# TypeTE: a tool to genotype mobile element insertions from whole genome resequencing data

Clement Goubert[1a], Jainy Thomas[2a], Lindsay M. Payer[3], Jeffrey M. Kidd[4], Julie Feusier[2], W. Scott Watkins[2], Kathleen H. Burns[3], Lynn B. Jorde[2*] and Cedric Feschotte[1*]

[1]Department of Molecular Biology and Genetics, 215 Tower Rd, Cornell University, Ithaca, New York 14853, USA

[2]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA

[3]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[4]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed: Tel: (+1) 607 255-8793; Fax: +1; Email: cf458@cornell.edu Correspondence may also be addressed to Tel: (+1)801-581-4566; Fax: (+1); Email: lbj@genetics.utah.edu

*Co-senior authors

[a]The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint first authors

## ABSTRACT

*Alu* retrotransposons account for more than 10% of the human genome, and insertions of these elements create structural variants segregating in human populations. Such polymorphic *Alu* are powerful markers to understand population structure, and they represent variants that can greatly impact genome function, including gene expression. Accurate genotyping of *Alu* and other mobile elements has been challenging. Indeed, we found that *Alu* genotypes previously called for the 1000 Genomes Project are sometimes erroneous, which poses significant problems for phasing these insertions with other variants that comprise the haplotype. To ameliorate this issue, we introduce a new pipeline -- TypeTE -- which genotypes *Alu* insertions from whole-genome sequencing data. Starting from a list of polymorphic *Alu*s, TypeTE identifies the hallmarks (poly-A tail and target site duplication) and orientation of *Alu* insertions using local re-assembly to reconstruct presence and absence alleles. Genotype likelihoods are then computed after re-mapping sequencing reads to the reconstructed alleles. Using a 'gold standard' set of PCR-based genotyping of >200 loci, we show that TypeTE improves genotype accuracy from 83% to 92% in the 1000 Genomes dataset. TypeTE can be readily adapted to other retrotransposon families and brings a valuable toolbox addition for population genomics.

## INTRODUCTION

Mobile element insertions (MEIs) are ubiquitous and are major contributors to genomic variation between and within species (Kidwell and Lisch 1997; Sudmant et al. 2015; Underwood, Henderson, and Martienssen 2017). Active ME families continuously generate new MEIs which segregate among individuals. Individual MEI generate structural variants (SV) between genomes (typically insertions) and can lead to complex chromosomal rearrangements through non-homologous recombination between copies (Jurka et al. 2004; Song and Boissinot 2007; Xing et al. 2009; Thomas, Perron, and Feschotte 2018). Both processes represent a substantial source of genomic instability, which has been implicated in more than 100 human genetic diseases (Hancks and Kazazian 2016) , and they are also fodder for the emergence of adaptive genetic novelties (Oliver, McComb, and Greene 2013; Chuong, Elde, and Feschotte 2017; Wallace et al. 2018; Horváth, Merenciano, and González 2017; Jangam, Feschotte, and Betrán 2017)(Oliver et al. 2013; Chuong et al. 2017; Horváth et al. 2017; Jangam et al. 2017).

In humans, recently mobilized transposable elements (TEs) include members of the LINE-1, *Alu*, SVA, and a few human endogenous retroviruses (HERVs) families. Together these elements make up to >30% of the human genome, but relatively few remain polymorphic, *i.e.* being either present or absent between two genomes (Mills et al. 2007; Hancks and Kazazian 2012). Such polymorphic MEIs (pMEIs) account for hundreds to thousands of loci per individual (Stewart et al. 2011; Hancks and Kazazian

2012; Sudmant et al. 2015). The extent of pMEIs segregating in the human population is yet to be determined, but *Alu* is known to be the most common source of human pMEIs. Thus far, a little less than 20,000 *Alu* copies have been identified as segregating among 2,504 humans sampled as part of the 1000 Genomes Project (Sudmant et al, 2015; 1000 GP, Gardner et al., 2017).

*Alu* elements are powerful markers for genetic and evolutionary studies of human populations. As non-autonomous retrotransposons, *Alu*s amplify through a copy-and-paste mechanism utilizing LINE-1 machinery (Dewannieux, Esnault, and Heidmann 2003) and are inherently incapable of precise excision, providing identical-by-descent loci virtually free of homoplasy (Doronina et al. 2019). Accordingly, *Alu* have been shown to effectively track human population history (Watkins et al. 2003; Jurka, Bao, and Kojima 2011; Stewart et al. 2011; Rishishwar, Tellez Villa, and Jordan 2015). Like most MEIs, *Alu* insertions in humans are usually thought of as neutral variants that achieve fixation in the population through genetic drift (Boissinot et al. 2006; Cordaux et al. 2006). Nevertheless, more than 70 *de novo Alu* insertions are known to cause genetic diseases (Hancks and Kazazian 2016), including neurological disorders (Larsen et al. 2018; Hueso et al., n.d.). Furthermore, polymorphic *Alu* insertions have been identified as candidate causative variants in common polygenic diseases (Payer et al. 2017), and a handful have been shown to alter mRNA splicing (Payer et al. 2019). Finally, worldwide reference pMEI datasets such as those produced by 1000 GP (Sudmant et al. 2015) can be used in conjunction with gene expression data (e.g. RNA-seq) to identify loci associated with changes in gene expression (S. Wang et al. 2016).

Together these studies suggest that pMEIs, and *Alu*s in particular, play an important, yet still underappreciated role in human phenotypic variation.

Recognizing the abundance and biological significance of MEIs, a growing number of software packages have been developed in the past few years to detect and map pMEIs in whole-genome resequencing (WGS) data relative to a reference genome (Goerner-Potvin and Bourque 2018). For studies of human pMEIs, Tea (Lee et al. 2012), Retroseq (Keane, Wong, and Adams 2013), Mobster (Thung et al. 2014), Tlex2 (Fiston-Lavier et al. 2015), RelocaTE2 (J. Chen et al. 2017), STEAK (Santander et al. 2017), MELT (Gardner et al. 2017), TranSurVeyor (Rajaby and Sung 2018), polyDetect (Jordan et al. 2018), and ERVcaller (X. Chen and Li 2019) are among the most recent software tools available. The algorithmic refinement dedicated to accurately detecting pMEIs, and *Alu*s in particular, in WGS data has led to an increase of the quality of the calls. Notably, the accurate detection of the presence or absence of a specific *Alu* at a precise breakpoint has improved substantially in recent years (Rishishwar, Mariño-Ramírez, and King Jordan 2016; Gardner et al. 2017; X. Chen and Li 2019).

Although the discovery of *Alu* and other pMEI alleles is generally benchmarked extensively when these methods are evaluated, far less attention has been paid to individual genotyping, *i.e.* determining whether the insertion is a homozygote or heterozygote for each individual locus. Genotyping accuracy is critical for phasing insertion polymorphisms with single nucleotide polymorphisms (SNPs) and relating insertions with expression quantitative trait loci (eQTL) and disease-risk loci identified by

genome wide association studies (GWAS). Similarly, accurate genotypes are necessary to infer how the effects of drift and selection influence allele frequencies. However, genotyping accuracy of pMEI released with the 1000 GP dataset has only been estimated using 250 bp Illumina reads (accuracy estimated to 98%) (Sudmant et al., 2015) . To our knowledge, only three pipelines, MELT (Gardner et al. 2017), polyDetect (Jordan et al. 2018), and ERVcaller (X. Chen and Li 2019) are maintained as tools that directly allow genotyping for non-reference pMEIs. However, MELT is the only one offering the option to directly genotype reference pMEIs (*i.e.* polymorphic elements that are annotated in the reference genome but still segregating in the population). None of these tools have been subject to a comprehensive evaluation of their genotyping performance. Given the ever-growing number of resequencing efforts, there is a pressing need to develop highly accurate genotyping tools to complement the diverse methods already available to detect the presence or absence of pMEI.

To address these issues, we have developed a new bioinformatics pipeline, TypeTE, which improves the genotyping of pMEIs located by other tools using whole genome resequencing data. Our method is based on the accurate recreation of both the presence and absence of pMEI alleles before the remapping of reads for genotyping. We benchmarked TypeTE with both low- and high-coverage data [(1000 GP phase 3 (Sudmant et al. 2015) and Simons Genome Diversity Project, (SGDP) (Mallick et al. 2016) respectively] and show, based on a collection of more than 200 PCR-based genotyping assays, that our method significantly improves genotype quality. In addition, we applied TypeTE to all polymorphic *Alu* insertions discovered in 445 human samples

146 both present in the 1000 GP phase 3 (low-coverage WGS) and the Genetic European
147 Variation in Disease Consortium (GEUVADIS; RNA sequencing) (Lappalainen et al.
148 2013). We thus provide a new high-quality genomic resource dedicated to the functional
149 and evolutionary analysis of polymorphic *Alu* insertions.
150

151 ## MATERIAL AND METHODS

152

153 ### Pipeline implementation

154 *Non-reference MEI*. TypeTE-*non-reference* is designed to genotype insertions not found
155 in the reference genome (Figure 1A,Supplementary Figure S1). Based on the
156 information provided in a vcf file (such as produced by MELT), the location and
157 orientation of each *Alu* insertion are first collected. For each breakpoint, reads that are
158 mapped in a window of 500 bp (250 bp upstream and downstream of the breakpoint)
159 are extracted. The mates of discordant reads (mapping somewhere else in the genome)
160 are also extracted from the BAM file of each individual. The reads from all individuals for
161 each locus are then combined, and a local *de-novo* assembly of all the reads is
162 attempted using SPAdes v3.11.1 (Bankevich et al. 2012). Minia (v2.0.7) (Chikhi and
163 Rizk 2013) is used an alternate assembler when SPAdes failed to generate an
164 assembly of the sequences ('scaffolds.fasta'). The genomic locations where mates of
165 discordant reads are mapped are identified and intersected with the respective
166 RepeatMasker track (we used the coordinates version hg19 for 1000 GP data and hg38
167 for the SGDP data; Repbase version 20140131). Using a majority rule, the most likely
168 *Alu* subfamily consensus for the copy inserted at that locus is identified. To verify
169 orientation and identify target site duplications (TSDs), homology-based searches are
170 performed. First, the identified *Alu* consensus is searched with blastn (v. 2.6.0+) against
171 the assembled contigs. Then, a second blastn is performed using the genome reference
172 sequence (500 bps window) against the assembled contigs. The contig with the highest
173 score from the query *Alu* and the reference sequence is selected and searched for
174 target site duplications flanking the MEI. To identify the strand of the MEI, the sequence

flanking the insertion in the contig is further compared with the reference sequence. For each MEI, the two alleles are reconstructed as follows: a new window of +/- 500bp is extracted upstream and downstream of the breakpoint predicted by MELT. This represents the “absence” allele. To recreate the “presence” allele, TypeTE first removes the predicted TSDs from the extracted reference sequence and inserts the fully assembled MEI with its two TSDs in the correct orientation. If the assembly fails to generate a complete sequence of the MEI with flanking TSDs, the TSD predicted by the TE detection program (in our case MELT) is duplicated and placed at the 5’ and 3’ end of the consensus MEI in the composite allele.

**<FIGURE 1>**

*Reference TE.* TypeTE-*reference* determines genotypes of *Alu*s in the reference genome that are polymorphic in other individuals (Figure 1B, Supplementary Figure S2). In this case, no reads are extracted from the original alignments to reconstruct the alternate allele. However, the exact coordinates and TSDs of each MEI in the reference genome are reassessed as follows: the breakpoints identified from MELT for the location of the reference TE are further refined using the corresponding RepeatMasker annotation track to identify the exact location and orientation of each TE inserted in the reference genome. At first, the longest reference *Alu* elements that are within +/-50 bps of the predicted MELT breakpoints are extracted. If none is found within that boundary, *Alu*s within +/-110 bps of the predicted breakpoints are collected. However, we did not find any difference in the number of elements identified after increasing the boundary up

198 to 200bps. The flanking sequence of the TE sequence is also extracted and TSDs and
199 their coordinates are identified whenever possible. Then, based on these new
200 coordinates, a region of +/- 500 bps upstream and downstream of the 5' and 3' end of
201 the MEI is extracted from the reference genome. This constitutes the "presence" allele.
202 The "absence" allele is defined by removing the TE sequence and one TSD from the
203 reference genome.
204
205 *Genotyping*. TypeTE automatically generates input files and parallelizes the method
206 developed by Wildschutte et al. (Wildschutte et al. 2015), called *insertion-genotype,* to
207 genotype each *Alu* insertion in every individual. Briefly, read-pairs with at least one read
208 mapping to the target locus are extracted and mapped against the reconstructed
209 insertion and empty site alleles using bwa (v. 0.7.16a) (H. Li and Durbin 2009). The
210 number of reads that align to each allele, and their associated mapping quality values
211 are tabulated and likelihoods for the three possible genotype states are calculated (H. Li
212 2011). Reads that map equally well to the empty and insertion alleles are assigned a
213 mapping quality of 0 by bwa (H. Li and Durbin 2009) and do not contribute to this
214 calculation. Additionally, read pairs are required to partially align to the repeat sequence
215 and pairs that align entirely within the target repeat sequence are ignored, since these
216 reads may not be specific to the targeted locus. By default, the genotype with the
217 highest likelihood is chosen, but the resulting likelihoods may optionally be used as
218 inputs to downstream programs which estimate genotypes based on patterns across
219 multiple samples and sites. After genotyping, individual per-sample VCFs are
220 concatenated.

221

222 **Evaluation of the 1000 GP genotypes quality and TypeTE performance**

223 *Genotype calling.* In order to evaluate the quality of the *Alu* genotype calls available in
224 the 1000 GP phase 3 structural variants (SV) dataset ([Sudmant et al. 2015], average
225 depth of coverage 7.4X), we gathered the genotypes available for both non-reference
226 (indicated by "<INS:ME:ALU>" in the available VCF file) and reference (tagged with
227 "SV_TYPE=DEL_ALU"). We ran TypeTE-reference and TypeTE-non-reference on the
228 same loci as well as MELT-discovery (non-reference) and MELT-deletion (reference)
229 using its version 2.1.4 (referred to as MELT2 for the remainder of the manuscript) in
230 order to take into account, the most recent changes added to its genotyping module.
231 Additionally, we tested the performances of TypeTE with samples from the SGDP
232 (Mallick et al. 2016), which has higher coverage (average 42X).

233

234 In the 1000 GP data, we ran TypeTE and MELT2 on 445 CEU, TSI, GBR, FIN and YRI
235 individuals, also present in the Geuvadis dataset (RNA-seq) (Lappalainen et al. 2013).
236 In the 1000 GP dataset released by Sudmant et al. (Sudmant et al. 2015), Alu
237 genotypes were produced by MELT (first version) for non-reference insertions.
238 However, polymorphic reference Alu insertions were first discovered along with other
239 genomic deletions with a set of SV detection tools (BreakDancer, Delly, CNVnator,
240 GenomeSTRiP, Variation-Hunter, SSF and Pindel), then genotyped with the same
241 algorithm as any other SV (Sudmant et al. 2015). Because the sample size we used
242 was smaller than the original one (n = 445 vs n = 2504), MELT2 did not recover all loci
243 genotyped by the 1000 GP and TypeTE. Also, probably because of changes in the

newer version, some Alu breakpoints were slightly different between the Sudmant et al. (Sudmant et al. 2015) dataset and the MELT2 output. Thus, in order to reconcile and compare the three datasets, bedtools intersect (v.1.5) (Quinlan et al. 2010) was used with a window of +/- 30bp around each original 1000 GP Alu breakpoint. Finally, the predicted genotypes were compared to PCR assays of 108 non-reference and 43 reference loci in 42 individuals from the CEU population (see next section).

For the SGDP data, reference and non-reference polymorphic *Alu* insertions were called using MELT2 in 14 publicly available individuals from the South Asian population for which we had access to DNA. The genotypes of the loci discovered were then determined using TypeTE and compared to 9 non-reference and 67 reference loci previously genotyped by PCR in these 14 samples (Watkins et al. 2003).

*PCR typing in a subset of 1KGP and SGDP dataset.* Non-reference *(108)* and reference *(43) Alu* loci identified in 1000 GP were tested in a 30-trio reference panel of CEPH CEU individuals (42 individuals were evaluated by PCR and sequenced in 1000 GP) (HAPMAPPT01, Coriell Institute for Medical Research). Primers flanking the *Alu* insertion site were selected using Primer3 (Untergasser et al. 2012). PCR amplifications were performed using One*Taq* Hot Start Quick-Load 2x Master Mix (New England BioLabs) using 3-step PCR (initial denaturation: 94$^{o}$C, 15", (94$^{o}$C, 15''; 57$^{o}$C, 15''; 68$^{o}$C, 30'') for 30 cycles; final extension 68$^{o}$C, 5'). Sequences for 20 new primer pairs are available in Table S1; the remainder are available in (Payer et al. 2017). Accuracy was evaluated by replication in duplicate samples and by evaluating the number of Mendelian errors in related individuals. Non-reference (9) and reference (67) *Alu* loci

were previously genotyped by PCR in 14 South Asian samples present in the SGDP dataset (Watkins et al. 2003). Primers around each *Alu* insertion were selected using Primer3 (Untergasser et al. 2012). PCR amplification was performed using three-step PCR (initial denaturation: 94$^{o}$C, 3'; (94$^{o}$C, 15''; 60$^{o}$C, 15''; 72$^{o}$C, 30'') for 30 cycles; final extension 72$^{o}$C, 5') in 1X PCR buffer (10mM Tris, pH 8.3, 50mM KCl, 1.5 mM MgCl2) with 200 uM dNTPs, 10 pmol each primer, and 1U Taq polymerase. Annealing temperature was adjusted for each primer set. DMSO (5-10%) was used to improve amplification for some loci.

**Effect of genotype corrections on the *Alu* insertion discovery**

In some cases, new genotyping changed the presence/absence status of an *Alu* insertion for a given genome. We define a false positive (FP) as a case in which an *Alu* copy is called present, either homozygote or heterozygote in one sample, while the PCR reported it absent. A false negative (FN) is recorded when an *Alu* is called absent (homozygote absent ) while it is called as either homozygote present or heterozygous by PCR. True positive (TP) and true negative (TN) are the same calls (presence/absence), respectively, being validated by PCR. For each dataset and method, we calculated the sensitivity (ability of the method to discover a MEI: TP/(TP+FN)), the precision (or positive predictive value: TP/(TP+FP)) as well as the F1 score as described by Rishishwar et al (Rishishwar, Mariño-Ramírez, and King Jordan 2016), which corresponds to the harmonic mean of sensitivity and precision and summarizes the overall performance of each method.

290 **Estimation of mappability scores**

291 The mappability scores are downloaded for the GRCh37/hg19 version of reference
292 assembly for 100mers
293 (ftp://hgdownload.soe.ucsc.edu/gbdb/hg19/bbi/wgEncodeCrgMapabilityAlign100mer.bw
294 ). The downloaded file is processed (Kent et al. 2010) and is converted to bed format
295 (Neph et al. 2012). These data are stored in an indexed mysql table. The mappability
296 scores for genomic regions in the flanking region (+/- 250bps) of the predicted *Alu*
297 breakpoint for non-reference insertions and flanking region (+/- 250bps) of the reference
298 *Alu* insertions are extracted from the table, and the mean of the mappability scores is
299 recorded in a dedicated table and is provided with the output files.

300

301 **Calculation of local read depth**

302 The average read depth at genomic regions in the flanking region (+/- 250bps) of the
303 predicted *Alu* breakpoint for non-reference insertions and flanking region (+/- 250bps) of
304 the reference *Alu* insertions is calculated using samtools (Version: 1.4.1). Only reads
305 with a mapping quality of 20 or more (mapped with > 99% probability) and bases with a
306 quality of 20 or more (base call accuracy of > 99%) are counted.

307

308 **Inbreeding Coefficient (*Fis*) estimates.**

309 In order to assess how genotype quality affects common population genetics summary
310 statistics, we computed the per locus inbreeding coefficient (*Fis*) for the loci assayed by
311 PCR. $F_{is}$ is a common metric used in population genetics to assess the excess ($F_{is} < 0$)

or the depletion ($F_{is}$ > 0) in heterozygotes relative to the expected genotypes proportion at Hardy-Weinberg equilibrium. Allele frequencies were calculated using the genotypes produced by each method (1000 GP, MELT2, TypeTE and PCR) as follows:

$$F_{is} = \frac{H_{exp} - H_{obs}}{H_{exp}}$$

with $H_{exp}$ = 2pq, p = presence allele insertion frequency, q = (1-p) $H_{obs}$ is the observed number of heterozygotes.

All statistical analyses were carried out with R version 3.5.1 (R Core Team 2018).

## RESULTS

### Concordance of the 1000 GP dataset genotypes with PCR assays

*Alu* genotype predictions in the 1000 GP phase 3 release (Sudmant et al. 2015) were called using MELTv1.0 (first version) for non-reference loci and a combination of SV tools (Sudmant et al. 2015) for reference insertions. To assess their accuracy, we compared them to an assembled collection of 108 non-reference and 43 reference loci genotyped by PCR (Figure 2A and Table 1) in 42 individuals (see methods). To ensure accuracy in genotyping validations, PCR assays were performed using all (30) trios of the CEPH CEU, and in all cases, no Mendelian errors in the transmission of alleles from parents to offspring were seen (see methods). Presence of both “empty” and “filled” alleles (with and without *Alu*) were confirmed by the presence of bands of expected size in the agarose gel electrophoresis and in most cases with Sanger sequencing. Upon comparing the genotype predictions to PCR assays, we found that the 1000 GP phase

335 3 release had an overall concordance rate with the PCR (total number of prediction
336 identical to the PCR generated genotypes/ total number of predictions) of 83.31%
337 (3649/4380) for non-reference *Alu* insertions and 80.72% (1248/1590) for reference
338 insertions.
339
340 **<FIGURE 2>**
341
342 **TypeTE pipeline overview**
343 In order to improve the quality of *Alu* genotyping by short read sequencing analysis, we
344 developed TypeTE which allows the re-genotyping of both reference and non-reference
345 *Alu* insertions. The pipeline is divided into two main modules: the *non-reference* module
346 genotypes *Alu* insertions absent from the reference genome, while the *reference*
347 module genotypes *Alu* insertions present in the reference genome. Details about the
348 implementation of each module are given in the Material and Methods section (Figure
349 1,Supplemental Figure S1 and Supplemental Figure S2). The basic principle of TypeTE
350 is to recreate the most accurate sequences for the two alleles of each insertion
351 (presence and absence). TypeTE currently uses a VCF file such as produced by a TE
352 discovery tool such as MELT to locate each individual TE insertion. The pipeline then
353 performs an independent analysis of each predicted locus and collects the information
354 regarding the insertion. After allele reconstruction (see Material and Methods), the
355 individual reads mapping to each insertion locus are extracted from the original
356 alignment file (bam) and mapped against the reconstructed alleles for genotyping using
357 an automated and parallelized version of the method developed by Wildschutte et al.

(2015). A new VCF file with the corrected genotypes and genotypes likelihoods is then produced.

**TypeTE performances**

In order to assess the accuracy of the predictions made by TypeTE, we ran the pipeline on a subset of 445 individuals of European and African ancestry included in the 1000 GP dataset (see Material and Methods). These samples were selected because they are both represented in the 1000 GP (WGS) and GEUVADIS (RNA-seq) datasets, allowing functional analyses of pMEIs. We also compared the performance of TypeTE with a recent version of MELT (version 2.1.4, abbreviated as MELT2) using the packages MELT-discovery and MELT-deletion on the same sample. TypeTE and MELT2 genotypes were then compared to 108 non-reference and 43 reference insertion for which we have collected or generated PCR genotypes. With non-reference insertions, MELT2 shows increased concordance with the PCR compared to the original 1000 GP calls, with 87.95% (vs 83.31%; +131/4298 accurate genotypes) of the predicted genotypes matching the experimental results. TypeTE further increases the concordance of the genotype prediction, achieving a rate of 92.14% (+325/4313 accurate genotypes compared to original 1000 GP release). For reference insertions, MELT2 had a lower concordance than the original 1000 GP predictions, with only 71% (vs. 80.72%; -374/1504 genotypes) of the genotypes matching the PCR results, while TypeTE achieved 91.56% concordance (+ 141/1575 genotypes). Note that the total

number of genotypes considered correspond to the total number of predictions available

| | 1000 GP |
|---|---|

and doesn't take into account here the missing genotypes.

We further tested the genotyping performance of MELT2 and TypeTE with the SGDP (Mallick et al. 2016) data, which benefits from a higher depth of coverage than the 1000 GP data (42x vs 7.4x). We tested the concordance of the predicted genotypes with 67 reference and 9 non-reference *Alu* loci in 14 individuals previously genotyped by PCR (Watkins et al. 2003). MELT2 has a concordance rate of 70.13% for reference loci while TypeTE matches the PCR results for 91.01% of the predicted genotypes (+181 correct genotypes; Figure 3 and Table 1). Finally, for the 9 non-reference loci that were experimentally genotyped, the concordance rate is 78.57% for MELT2 and 94.44% for TypeTE (+ 20 correct genotypes).

**<FIGURE3>**

In order to analyze in detail, the genotyping performances of each method, we calculated the concordance rate by genotype category (0 or (0/0): homozygote absent, 1 or (0/1): heterozygote, 2 or (1/1): homozygote present) corresponding to the percent of correct genotypes in one category to the total number of calls for this category (Table 1). Additionally, we report the percentage of unascertained loci (NA genotypes) for each method.

| | non-reference insertions (n=108x42) | | | | | reference insertions (n=43x42) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | hom ref (0) | het (1) | hom alt (2) | NAs | overall | hom ref (2) | het (1) | hom alt (0) | NAs | **overall** |
| 1000 GP | 98.92% | 98.28% | 23.49% | -* | 83.31% | 97.71% | 90.00% | 41.84% | -* | **80.72%** |
| MELT 2.1.4 | 99.02% | 92.27% | 68.01% | 1.90% | 87.95% | 98.63% | 37.97% | 26.62% | 5.37% | **71.00%** |
| TypeTE | 98.44% | 89.28% | 93.61% | 1.54% | 92.14% | 91.46% | 84.54% | 87.82% | 1.04% | **91.56%** |
| | SGDP | | | | | | | | | |
| | non-reference insertions (n=9x14) | | | | | reference insertions (n=67x14) | | | | |
| 1000 GP | - | - | - | - | | - | - | - | - | |
| MELT 2.1.4 | 92.31% | 94.87% | 9.09% | 0.00% | 79.57% | 97.93% | 88.26% | 2.95% | 0.00% | **70.13%** |
| TypeTE | 92.31% | 94.87% | 100.00% | 0.00% | 94.44% | 91.19% | 87.58% | 91.14% | 0.00% | **91.01%** |

NA: Not Applicable as no genotypes reported
*no NA genotype has been reported in the Sudmant et al. (2015) dataset

Table 1. Genotype prediction accuracy (%) for each category of insertions when compared with PCR generated genotypes

We then investigated how the concordance between predicted and PCR genotypes is distributed across loci and individuals by calculating the average concordance rate at each locus (total number of correct genotypes at a locus / total number of individuals with a predicted genotype). Regardless of the genotype category (reference / non-reference), TypeTE has a higher average concordance rate per locus, as well as lower variance for this value, than the other methods (Figure 4). The greatest improvement was when the genotypes of reference insertions were compared to MELT2, where the concordance rate of TypeTE is always significantly higher (Tukey's HSD, $P < 0.05$).

**<FIGURE 4>**

For each locus assayed by PCR in the 1000 GP dataset, we also examined whether the mappability and local read coverage affect genotyping predictions for TypeTE. We do not find a significant correlation between genotype concordance and the mappability score (0.1 - 1) computed in a 500-bp window around the MEI breakpoints

(Supplementary Figure S3. Pearson's product-moment correlation, r = 0.20, *P* = 0.281 for non-reference loci and r = 0.13, *P* = 0.414 for reference loci). We also found that the average depth of coverage for a given locus (4.69 X – 10.01 X) is not correlated to genotyping concordance for both reference (r = 0.12, *P* = 0.4538) and non-reference insertions (r = -0.01, *P* = 0.957) (Supplementary Figure S4). We conclude that at least for the loci tested by PCR, the level of repetitiveness of the flanking sequence of individual *Alu* insertions and the local read depth do not appear to influence the genotyping performance of TypeTE.

**Effect of genotype corrections on variant discovery**

Different methods can assign different genotypes for some loci due to the inherent differences in their approach or due to locus specific features. For example, a heterozygous locus for the presence of *Alu* can be genotyped either as homozygous presence or absence by different methods. We first converted the genotypes into presence/absence calls in order to assess sensitivity, precision (positive predictive value), and the overall detection accuracy, summarized by the F1 score (harmonic mean of sensitivity and precision, see Material and Methods) for each method considering PCR results as true genotypes. TypeTE received the highest F1 score in each dataset (1000 GP or SGDP) and for both types of insertion (reference or non-reference) (Fig 5). The small number of loci tested for the SGDP-non-reference dataset (n = 9) did not allow us to find significant differences between the methods; however, we show that the increased F1 score of TypeTE with the 1000 GP-non-reference loci is due

444 to a significant increase of the sensitivity compared to the other methods. Interestingly,
445 the higher F1 score of TypeTE with reference insertions (both for the 1000 GP and
446 SGDP datasets) is, in these cases, due to significantly higher precision (TP/(TP+FP)).
447
448 **<FIGURE 5>**
449
450 **Influence of re-genotyping on population genetics statistics**
451 To illustrate the importance of accurately genotyping of *Alu*s, we calculated the
452 population-wise inbreeding coefficient ($F_{is}$) for each locus in 42 individuals of the CEU
453 cohort (1000 GP) and 14 individuals of the South Asian cohort (SGDP). Compared to
454 the original 1000 GP and MELT2 genotypes, the $F_{is}$ values calculated with TypeTE
455 genotypes are concordant with the ones based on PCR genotypes. These results are
456 even more striking when only reference loci are considered: while TypeTE and PCR
457 estimates of $F_{is}$ are centered at 0, MELT2 and 1000 GP genotypes suggest a clear
458 deviation of most loci from Hardy-Weinberg equilibrium (Figure 6). We note that
459 estimates of the $F_{is}$ are more variable using the SGDP data, which can be explained by
460 its smaller sample size and a higher population subdivision (e.g. castes) than the 1000
461 GP dataset.
462
463 **<FIGURE 6>**
464
465 **Influence of the dataset quality on genotype prediction**
466

To discover factors specific to each dataset that influence genotype prediction, we compared the results obtained in the 1000 GP dataset (average depth of 7.4 X) with the results from analysis of the SGDP (average depth of 42X) to the respective PCR genotypes. The provenance of the dataset does not influence the variant discovery abilities of MELT2 and TypeTE (Supplementary Figure S5). However, we observe that the percentage of unascertained loci differs between the 1000 GP and SGDP datasets. About 1% to 5.4% of genotypes are not ascertained by either MELT2 and TypeTE in the 1000 GP dataset, probably due to low coverage. Conversely, all SGDP loci are called in every individual for the SGDP dataset (Table 1).

## DISCUSSION

The purpose of TypeTE is to provide automatic and reliable genotyping of pMEIs, especially *Alu*s from short read, whole genome or targeted interval sequencing. To our knowledge, MELT (Gardner et al. 2017) is the only tool with continued support and documentation that allows direct genotyping of both reference and non-reference pMEI. While its performance for variant discovery has made it a popular tool for pMEI mapping, to our knowledge its performance at genotyping has never been comprehensively tested. Moreover, there was no formal testing of the genotype quality concerning pMEI reported in the Phase 3 release of the 1000 GP. Thus, we thoroughly tested the *Alu* genotype predictions made for the 1000 GP, a recent version of MELT (v. 2.1.4), and TypeTE by assembling the results of more than 200 locus-specific PCR genotyping assays.

Combining reference and non-reference pMEI, our analysis indicates that 82% of the genotypes reported by the 1000 GP are consistent with the PCR assays which we consider the 'gold standard' and wherein accuracy was evaluated by comparing duplicate samples and verifying the absence of Mendelian errors in related individuals. This estimate of genotyping accuracy is much lower than a previous estimate of 98% based on long (250bp) Illumina reads (Sudmant et al. 2015). Genotypes reported by the 1000 GP were estimated using the first version of MELT for non-reference loci, but genotyping methods developed for other structural variation (indels, inversions, etc.) were used for reference insertions. While MELT2 appears to offer a noticeable improvement over its first version for genotyping non-reference pMEI, its overall genotyping performance is diminished when applied to reference loci, with genotyping errors reaching more than 20% in based on our PCR assays. For both categories of loci, most errors are caused by the underestimation of homozygous genotypes carrying the alternative allele relative to the reference genome (Table 1). We note that for non-reference insertions, MELT's genotyping algorithm benefited from improvements deployed in the version tested (v2.1.4) compared to its original release, in particular to detect homozygous insertion (1/1). However, this increased sensitivity to detect pMEI alleles from read alignments seems to be accompanied by a reduced power to detect "absence" alleles for reference insertions (MELT-deletion module). Such errors are consequential for population genetics analysis because they lead to inaccurate estimation of population genetics parameters. For example, calculation of the inbreeding coefficient ($F_{is}$) shows that the original release of the 1000 GP genotypes

513 was overestimating heterozygotes, leading to negative and likely inaccurate values of
514 $F_{is}$ (Figure 6). Genotypes obtained with MELT2 improve these estimates for non-
515 reference insertions, but the results appear less accurate when computed from a small
516 sample and they are more inaccurate for reference insertions. These issues underscore
517 the need for a tool dedicated to the genotyping of pMEI.
518
519 Toward this goal we developed TypeTE and applied it to genotype both reference and
520 non-reference *Alu* insertions. Our benchmarking data show that TypeTE has an
521 average concordance rate of 91% or greater with PCR-based genotyping. Importantly,
522 TypeTE maintains a genotyping accuracy greater than 84% under all genotyping
523 scenarios. While TypeTE performs better than MELT v1 (1000 GP) and MELT2 for non-
524 reference insertions, the most significant improvement is for reference insertions. In
525 particular, the genotypes predicted by 1000 GP and MELT2 never reached more than
526 41.8% concordance with the experimental results when the PCR called a homozygote
527 absence (0/0); by contrast, TypeTE predicted these genotypes with more than 87%
528 concordance in the two datasets tested (1000 GP and SGDP). Consequently,
529 calculation of $F_{is}$ based on TypeTE genotypes shows better concordance with that
530 based on PCR-derived genotypes, and fits the neutral expectation as we observe no
531 deviation from Hardy-Weinberg equilibrium for a single human population (Hosking et
532 al., 2014).
533
534 The principal difference between TypeTE and MELT derives from characteristics of the
535 actual data on which the genotyping is performed. While both methods implement the

core genotyping algorithm described by Li (H. Li 2011), TypeTE relies on a strategy based on re-alignment of the reads against both presence and absence alleles before computation of the genotype likelihoods, an approach initially introduced by Wildschutte et al. (Wildschutte et al. 2015). Furthermore, TypeTE facilitates the genotyping with no user intervention by using as input the 'vcf' produced by MELT (or virtually any other pMEI detection software) to generate a new 'vcf' output file delivering the predicted genotypes. TypeTE also uses recently developed assemblers (SPAdes (Bankevich et al. 2012) and Minia (Chikhi and Rizk 2013)) and use reads from all individuals for a locus for local MEI assembly which, in our hands, showed a higher rate of assembly than the CAP3 assembler (Huang and Madan 1999) used previously (Wildschutte et al. 2015). In addition, TypeTE can also genotype more pMEIs than previous studies based solely on *de-novo* MEI assembly (Wildschutte et al. 2015): if an incomplete Alu is assembled, TypeTE subsidize it with the exact consensus sequence based on TE's read identity with RepBase. This additional step is performed by retrieving the subfamily on which most discordant mates align in the assembly. Here, we show that reconstruction of alternative alleles (either by local assembly or consensus-based) -- a major difference with MELT -- significantly improves the accuracy of *Alu* genotyping. Finally, TypeTE predicts the TSD accompanying each insertion and the pMEI orientation, which ensures optimal reconstruction of the two alleles. Collectively these implementations enable TypeTE to generate highly accurate *Alu* insertion genotypes.

We further tested whether the quality of the starting dataset (in particular sequencing depth) influenced TypeTE performance. By comparing results on the 1000 GP and

559 SGDP datasets, which use different sequencing depth (on average 7.4X vs 42X,
560 respectively), to the PCR genotypes, we found that TypeTE performs equally regardless
561 of coverage depth (at least for reference insertions, for which we had enough loci to
562 compare between datasets). In fact, using both non-reference and reference *Alu*
563 insertions genotyped with TypeTE in the 1000 GP dataset, we showed that the average
564 sequence coverage of the region flanking these loci does not seem to influence
565 genotyping accuracy. Thus, TypeTE can support the analysis of large population
566 dataset without stringent or highly uniform coverage requirements.
567
568 While TypeTE offers significant improvements over MELT, it still fails to genotype
569 accurately some of the loci we experimentally assayed (16/227). Neither low
570 sequencing coverage nor mappability issues could be readily implicated as hindering
571 genotyping of these loci. We believe that other locus-specific idiosyncrasies prevent the
572 ability of TypeTE to produce an accurate allele call. For instance, earlier tests on the
573 pipeline showed that a 1-bp insertion at the end of the element in one allele or a slight
574 error in the TSD prediction could dramatically affect the re-mapping and genotype
575 predictions. A specific assessment of the bioinformatic methods aimed to identify TSDs
576 should be able to improve this issue. Identifying boundaries of *Alu* insertion in low
577 complexity (especially A-rich) regions is challenging due to inter-individual differences in
578 the length of the poly-A tail of the element, and according to our tests, Repeatmasker
579 often fails to identify the exact boundaries of such reference elements. Even though our
580 pipeline in principle considers such subtle sequence variation, at least for one locus, we
581 found that the TSD was overlapping the annotated poly-A region. Implementing

582 changes to identify similar instances could mitigate genotyping miscalls for those loci.
583 Additionally, our ability to evaluate the concordance of genotype predictions in low-
584 complexity and highly repetitive regions was restrained to PCR-accessible loci. We
585 have also noticed that altering the parameters or method for local *de novo* assembly
586 improved the assembly of certain TE loci. An automated approach to customize the
587 assembly parameters for each locus that failed with the standard approach would
588 enhance the reconstruction of non-reference TE sequences. Identifying proper
589 orientation of insertions is also crucial in accurately genotyping the insertions and we
590 are also contemplating a read-based approach to identify the orientation of insertions in
591 addition to the current assembly-based approach. Collecting more benchmarking data
592 might allow us to characterize more finely these issues and adapt the pipeline
593 accordingly. Notwithstanding these peculiar instances, TypeTE has the lowest error rate
594 of all methods tested and as such it represents a valuable advance in the field.
595
596 The task and challenges of pMEI genotyping have been largely overlooked thus far, yet
597 we show here that inaccurate genotyping of pMEIs can significantly bias population
598 genetics inferences. It is presumably because of these issues that reference pMEIs
599 have been entirely ignored in previous population genomics studies using pMEIs (L.
600 Wang et al. 2016; L. Wang, Norris, and Jordan 2017)). By increasing genotyping
601 accuracy for both reference and non-reference insertions, TypeTE will enhance future
602 studies using pMEI as markers or structural variants in the human population. Notably,
603 our results now offer a dataset of genotyped *Alu* insertions for 445 samples of the 1000
604 GP that is complemented by a wealth of functional data including RNA-seq

(Lappalainen et al. 2013), DNA methylation (Pai et al. 2011), DNase I accessibility (Degner et al. 2012), and ATAC-seq (Kumasaka, Knights, and Gaffney 2016, 2019). We anticipate that these resources will open new avenues to explore the cis-regulatory influence of pMEIs in humans (L. Wang et al. 2016; L. Wang, Norris, and Jordan 2017; Rishishwar et al. 2018). The modularity of TypeTE allows one to easily combine new assemblers to improve the reconstruction of each pMEI, but it is also possible to skip this step and only use consensus sequence of MEI to speed up the computation time. The design of TypeTE makes it compatible with any data produced by pMEI detection tools and in principle it can be readily adapted to genotype insertions from any other retroelement families in virtually any species.

## DATA AVAILABILITY

TypeTE is freely available in the Github repository https://github.com/clemgoub/TypeTE

## FUNDING

This work was supported by funds from the National Institutes of Health (R35 GM122550, R01 GM059290) to C.F and (GM118335 and GM059290) to L.B.J. Funding for open access charge: The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

## CONFLICT OF INTEREST

The authors do not declare any conflict any interest.

## REFERENCES

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77.

Boissinot, Stephane, Jerel Davis, Ali Entezam, Dimitri Petrov, and Anthony V. Furano. 2006. "Fitness Cost of LINE-1 (L1) Activity in Humans." *Proceedings of the National Academy of Sciences of the United States of America* 103 (25): 9590–94.

Chen, Jinfeng, Travis R. Wrightsman, Susan R. Wessler, and Jason E. Stajich. 2017. "RelocaTE2: A High Resolution Transposable Element Insertion Site Mapping Tool for Population Resequencing." *PeerJ* 5 (January): e2942.

Chen, Xun, and Dawei Li. 2019. "ERVcaller: Identifying Polymorphic Endogenous Retrovirus and Other Transposable Element Insertions Using Whole-Genome Sequencing Data." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz205.

Chikhi, Rayan, and Guillaume Rizk. 2013. "Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter." *Algorithms for Molecular Biology: AMB* 8 (1): 22.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2017. "Regulatory Activities of Transposable Elements: From Conflicts to Benefits." *Nature Reviews. Genetics* 18 (2): 71–86.

Cordaux, Richard, Jungnam Lee, Liv Dinoso, and Mark A. Batzer. 2006. "Recently Integrated Alu Retrotransposons Are Essentially Neutral Residents of the Human Genome." *Gene* 373 (May): 138–44.

Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. "DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation." *Nature* 482 (7385): 390–94.

Dewannieux, Marie, Cécile Esnault, and Thierry Heidmann. 2003. "LINE-Mediated Retrotransposition of Marked Alu Sequences." *Nature Genetics* 35 (1): 41–48.

Doronina, Liliya, Olga Reising, Hiram Clawson, David A. Ray, and Jürgen Schmitz. 2019. "True Homoplasy of Retrotransposon Insertions in Primates." *Systematic Biology* 68 (3): 482–93.

Fiston-Lavier, Anna-Sophie, Maite G. Barrón, Dmitri A. Petrov, and Josefa González. 2015. "T-lex2: Genotyping, Frequency Estimation and Re-Annotation of Transposable Elements Using Single or Pooled next-Generation Sequencing Data." *Nucleic Acids Research* 43 (4): e22.

Gardner, Eugene J., Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, 1000 Genomes Project Consortium, and Scott E. Devine. 2017. "The Mobile Element Locator Tool (MELT): Population-Scale Mobile Element Discovery and Biology." *Genome Research* 27 (11): 1916–29.

Goerner-Potvin, Patricia, and Guillaume Bourque. 2018. "Computational Tools to Unmask Transposable Elements." *Nature Reviews. Genetics* 19 (11): 688–704.

Hancks, Dustin C., and Haig H. Kazazian Jr. 2012. "Active Human Retrotransposons: Variation and Disease." *Current Opinion in Genetics & Development* 22 (3): 191–

203.———. 2016. “Roles for Retrotransposon Insertions in Human Disease.” *Mobile DNA* 7 (May): 9. 1.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu C-F. 2004 Detection of genotyping errors by Hardy–Weinberg equilibrium testing. Eur. J. Hum. Genet. 12, 395–399. (doi:10.1038/sj.ejhg.5201164)

Horváth, Vivien, Miriam Merenciano, and Josefa González. 2017. “Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response.” *Trends in Genetics: TIG* 33 (11): 832–41.

Huang, X., and A. Madan. 1999. “CAP3: A DNA Sequence Assembly Program.” *Genome Research* 9 (9): 868–77.

Hueso, Miguel, Josep M. Cruzado, Joan Torras, and Estanis Navarro. n.d. “ALUminating the Path of Atherosclerosis Progression: Chaos Theory Suggests a Role for Alu Repeats in the Development of Atherosclerotic Vascular Disease.” https://doi.org/10.20944/preprints201804.0051.v1.

Jangam, Diwash, Cédric Feschotte, and Esther Betrán. 2017. “Transposable Element Domestication As an Adaptation to Evolutionary Conflicts.” *Trends in Genetics: TIG* 33 (11): 817–31.

Jordan, Vallmer E., Jerilyn A. Walker, Thomas O. Beckstrom, Cody J. Steely, Cullen L. McDaniel, Corey P. St Romain, Baboon Genome Analysis Consortium, et al. 2018. “A Computational Reconstruction of Phylogeny Using Insertion Polymorphisms.” *Mobile DNA* 9 (April): 13.

Jurka, Jerzy, Weidong Bao, and Kenji K. Kojima. 2011. “Families of Transposable Elements, Population Structure and the Origin of Species.” *Biology Direct* 6 (September): 44.

Jurka, Jerzy, Oleksiy Kohany, Adam Pavlicek, Vladimir V. Kapitonov, and Michael V. Jurka. 2004. “Duplication, Coclustering, and Selection of Human Alu Retrotransposons.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (5): 1268–72.

Keane, Thomas M., Kim Wong, and David J. Adams. 2013. “RetroSeq: Transposable Element Discovery from next-Generation Sequencing Data.” *Bioinformatics* 29 (3): 389–90.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. 2010. “BigWig and BigBed: Enabling Browsing of Large Distributed Datasets.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq351.

Kidwell, M. G., and D. Lisch. 1997. “Transposable Elements as Sources of Variation in Animals and Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 7704–11.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. “Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq.” *Nature Genetics* 48 (2): 206–13.

———. 2019. “High-Resolution Genetic Mapping of Putative Causal Interactions between Regions of Open Chromatin.” *Nature Genetics* 51 (1): 128–37.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ’t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468): 506–11.

Larsen, Peter A., Kelsie E. Hunnicutt, Roxanne J. Larsen, Anne D. Yoder, and Ann M.

720 Saunders. 2018. “Warning SINEs: Alu Elements, Evolution of the Human Brain, and
721 the Spectrum of Neurological Disease.” *Chromosome Research: An International*
722 *Journal on the Molecular, Supramolecular and Evolutionary Aspects of*
723 *Chromosome Biology* 26 (1-2): 93–111.
724 Lee, Eunjung, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace
725 J. Luquette 3rd, Jens G. Lohr, et al. 2012. “Landscape of Somatic
726 Retrotransposition in Human Cancers.” *Science* 337 (6097): 967–71.
727 Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association
728 Mapping and Population Genetical Parameter Estimation from Sequencing Data.”
729 *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr509.
730 Li, H., and R. Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-
731 Wheeler Transform.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp324.
732 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor
733 Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data
734 Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.”
735 *Bioinformatics* 25 (16): 2078–79.
736 Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando
737 Racimo, Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300
738 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.
739 Mills, Ryan E., E. Andrew Bennett, Rebecca C. Iskow, and Scott E. Devine. 2007.
740 “Which Transposable Elements Are Active in the Human Genome?” *Trends in*
741 *Genetics*. https://doi.org/10.1016/j.tig.2007.02.006.
742 Neph, Shane, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman,
743 Audra K. Johnson, Eric Rynes, et al. 2012. “BEDOPS: High-Performance Genomic
744 Feature Operations.” *Bioinformatics* 28 (14): 1919–20.
745 Oliver, Keith R., Jen A. McComb, and Wayne K. Greene. 2013. “Transposable
746 Elements: Powerful Contributors to Angiosperm Evolution and Diversity.” *Genome*
747 *Biology and Evolution* 5 (10): 1886–1901.
748 Pai, Athma A., Jordana T. Bell, John C. Marioni, Jonathan K. Pritchard, and Yoav Gilad.
749 2011. “A Genome-Wide Study of DNA Methylation Patterns and Gene Expression
750 Levels in Multiple Human and Chimpanzee Tissues.” *PLoS Genetics* 7 (2):
751 e1001316.
752 Payer, Lindsay M., Jared P. Steranka, Daniel Ardeljan, Janiece Walker, Kathryn C.
753 Fitzgerald, Peter A. Calabresi, Thomas A. Cooper, and Kathleen H. Burns. 2019.
754 “Alu Insertion Variants Alter mRNA Splicing.” *Nucleic Acids Research* 47 (1): 421–
755 31.
756 Payer, Lindsay M., Jared P. Steranka, Wan Rou Yang, Maria Kryatova, Sibyl
757 Medabalimi, Daniel Ardeljan, Chunhong Liu, Jef D. Boeke, Dimitri Avramopoulos,
758 and Kathleen H. Burns. 2017. “Structural Variants Caused by Insertions Are
759 Associated with Risks for Many Human Diseases.” *Proceedings of the National*
760 *Academy of Sciences of the United States of America* 114 (20): E3984–92.
761 Rajaby, Ramesh, and Wing-Kin Sung. 2018. “TranSurVeyor: An Improved Database-
762 Free Algorithm for Finding Non-Reference Transpositions in High-Throughput
763 Sequencing Data.” *Nucleic Acids Research* 46 (20): e122.
764 Rishishwar, Lavanya, Leonardo Mariño-Ramírez, and I. King Jordan. 2016.
765 “Benchmarking Computational Tools for Polymorphic Transposable Element

766 Detection." *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbw072.
767 Rishishwar, Lavanya, Carlos E. Tellez Villa, and I. King Jordan. 2015. "Transposable
768 Element Polymorphisms Recapitulate Human Evolution." *Mobile DNA* 6
769 (November): 21.
770 Rishishwar, Lavanya, Lu Wang, Jianrong Wang, Soojin V. Yi, Joseph Lachance, and I.
771 King Jordan. 2018. "Evidence for Positive Selection on Recent Human
772 Transposable Element Insertions." *Gene*.
773 https://doi.org/10.1016/j.gene.2018.06.077.
774 Santander, Cindy G., Philippe Gambron, Emanuele Marchi, Timokratis Karamitros, Aris
775 Katzourakis, and Gkikas Magiorkinis. 2017. "STEAK: A Specific Tool for
776 Transposable Elements and Retrovirus Detection in High-Throughput Sequencing
777 Data." *Virus Evolution* 3 (2): vex023.
778 Song, Mingzhou, and Stéphane Boissinot. 2007. "Selection against LINE-1
779 Retrotransposons Results Principally from Their Ability to Mediate Ectopic
780 Recombination." *Gene* 390 (1-2): 206–13.
781 Stewart, Chip, Deniz Kural, Michael P. Strömberg, Jerilyn A. Walker, Miriam K. Konkel,
782 Adrian M. Stütz, Alexander E. Urban, et al. 2011. "A Comprehensive Map of Mobile
783 Element Insertion Polymorphisms in Humans." *PLoS Genetics* 7 (8): e1002236.
784 Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej
785 Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural
786 Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.
787 Thomas, Jainy, Hervé Perron, and Cédric Feschotte. 2018. "Variation in Proviral
788 Content among Human Genomes Mediated by LTR Recombination." *Mobile DNA* 9
789 (December): 36.
790 Thung, Djie Tjwan, Joep de Ligt, Lisenka E. M. Vissers, Marloes Steehouwer, Mark
791 Kroon, Petra de Vries, Eline P. Slagboom, Kai Ye, Joris A. Veltman, and Jayne Y.
792 Hehir-Kwa. 2014. "Mobster: Accurate Detection of Mobile Element Insertions in next
793 Generation Sequencing Data." *Genome Biology* 15 (10): 488.
794 Underwood, Charles J., Ian R. Henderson, and Robert A. Martienssen. 2017. "Genetic
795 and Epigenetic Variation of Transposable Elements in Arabidopsis." *Current Opinion
796 in Plant Biology* 36 (April): 135–41.
797 Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth,
798 Maido Remm, and Steven G. Rozen. 2012. "Primer3--New Capabilities and
799 Interfaces." *Nucleic Acids Research* 40 (15): e115.
800 Wallace, Amelia D., George A. Wendt, Lisa F. Barcellos, Adam J. de Smith, Kyle M.
801 Walsh, Catherine Metayer, Joseph F. Costello, Joseph L. Wiemels, and Stephen S.
802 Francis. 2018. "To ERV Is Human: A Phenotype-Wide Scan Linking Polymorphic
803 Human Endogenous Retrovirus-K Insertions to Complex Phenotypes." *Frontiers in
804 Genetics* 9 (August): 298.
805 Wang, Lu, Emily T. Norris, and I. K. Jordan. 2017. "Human Retrotransposon Insertion
806 Polymorphisms Are Associated with Health and Disease via Gene Regulatory
807 Phenotypes." *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2017.01418.
808 Wang, Lu, Lavanya Rishishwar, Leonardo Mariño-Ramírez, and I. King Jordan. 2016.
809 "Human Population-Specific Gene Expression and Transcriptional Network
810 Modification with Polymorphic Transposable Elements." *Nucleic Acids Research*.
811 https://doi.org/10.1093/nar/gkw1286.

Wang, Su, Chongzhi Zang, Tengfei Xiao, Jingyu Fan, Shenglin Mei, Qian Qin, Qiu Wu, et al. 2016. “Modeling Cis-Regulation with a Compendium of Genome-Wide Histone H3K27ac Profiles.” *Genome Research* 26 (10): 1417–29.

Watkins, W. Scott, Alan R. Rogers, Christopher T. Ostler, Steve Wooding, Michael J. Bamshad, Anna-Marie E. Brassington, Marion L. Carroll, et al. 2003. “Genetic Variation among World Populations: Inferences from 100 Alu Insertion Polymorphisms.” *Genome Research* 13 (7): 1607–18.

Wildschutte, Julia H., Alayna Baron, Nicolette M. Diroff, and Jeffrey M. Kidd. 2015. “Discovery and Characterization of Alu Repeat Sequences via Precise Local Read Assembly.” *Nucleic Acids Research* 43 (21): 10292–307.

Xing, Jinchuan, Yuhua Zhang, Kyudong Han, Abdel Halim Salem, Shurjo K. Sen, Chad D. Huff, Qiong Zhou, et al. 2009. “Mobile Elements Create Structural Variation: Analysis of a Complete Human Genome.” *Genome Research* 19 (9): 1516–26.

## TABLE AND FIGURES LEGENDS

**Figure 1: Overview of the TypeTE pipeline.** TypeTE is divided in two main scripts. The first (A) genotypes non-reference insertion (TypeTE-nonref) and the second (B) genotypes reference pMEI (TypeTE-ref). (A) TypeTE-ref creates the reference allele (REF) by extracting +/- 500 bps from the *Alu* predicted breakpoint. The alternate allele (ALT), corresponding to the pMEI presence is made by 1-2) removing the predicted TSD from the +/- 500 bps extracted sequence. Then, for each locus, read pairs (including discordant mates) are extracted from the individual bam files and are pooled for local assembly (3). If TSDs are identified in the assembly, the sequence is then inserted onto the flanking (4). In case the assembly is incomplete, the Repbase consensus for the predicted TE family is inserted instead (4). (B) The REF allele is created after extraction of +/- 500 bps from the 5’ and 3’ ends of the adjusted *Alu* position (including TSDs). The ALT allele is then created removing the *Alu* sequence and 1 TSD from the same extracted sequence. (C) Genotyping. For each locus, read-

844 pairs of each sample are extracted in a 500 bps window centered on the predicted
845 breakpoint. For each sample, these reads are then mapped to the two alleles and
846 genotype likelihood are computed.
847
848 **Figure 2: Comparison of the predicted genotypes in the 1000 GP dataset with**
849 **PCR-assays in 42 CEU individuals.** Each vertical bar represents one locus, and
850 match or error regarding the genotype for each individual are piled up on the Y axis and
851 color coded according to the legend. NA values (no genotype predicted or failed PCR)
852 are removed from the plot. >
853
854 **Figure 3: Comparison of the predicted genotypes in the SGDP dataset with PCR-**
855 **assays in 14 South Asian individuals.** Each vertical bar represents one locus, and
856 match or error regarding the genotype for each individual are piled up on the Y axis and
857 color coded according to the legend. NA values (no genotype predicted or failed PCR)
858 are removed from the plot. >
859
860 **Figure 4: Average error rate per locus across methods and datasets.** Different
861 letters indicate significant difference. Tukey's HSD, $P < 0.05$; NS: Not significant>
862
863 **FIGURE 5: Effect of method and dataset on variant discovery performance.**
864 Sensitivity, precision and F1 score are compared for each dataset (1000 GP and
865 SGDP) according to the type of insertion (non-reference vs reference) and the
866 genotyping method used (1000 GP, MELT2.1.4 and TypeTE). Error bars: 95%

867 confidence interval. Non-overlapping intervals denotes a significant difference between
868 scores.
869
870 **FIGURE 6: Per locus inbreeding coefficient (Fis).** The Fis is estimated for each locus
871 using the alleles frequencies given by each method (1000 GP: original 1000 GP
872 genotypes, MELT2, TypeTE and PCR assays) and for each of the 1000 GP (n = 42
873 individuals) and SGDP (n = 9 individuals) datasets. Red dashed-line: expected Fis at
874 Hardy-Weinberg equilibrium (Fis = 0).

input 1:

**input.vcf**
breakpoints,
TSD,
strand

**TypeTE *non-reference***

A

1 ±500bp REF → 2 REF

1 REF → 2 ±500bp → 3 Local de-novo assembly of TE insertion → 4 ALT

pool reads from individual bams

*input 2:*

indX.bam
ind2.bam
ind1.bam

*Use MELT MEI strand and RepBase consensus if assembly fails*

**TypeTE *reference***

B

1 ±500bp REF → 2 REF

1 ±500bp REF → Adjust TE and TSDs coordinates → 2 ALT

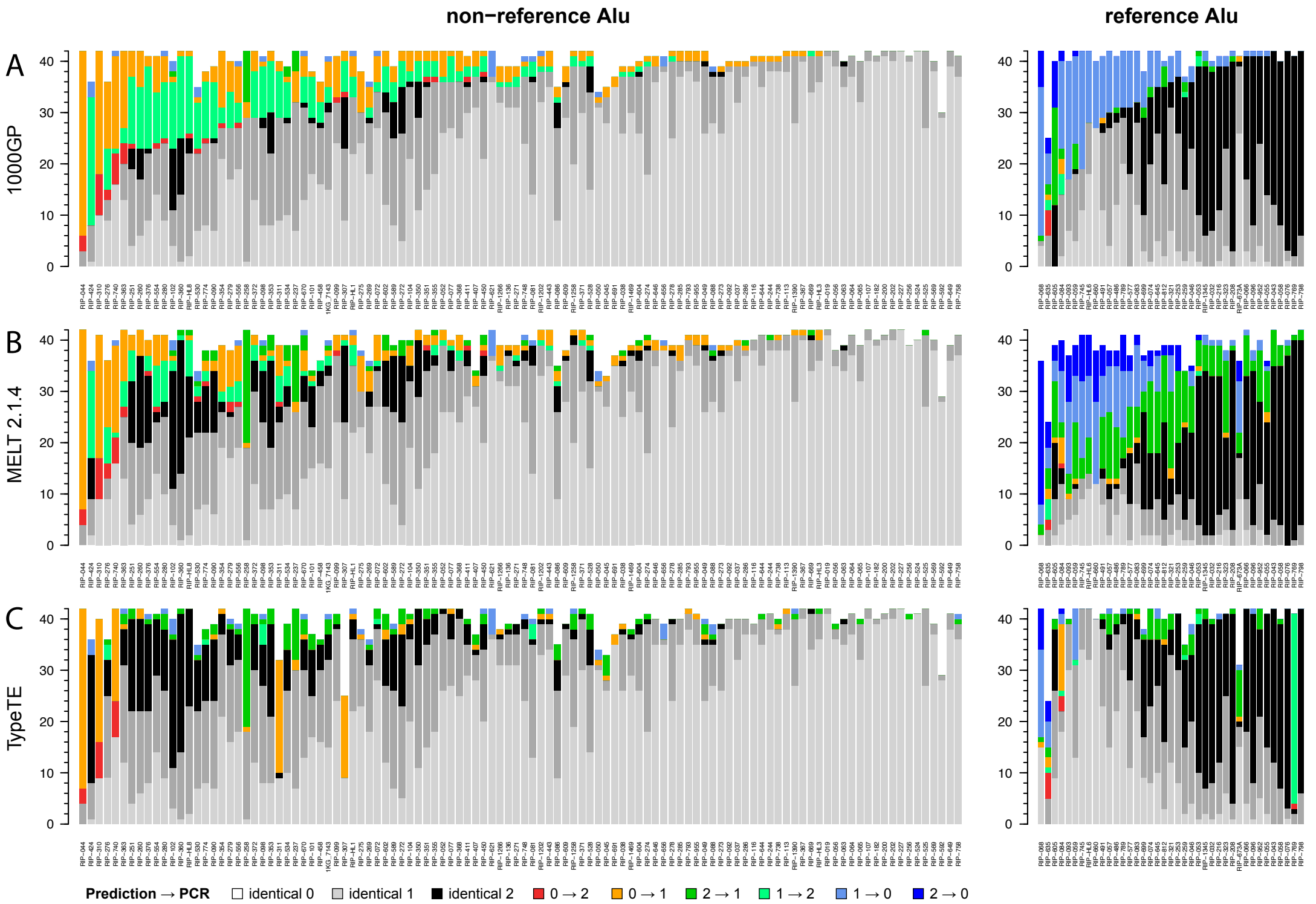
C

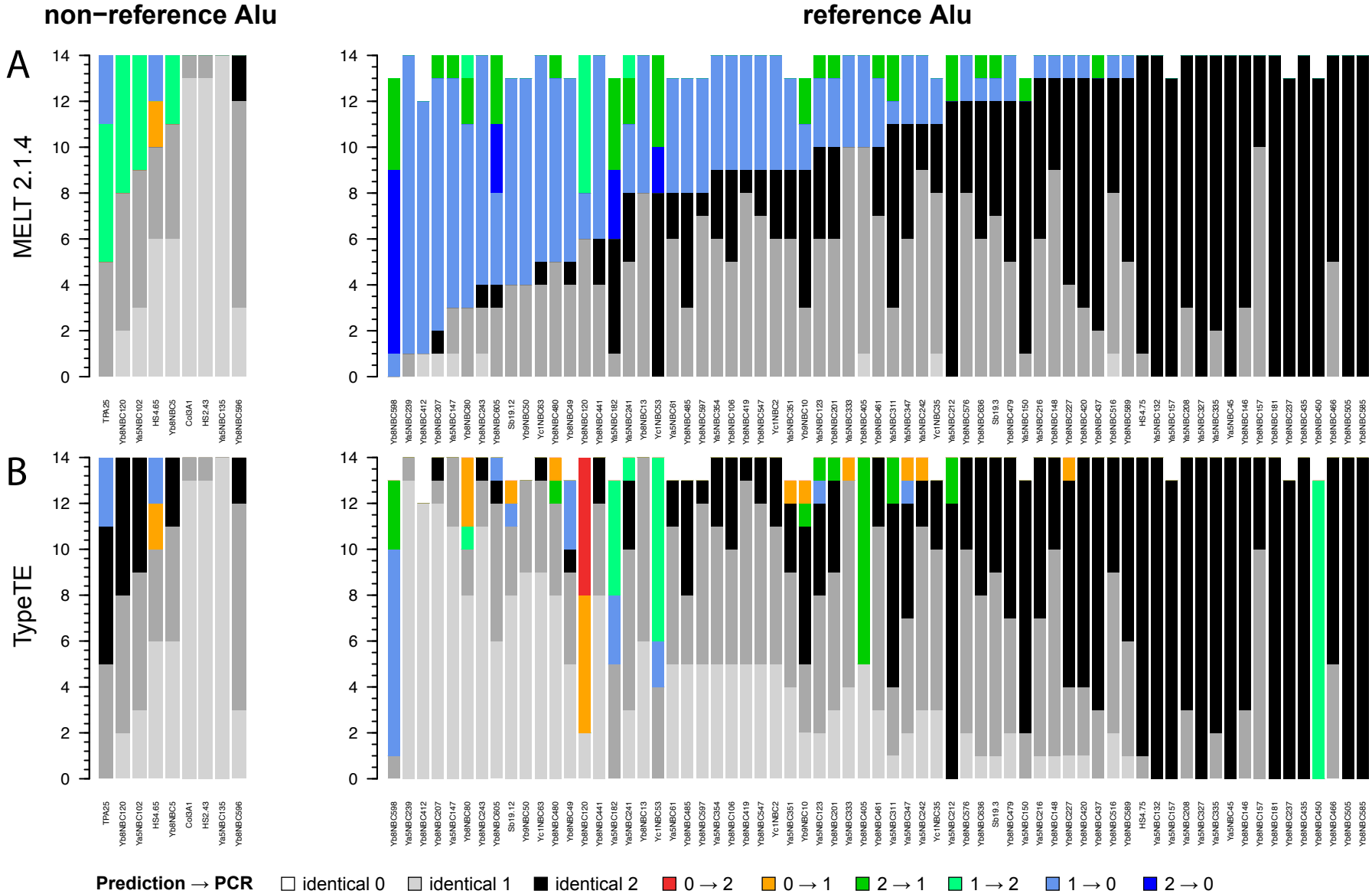
**Genotyping**
(Wildschutte et al. 2015)

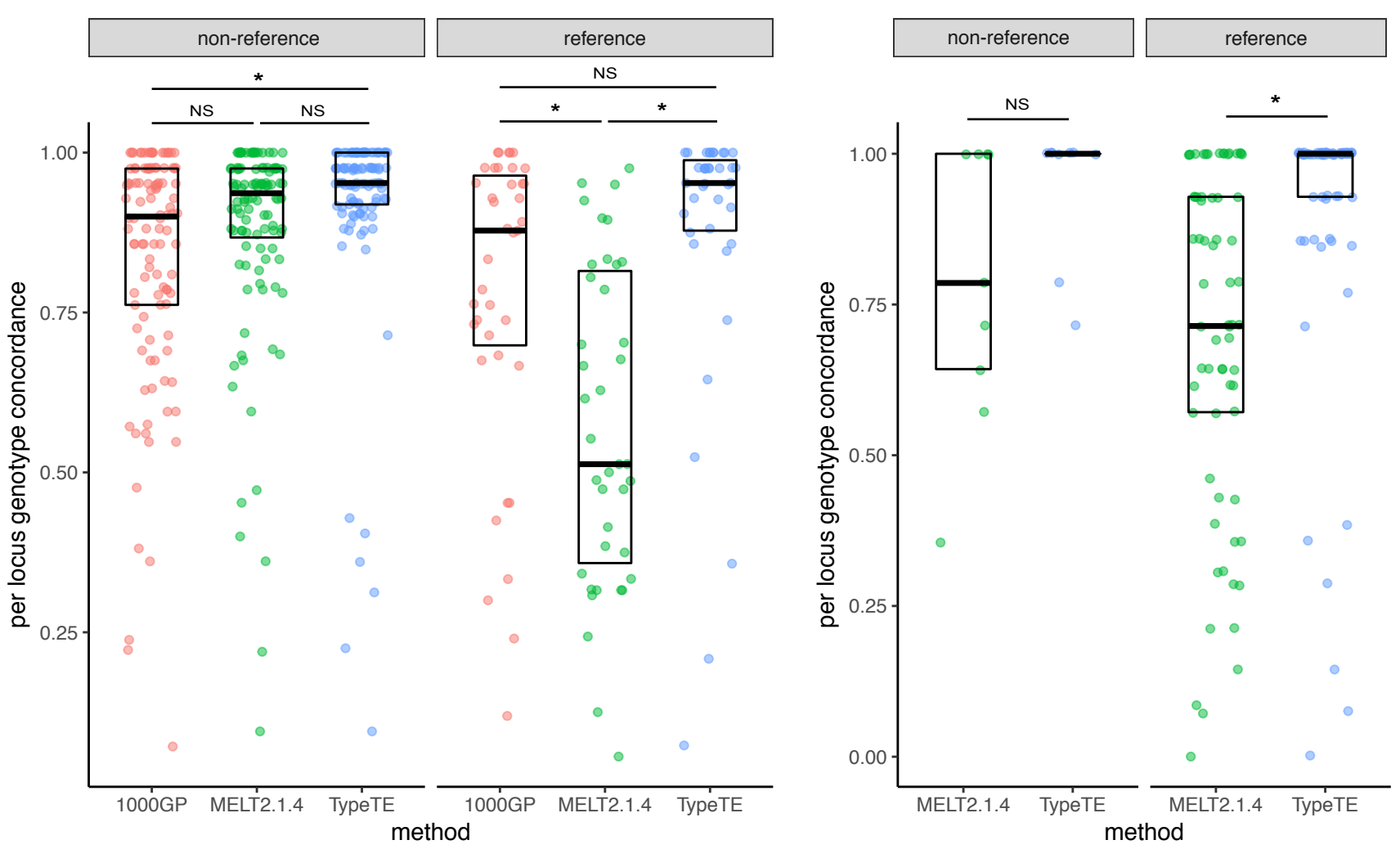
1/0 1/1 0/0

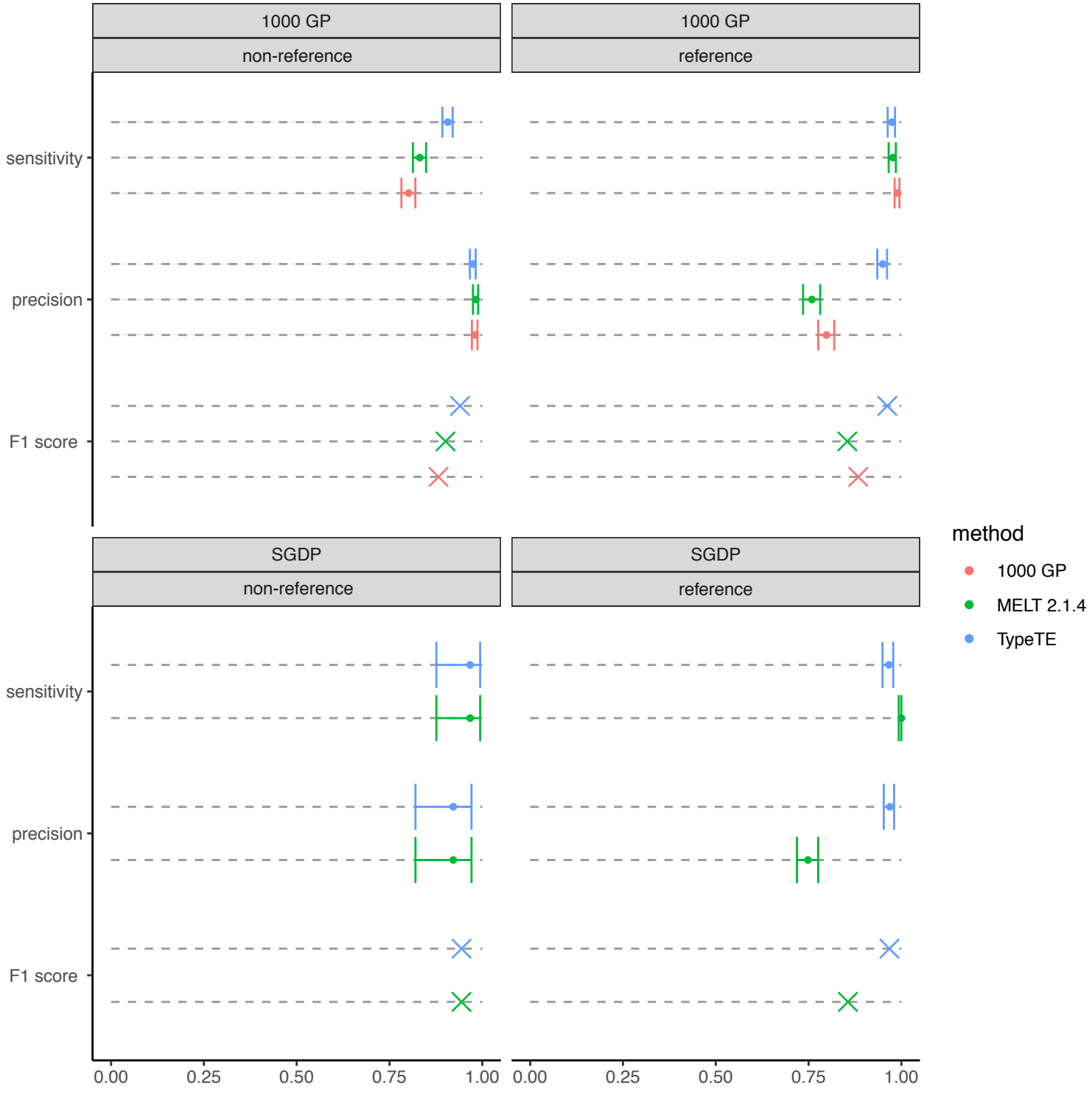
**expected *Alu* insertion**

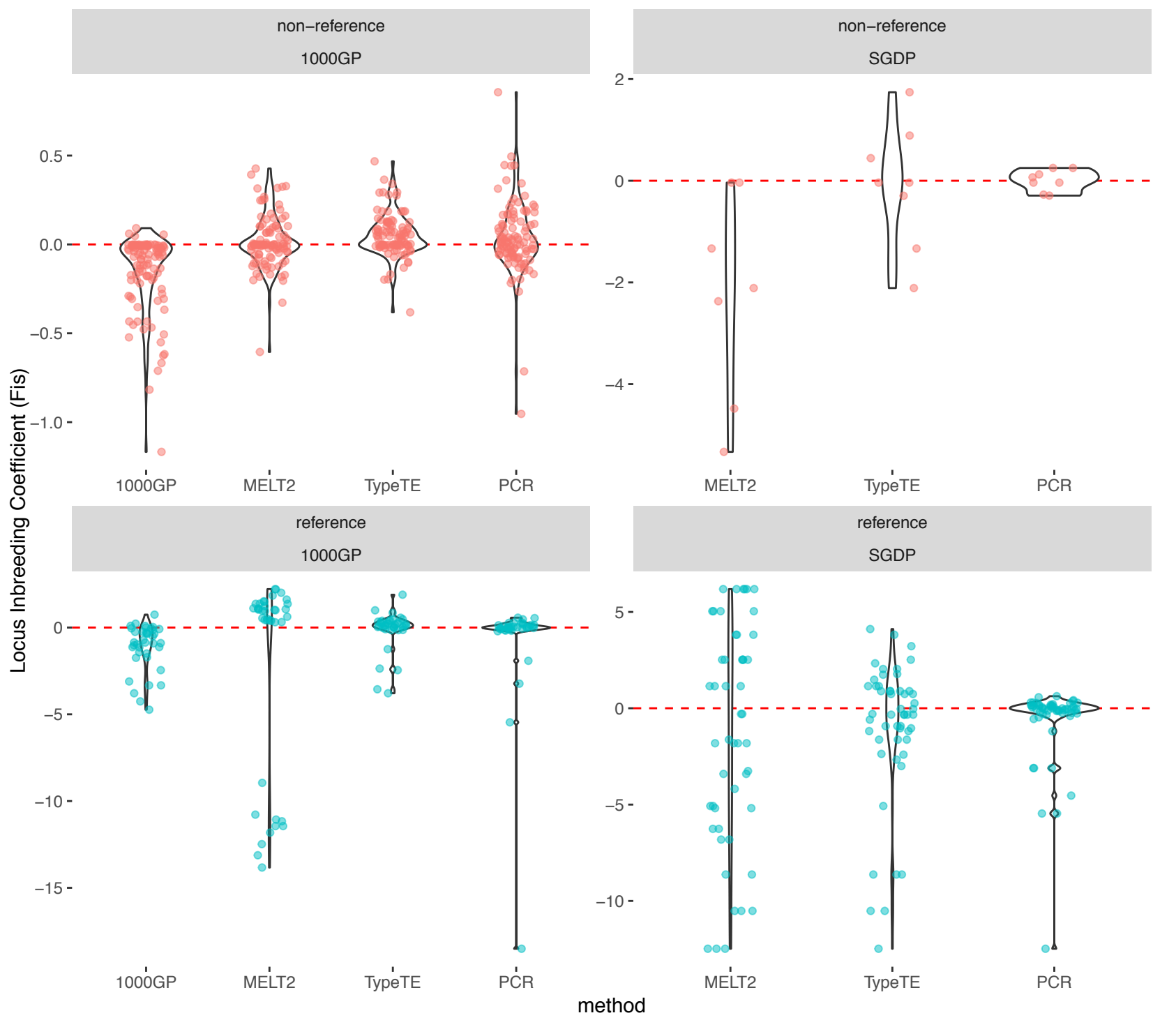
flanking | TSD | ALU | $AAA_n$ | TSD


non−reference Alu
reference Alu
A
1000GP
B
MELT 2.1.4
C
TypeTE
0
10
20
30
40
Prediction → PCR
identical 0
identical 1
identical 2
0 → 2
0 → 1
2 → 1
1 → 2
1 → 0
2 → 0


non−reference Alu
reference Alu
A
MELT 2.1.4
B
TypeTE
0
2
4
6
8
10
12
14
TPA25
Yb8NBC120
Ya5NBC102
HS4.65
Yb8NBC5
Col3A1
HS2.43
Ya5NBC135
Yb8NBC596
Yb8NBC598
Ya5NBC239
Yb8NBC412
Yb8NBC207
Yb8NBC147
Yb8NBC80
Yb8NBC243
Yb8NBC605
Sb19.12
Yb9NBC50
Yc1NBC83
Yb8NBC480
Yb8NBC49
Yb8NBC120
Yb8NBC441
Ya5NBC182
Ya5NBC241
Yb8NBC13
Yc1NBC53
Ya5NBC61
Yb8NBC485
Yb8NBC597
Ya5NBC354
Yb8NBC106
Yb8NBC419
Yb8NBC547
Yc1NBC2
Ya5NBC351
Yb9NBC10
Ya5NBC123
Yb8NBC201
Ya5NBC333
Yb8NBC405
Yb8NBC461
Ya5NBC311
Ya5NBC347
Ya5NBC242
Yc1NBC35
Ya5NBC212
Yb8NBC576
Yb8NBC636
Sb19.3
Yb8NBC479
Ya5NBC150
Ya5NBC216
Yb8NBC148
Yb8NBC227
Yb8NBC420
Yb8NBC437
Yb8NBC516
Yb8NBC589
HS4.75
Ya5NBC132
Ya5NBC157
Ya5NBC208
Ya5NBC327
Ya5NBC335
Ya5NBC45
Yb8NBC146
Yb8NBC157
Yb8NBC181
Yb8NBC237
Yb8NBC435
Yb8NBC450
Yb8NBC466
Yb8NBC505
Yb8NBC585
Prediction → PCR
identical 0
identical 1
identical 2
0 → 2
0 → 1
2 → 1
1 → 2
1 → 0
2 → 0


non-reference
reference
non-reference
reference
*
NS
NS
NS
*
*
NS
*
per locus genotype concordance
1.00
0.75
0.50
0.25
0.00
1000GP
MELT2.1.4
TypeTE
1000GP
MELT2.1.4
TypeTE
MELT2.1.4
TypeTE
MELT2.1.4
TypeTE
method
method


1000 GP
non-reference
1000 GP
reference
SGDP
non-reference
SGDP
reference
sensitivity
precision
F1 score
0.00
0.25
0.50
0.75
1.00
method
1000 GP
MELT 2.1.4
TypeTE


non−reference
1000GP
non−reference
SGDP
reference
1000GP
reference
SGDP
Locus Inbreeding Coefficient (Fis)
0.5
0.0
−0.5
−1.0
2
0
−2
−4
5
−5
−10
−15
1000GP
MELT2
TypeTE
PCR
method