattern, and quality of call.

Table S3. Example SCRAMble output

Insertion	MEI Family	Insertion Direction	Clipped Reads In Cluster	Alignment Score	Alignment Percent Length	Alignment Percent Identity
chr3:29222187	alu	Plus	14	233.8	100	100
chr8:71752636	l1	Minus	12	235.8	100	100
chr10:3083763	sva	Plus	15	186.0	99.2	95.1

Clipped Sequence	Clipped Side	Start In MEI	Stop In MEI	polyA Position
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACT	right	1	118	29222173
CCCTAGTGAGATGAACCCGGTACCTCAGATGGAAATGCAGAAATCACCGTCTTCTGCGTCGCTCACGCTGGGAGCTGTAGACCGGAGCTGTTCCTATTCGGCCATCTTGGCTCCTCCCC	left	3	121	71752649
GAGTGCTCAATGGTGCCCAGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTACAACCTACACCTCCCAGCCGCCTGCCTTGGCCTCCCAAAGTGCCGAGATTGCAGCCTCTGCCCGGCCGCCAC	right	327	446	3083749

polyA Seq	polyA Supporting Reads	TSD	TSD length
AAAAAAAAAAAAAAAAAAAA	6	AAGAACACAGAACC	14
TTTTTTTTTTTTTTTTTTTT	3	TCAAGACACTTTT	13
AAAATAAATTAAAAAAAAAAAAAAAAAAA	7	AAAAGAAAAATGGT	14

SCRAMble creates a tab delimited file as output with MEI calls. Features of the MEI call such as alignments score, number of clipped reads, and alignment percent length are available for post hoc filtering if desired.

Table S4. SCRAMble recall of MEIs on the Genome In A Bottle sample

	Alu	L1	SVA
SCRAMble default (conservative)	1,066/1,237 (86.2%)	141/157 (89.8%)	40/73 (54.8%)
SCRAMble high sensitivity	1,134/1,237 (91.7%)	151/157 (96.2%)	58/73 (79.5%)
SCRAMble high sensitivity / limit to GIAB benchmark regions	1,084/1,173 (92.4%)	145/150 (96.7%)	55/70 (78.6%)

SCRAMble high sensitivity is run with a minimum of 2 clipped reads (instead of the default 5) and removing the minimum 70% alignment read length requirement.

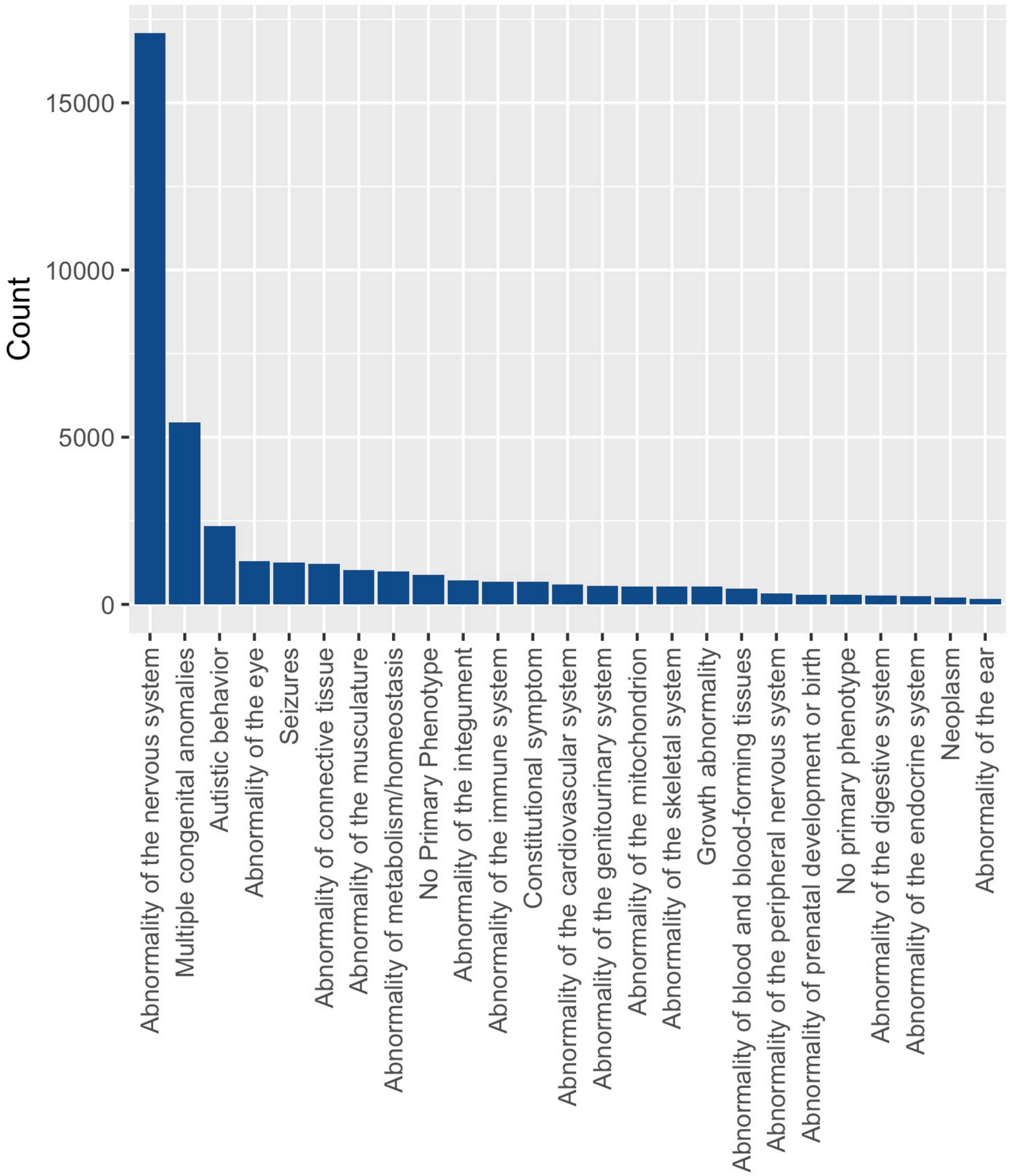
Table S5. Comparison of MEI callers

	SCRAMble	MELT	Mobster
Median number of Calls Per Sample (N=1,075 samples)	18	5	10
Median Number of Calls Per Sample in Targeted Regions (N=1,075 samples)	7	2	5
Recall of published 1000 Genomes MEI polymorphisms (N=1,075 samples)	5,576	402	3,827
Recall on pathogenic MEIs identified in this study (N=12 samples*)	12 (100%)	6 (50%)	8 (66.7%)
Median runtime on a single exome (N=12 samples)	29 min	37 min	23 min

*12 of the 14 MEI positive samples from Table 1 were re-sequenced for this analysis. The remaining 2 samples did not have enough DNA.

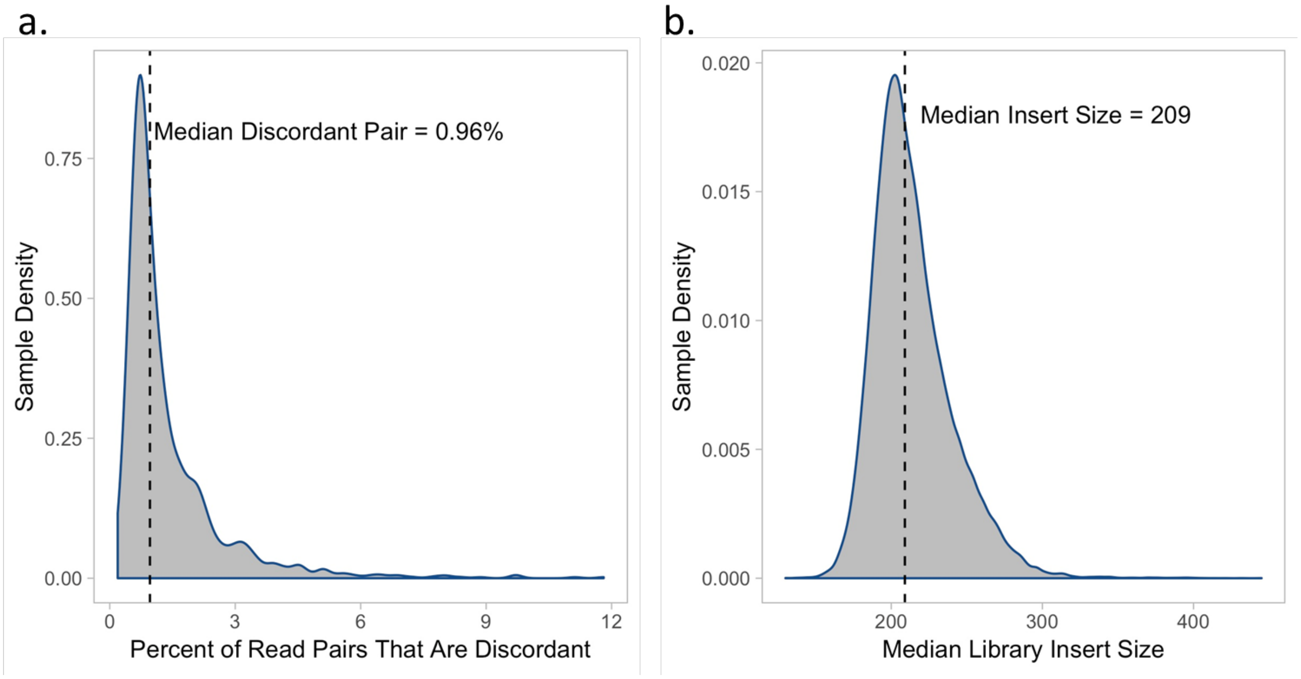
Supplemental Figures

Figure S1. Top 25 Primary Phenotypes in Cohort



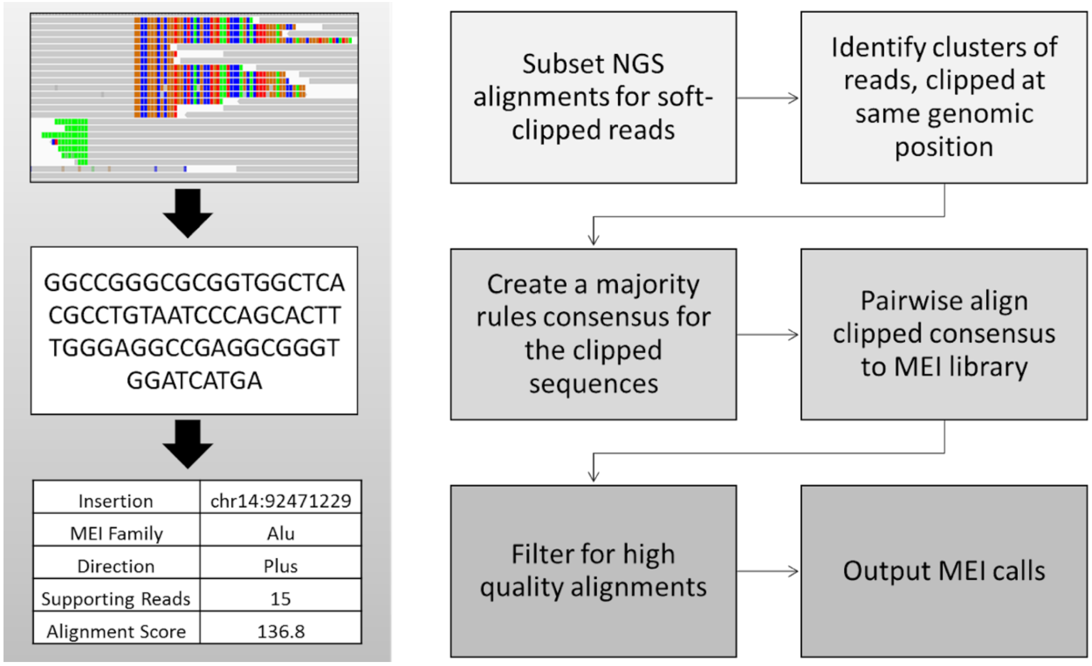
Bar plot of the 25 most common primary phenotypes in the cohort.

Figure S2. Features of our exome sequencing library that affect MEI detection



The relatively low discordant read pair rate and small insert sizes of our clinical exomes may help explain why 1) SCRAMble identifies more MEIs per person than read pair based methods, and, 2) why MELT found more MEIs per person in the cohort described in Gardner et al.[1], than in this cohort.

Figure S3. SCRAMble workflow

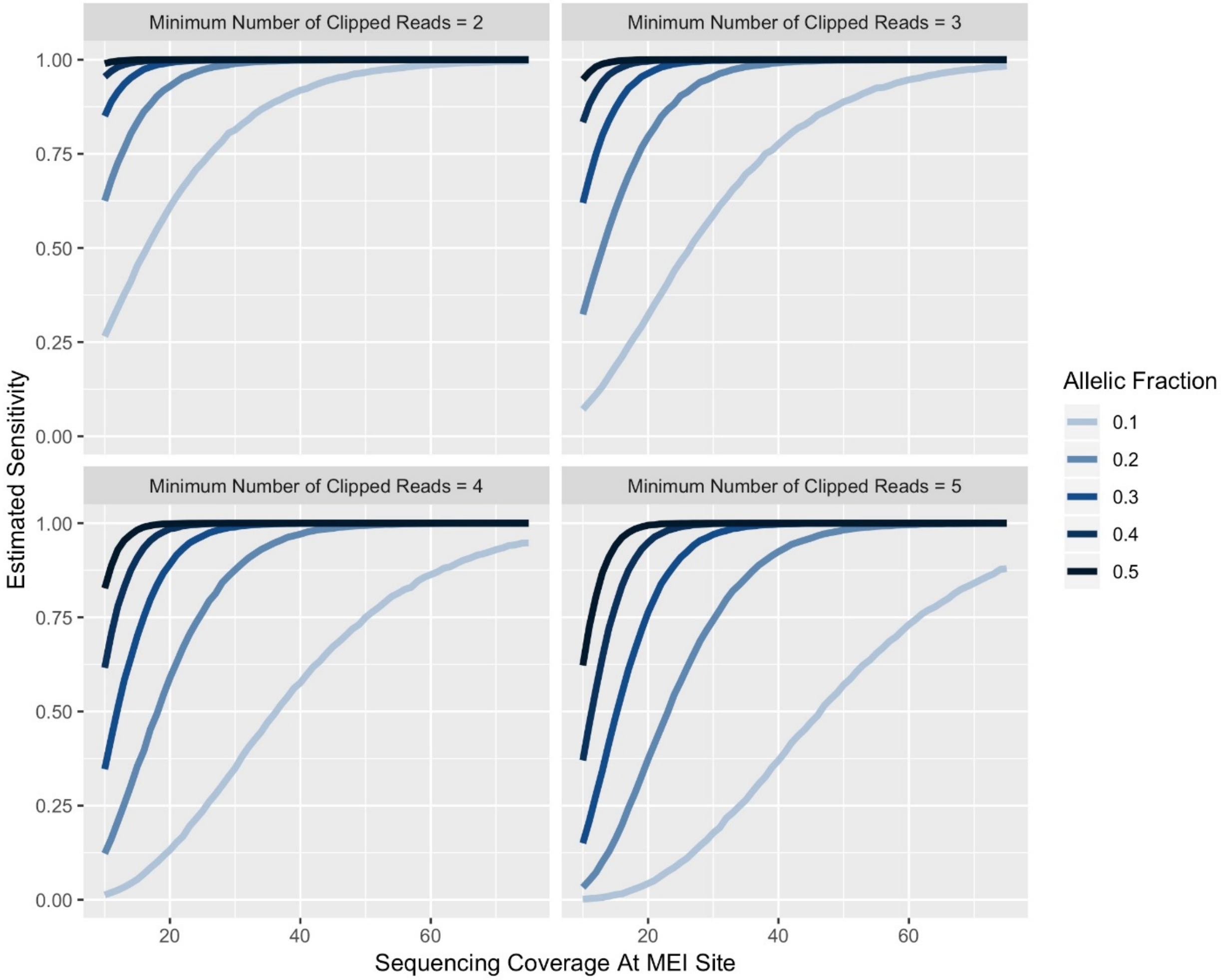

Subset NGS alignments for soft-clipped reads
Identify clusters of reads, clipped at same genomic position
Create a majority rules consensus for the clipped sequences
Pairwise align clipped consensus to MEI library
Filter for high quality alignments
Output MEI calls

GGCCGGGCGCGGTGGCTCA
CGCCTGTAATCCCAGCACTT
TGGGAGGCCGAGGCGGGT
GGATCATGA

Insertion	chr14:92471229
MEI Family	Alu
Direction	Plus
Supporting Reads	15
Alignment Score	136.8

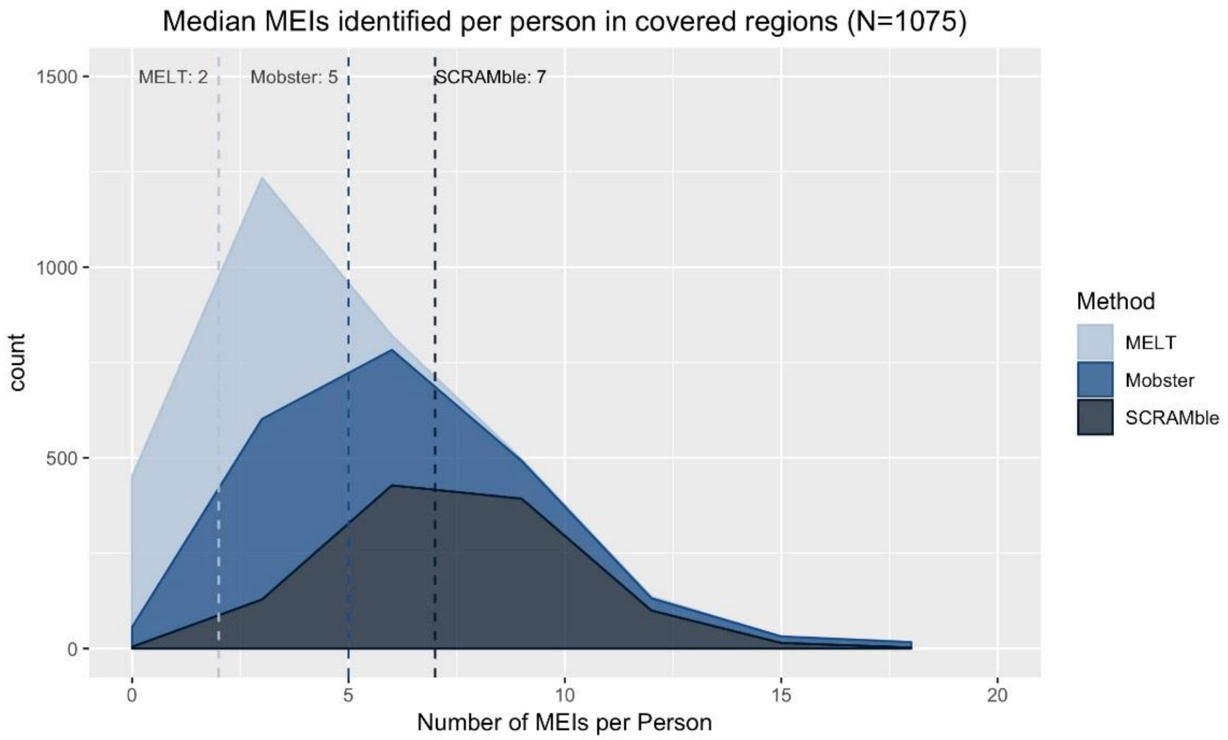
Starting from aligned sequences (BAM files), SCRAMble identifies reads which partially align to the reference genome (i.e., soft-clipped reads). It then forms clusters of soft-clipped reads and builds a consensus sequence of the clipped reads for each cluster site. The clipped consensus sequences, and their reverse complements, are aligned to a reference library of MEI sequences for L1, Alu, and SVA. When run at default settings, alignments with scores > 50 and percent identity >90 are used to define MEIs. SCRAMble then scans a 200 bp window for additional clusters that provide evidence of a second MEI breakpoint in order to identify the target site duplication. On the left panel, top, is an example of how clipped reads appear when a BAM file is viewed in IGV. The left, middle panel shows an example clipped consensus sequence for an Alu. The left, bottom panel shows an example of some of the information that SCRAMble provides when an MEI call is made.

Figure S4. Estimated sensitivity of SCRAMble given sequencing coverage at MEI variant site and allelic fraction of clipped reads.



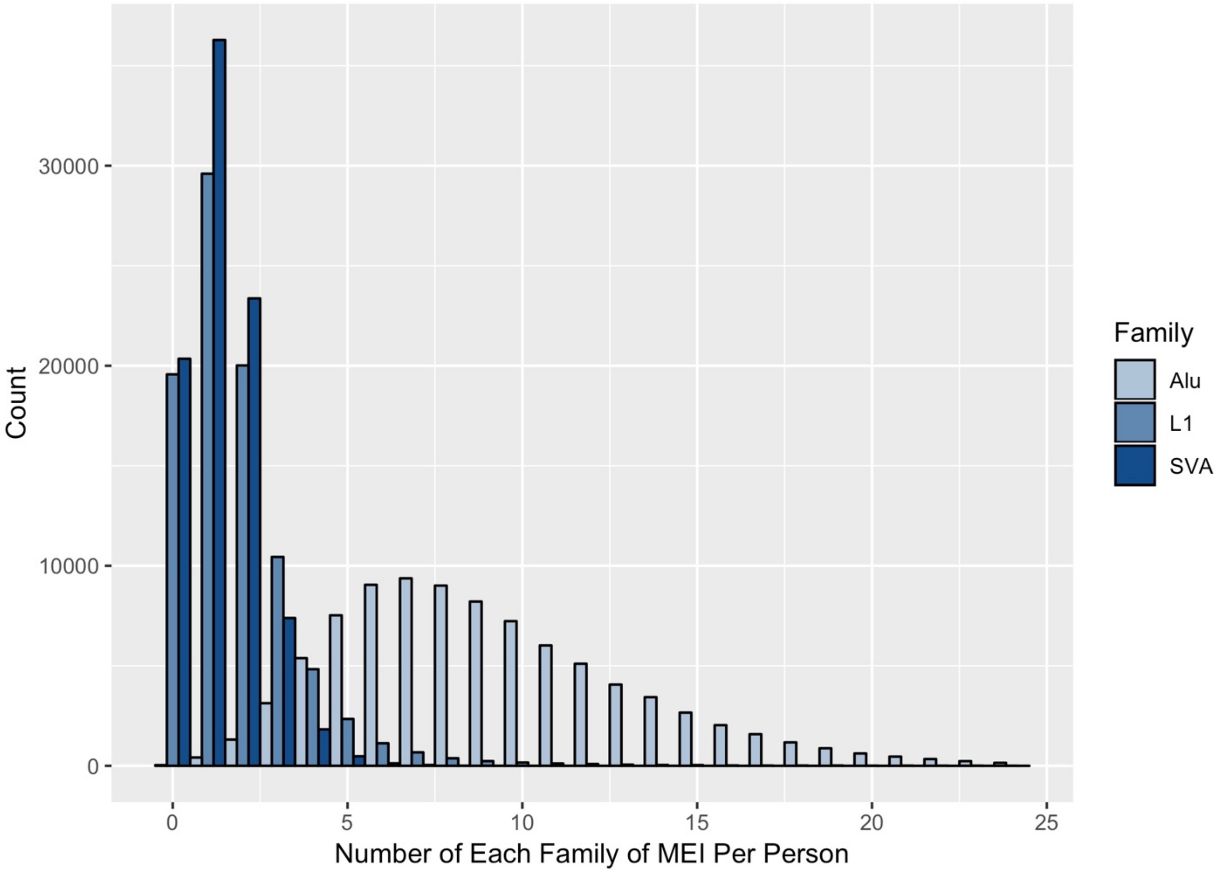
Simulated data based on allelic fraction probabilities were used to determine how often at each sequencing coverage we would expect to see the minimum number of reads required for SCRAMble to detect a given MEI. This analysis assumes that the clipped consensus is long enough to meet SCRAMble's alignment score requirement.

Figure S5. Median number of MEIs detected by MELT, Mobster, and SCRAMble



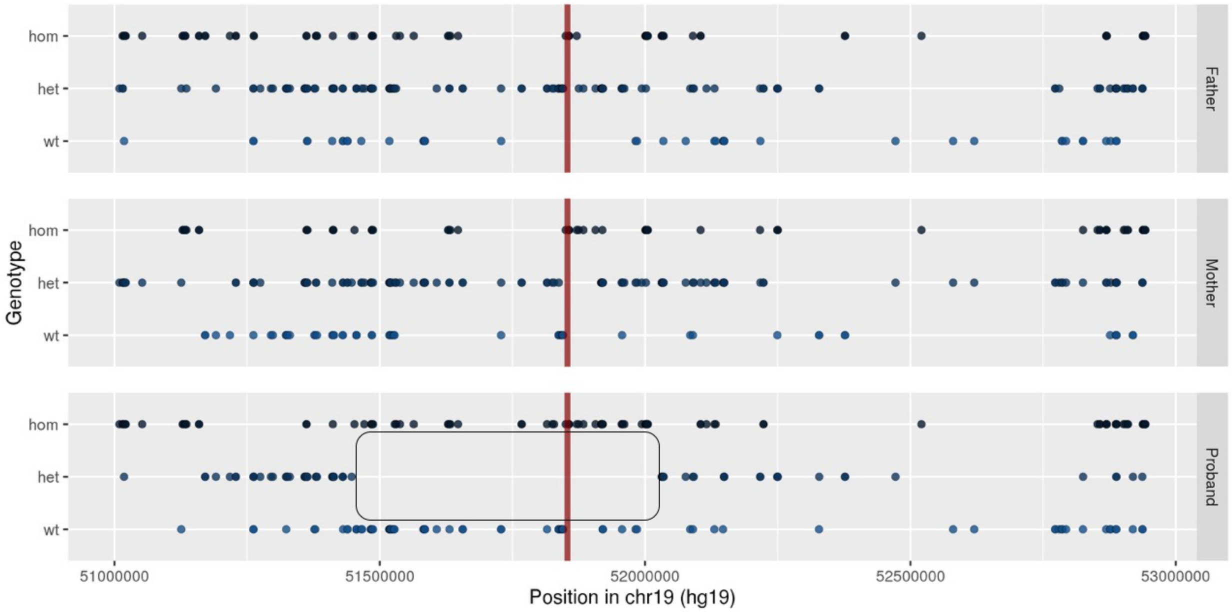
Smoothed histogram of the number of MEIs detected using SCRAMble and using two discordant read-pair based methods, MELT[6] and Mobster[9] in covered regions (within 250 bp of the center of a target probe). In our samples, SCRAMble identified more MEIs per person than either of the other two methods.

Figure S6. Number of MEI calls per person



Histogram of the number of calls of each MEI family found in each individual in the cohort.

Figure S7. Run of homozygosity surrounding of *ETFB* Alu insertion



The location of wildtype (wt), heterozygous (het), and homozygous (hom) variants along chromosome 19 (hg19) are plotted for all members of the trio with the exon 4 *ETFB* pathogenic MEI. The location of the MEI is noted with a red line. Mother and father are heterozygous carriers while the prenatal proband is homozygous for the MEI. A run of homozygosity (ROH) in the proband around the MEI variant site is circled. Given the size of the ROH block and local recombination rates, it is estimated that the MEI occurred 35 generations ago. No other individuals in the cohort have this MEI.

Supplemental References

1. Gardner EJ, Prigmore E, Gallone G, et al. Contribution of retrotransposition to developmental disorders. *Nat Commun*. 2019;10(1):4630. doi:10.1038/s41467-019-12520-y
2. Chen J-M, Chuzhanova N, Stenson PD, Férec C, Cooper DN. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Hum Mutat*. 2005;25(2):207-221. doi:10.1002/humu.20133
3. Zook JM, Hansen NF, Olson ND, et al. A robust benchmark for germline structural variant detection. *bioRxiv*. 2019. doi:10.1101/664623
4. Iskow RC, McCabe MT, Mills RE, et al. Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. *Cell*. 2010;141(7):1253-1261. doi:10.1016/j.cell.2010.05.020
5. Lee E, Iskow R, Yang L, et al. Landscape of Somatic Retrotransposition in Human Cancers. *Science (80-)*. 2012;337(6097):967-971. doi:10.1126/science.1222077
6. Gardner EJ, Lam VK, Harris DN, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*. 2017;27(11):1916-1929. doi:10.1101/gr.218032.116
7. Collins RL, Brand H, Karczewski KJ, et al. An open resource of structural variation for medical and population genetics. *bioRxiv*. 2019:578674. doi:10.1101/578674
8. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat*. 2006;27(4):323-329. doi:10.1002/humu.20307
9. Thung DT, de Ligt J, Vissers LE, et al. Mobster: accurate detection of mobile element

insertions in next generation sequencing data. *Genome Biol*. 2014;15(10):488. doi:10.1186/s13059-014-0488-x

10. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genet Med*. 2016;18(7):696-704. doi:10.1038/gim.2015.148

11. Ying D, Sham PC, Smith DK, Zhang L, Lau YL, Yang W. HaploShare: identification of extended haplotypes shared by cases and evaluation against controls. *Genome Biol*. 2015;16(1):92. doi:10.1186/s13059-015-0662-9