GraffiTE: a Unified Framework to Analyze Transposable Element Insertion Polymorphisms using Genome-graphs

Cristian Groza[1], Xun Chen[2], Travis J. Wheeler[3], Guillaume Bourque[2,4,5,6] and Clément Goubert[6]*

[1]Quantitative Life Sciences, McGill University, Montréal, QC, Canada
[2]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan
[3]R. Ken Coit College of Pharmacy, University of Arizona, Tucson, AZ, USA
[4]Canadian Centre for Computational Genomics, McGill University, Montréal, QC, Canada
[5]Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC, Canada
[6]Human Genetics, McGill University, Montréal, QC, Canada
***Corresponding author:** goubert.clement@gmail.com

Abstract:

Transposable Elements (TEs) are abundant and mobile repetitive DNA sequences evolving within and across their hosts' genomes. Active TEs cause insertion polymorphism and contribute to genomic diversity. Here, we present GraffiTE, a flexible and comprehensive pipeline for detecting and genotyping polymorphic mobile elements (pMEs). By integrating state-of-the-art SV detection algorithms and graph-genome frameworks, GraffiTE enables the accurate identification of pMEs from genomic assemblies and long-read as well as the precise genotyping of these variants using short- or long-read data. Performance evaluations using simulated and benchmark datasets demonstrate high precision and recall rates. Notably, we demonstrate the versatility of GraffiTE by analyzing the human reference pangenome, 30 *Drosophila melanogaster* genomes, and multiple cultivars of the emerging crop model *Cannabis sativa*, where pMEs are undocumented. These analyses reveal the landscapes of pMEs and their frequency variations across individuals, strains, and cultivars. GraffiTE provides a user-friendly interface, allowing non-expert users to perform comprehensive pME analyses, including in models with limited TE prior knowledge. The pipeline's extensible design and compatibility with various sequencing technologies make it a valuable integrative framework for studying TE dynamics and their impact on genome evolution. GraffiTE is freely available at https://github.com/cgroza/GraffiTE.

Keywords: Transposable elements, insertion polymorphism, graph-genomes, structural variation, genotyping, genome evolution

Introduction

Transposable Elements (TEs) represent a heterogeneous collection of repetitive sequences found in virtually all eukaryotic species[1,2]. A unifying characteristic of TEs is their initial ability to multiply and colonize new loci within their host genome through transposition[3]. Active TE families (sets of copies descending from the same source sequence) are often quickly silenced through the rapid evolution of host defense mechanisms[4]. However, transposition of evolutionarily young TE families acts as a major source of structural variants (SVs) in a multitude of species[5]. As an example, and with only three TE families currently active, the human population harbors tens of thousands of polymorphic loci (hereinafter referred as polymorphic mobile elements, or pMEs) differentially present or absent between individuals[6,7]. This number is even higher in *Drosophila melanogaster*, whose genome hosts many fewer TEs than the human genome (~20 and ~50%, respectively) but possesses a significantly higher proportion of families remaining active[8].

With the advancement of sequencing technologies, and fostered by the intricate involvement of TEs with a wide array of biological processes[2], the study of pMEs has risen in popularity. Indeed, as segregating variants, pMEs can be used as genetic markers, either to describe the population structure or to search for selection signatures[9–13]. Furthermore, the mutagenic effect of new insertions, coupled with the dissemination of their own regulatory and coding sequences among genomes, has been shown to affect regulatory networks and hosts' phenotypes from diseases to adaptation[14–18]. In effect, pMEs are a significant contributor to genomic variation, constantly delivering raw material for evolution.

The detection and genotyping of pMEs has proven to be a challenging task, due to their repetitive nature and variable size (from hundreds to thousands of bp)[19]. Nevertheless, a plethora of methods have been published, in particular targeting short-read paired-end datasets[6,20–27]. With the increased availability of long-read sequencing, new tools have emerged to detect and genotype SVs, including pMEs, with increased sensitivity[28–32]. Indeed, long reads are able to capture full-length pME sequences in context, i.e. with enough flanking sequence to anchor them with confidence in a reference genome. With long-read sequencing growing in popularity and affordability, an increasing number of projects are producing multiple high-quality, chromosome-level assemblies for a given species, referred to as pangenomes[33–37].

Pangenomics, the integrative analysis of multiple related genomes simultaneously, has become the gold-standard in comparative genomics[38]. This new paradigm allows for a better representation of the genetic diversity in a given model, by including genomic variants previously ignored or missed, due to a lack of information stored in a single reference genome. As a consequence, pangenomes can dramatically enhance the estimation of structural variants' (SVs) frequencies, which is for example extremely relevant in the context of rare diseases diagnostic[39], agriculture[40,41], and association studies in general[42].

Among the tools developed for pangenome analysis, graph-genomes are flexible data structures enabling representation of genomic variation, from single nucleotide polymorphisms (SNVs) to large structural variants (SVs), in a single graph. In graph-genomes, variants are represented by bubbles of divergent sequences (or nucleotides) and shared segments are collapsed into a single node[43]. Graph-genomes allow high-precision mapping of genomic, transcriptomic[44] and epigenomic[45,46] variants, enabling research in evolutionary and functional genomics.

Existing pME methods that can apply to pangenome projects may be restricted to the detection of non-reference variants (i.e. TE absent from a reference genome, for example TLDR and TELR[28,29]), or specialized to a specific type of TE or organism[31]. Furthermore, direct detection of pMEs from alternative genomic assemblies is rare (though this feature – restricted to a single alternative assembly at a time – exists in TrEMOLO[30]). Finally, to our knowledge, none of the existing methods offer a graph-genome framework to analyze pMEs and their impacts in a reliable yet flexible way.

To fill this gap, we have created GraffiTE (pronounced "gruh·FEE·tee", like the popular street art form, *graffiti*). GraffiTE is a general-purpose pipeline for calling and genotyping pMEs, applicable to any model for which a list of consensus TE sequences of interest is available. GraffiTE makes available different state-of-the-art methods to detect and report pMEs from genomic assemblies or long-reads, and performs graph genotyping using short or long read datasets. The GraffiTE workflow is managed using the Nextflow framework[47]. GraffiTE was built with non-expert users in mind: a complete analysis of multiple genomic samples can be performed in a single command, while dependencies are directly available through containerization. Furthermore, variant handling and comparison and sharing is facilitated through the use of standardized

variant call format (VCF). Finally, advanced configuration and optimisation are available for high performance clusters or cloud deployment.

After a brief presentation of the method, we will present the results of a synthetic and real data benchmark to demonstrate the performance of the tool. Following that, we show the versatility of the tool by searching for pMEs using 47 diploid human genome assemblies, 30 haploid *Drosophila melanogaster* assemblies alongside their original long-read sets, and finally, demonstrate how the tool can be applied to a species with limited knowledge about TEs, the emerging agricultural model *Cannabis sativa*.

Results

A unified framework to study TE insertion polymorphisms

GraffiTE organizes a complex collection of SVs and pangenome analysis software in order to identify, extract, and analyze genomic variants likely to represent mobile element insertion polymorphisms (pMEs). The pipeline is implemented in the Nextflow framework, and relies on an Apptainer container (formerly known as Singularity[48]) to provide all the required dependencies.

In brief, the pipeline is divided into three main steps (Figure 1). First, SVs are searched between either (i) alternative assemblies and a reference genome or (ii) long-read sets and a reference genome (both searches can be also combined). Second, SVs found in individual samples are merged and further filtered to retain pMEs. Alternatively, a user-provided VCF file reporting both reference and alternative alleles sequences for each variant can be used as input. Finally, each pME detected can be further genotyped by mapping short or long reads against a TE graph-genome. This graph-genome represents each identified pME as a bubble, i.e. providing alternate paths in the graph, where both presence and absence alleles are available for read mapping and genotyping. The first two steps focus on pME detection, i.e. establishing whether a given TE copy is present or absent in a sample. The third step, called pME genotyping, is aimed to report whether each pME detected is homozygous or heterozygous.

The ability to handle and combine heterogeneous data types (genome assemblies, long- and short- read sets) is a unique aspect of the workflow, and addresses a common scenario of many research projects. For instance, one may use a limited number of genome assemblies or long read sets to establish a high-quality catalog of pME, while genotyping them later at population scale using short-reads. In large pangenome projects, assembled genomes are often more convenient to use than raw read sets due to storage limitations. GraffiTE can be directly applied to hundreds of assemblies, and pME detection can be performed without additional data. Another use of GraffiTE can be to annotate and retain pME from a pre-established collection of SVs (stored in a VCF file, see Methods), with the possibility to further perform genotyping on it. Finally, a complete analysis, including pME detection and genotyping can be performed using long-reads set only.

The modular aspect of Nextflow makes it easy to swap or bypass components of the pipeline, and thus will allow seamless upgrade and/or modification with new softwares and routines as needed (Table 1 and Figure S1). The program and a detailed documentation (including installation, usage and output description) is available online at https://github.com/cgroza/GraffiTE

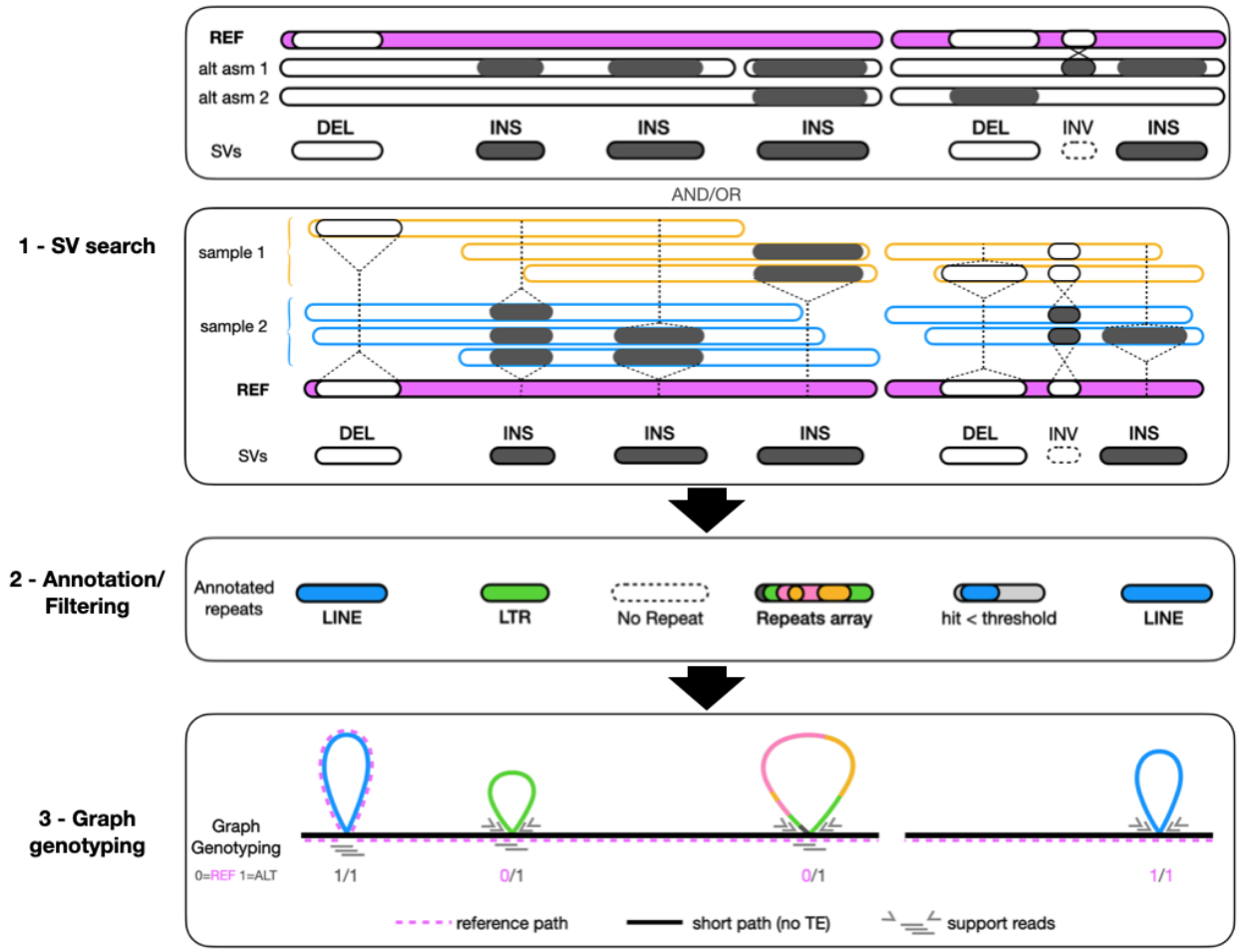

Figure 1. GraffiTE workflow. The pipeline is divided in three main steps: **1- SV search.** Alternative assemblies ("alt asm") and/or long-read sets (sample 1, sample 2) provided in input are pseudo-aligned to a reference genome (represented in pink) using minimap2. Subsequently, SVs are searched with Svim-asm (from assemblies) or Sniffles 2 (long-read sets) and reported in a VCF file. Insertions (dark gray) and deletion (white) are then retained while other types of SVs (such as inversions) are discarded. **2- Annotation/Filtering.** SVs found in individual assemblies or read-sets and representing the same loci are merged with the software SURVIVOR, and further filtered to retain likely pME using RepeatMasker and a user-provided repeat library. **3- Graph genotyping.** Based on annotation thresholds, likely pMEs are retained and a graph-genome representing each of them as a bubble is generated (TE-graph-genome). The reference genome path in the graph (pink dashed line) can either skip (insertions) or include (deletions) a given pME. Finally, short- or long-read sets can be used to genotype each pME using graph aligners bundled in GraffiTE. Additional details are provided in Figure S1.

Table 1. Software and methods implemented in GraffiTE. GraffiTE steps are labeled according to Figure 1.

software	version	purpose	GraffiTE step	reference
minimap2	2.24-r1155-dirty	assembly or long-read mapping over a reference genome	1 - SV search	[49]
SVIM-asm	1.0.3	SV calling from assembly-to-genome alignments	1 - SV search	[50]
Sniffles2	2.0.7	SV calling from long reads-to-genome alignments	1 - SV search	[51]
samtools	1.16.1-49-geb703e0	alignment file sorting	1 - SV search	[52]
SURVIVOR	1.0.7	SV merging from individual calls	1 - SV search	[53]
RepeatMasker	4.1.4	TE annotation	2 - Annotation / Filtering	[54]
OneCodeToFindThemAll	1 (adapted)	TE annotation parsing	2 - Annotation / Filtering	[55]
blastn	2.9.0+	TSD search module	2 - Annotation / Filtering	[56]
bedtools	2.27.1	TSD search module, TE annotation parsing	2 - Annotation / Filtering	[57]
Pangenie	2.1.0	short-read mapping to TE pangenome graph	3 - Graph genotyping	[58]
vg	1.45.0 "Alpicella"	graph induction for Giraffe and GraphAligner	3 - Graph genotyping	[59]
Giraffe	(part of vg)	short-read mapping to TE pangenome graph	3 - Graph genotyping	[60]
GraphAligner	bioconda 1.0.13	long-read mapping to TE pangenome graph	3 - Graph genotyping	[61]
bcftools	1.16-81-g8f54545	vcf file handling and annotation	other	[62]
tabix	1.16-40-g114f5eb	vcf file handling and annotation	other	[63]

Simulations demonstrate effective pME detection from different data sources.

The first step in the GraffiTE pipeline performs SV search between (i) alternative assemblies and reference genome, (ii) long-read sets and a reference genome or (iii) both (Figure 1, “1-SV search”). This SV search is followed by annotation of the variants’ sequences and filtering to retain likely pME. These analyses produce a VCF file, unified across all samples, listing putative pME presence/absence among samples and their annotations. To test the ability of the algorithms implemented in GraffiTE to detect pMEs from genomic assemblies and long-read sets, we simulated insertions and deletions of known active families (Alu, LINE-1 and SVA elements) in the human genome chromosome 22 (see Methods and Supplementary Methods 1.a.). We introduced background noise in the simulated pMEs by generating random non-pME SVs (insertions or deletions) based on the length distribution of those observed in the Genome In a Bottle Benchmark set[64]. In all the combinations of tools and data type tested (Figure 2 A, top), F1 scores computed from the raw outputs averaged 0.870±0.017 s.d., mainly driven by a high recall (mean: 0.961±0.034 s.d.). We observed that the lowest scores were recorded when the SV search was performed with Sniffles 2 on simulated PacBIO HiFi reads (GT-sn [hifi]). This is unexpected since the HiFi reads are usually more accurate than the ONT reads but is

probably due to the fact that the simulated HiFi reads were shorter on average than the simulated ONT reads, which is also true of real datasets. While read length was taken into account when controlling for coverage, as expected, it still has a strong influence on accuracy of SV detection compared to accuracy. Nevertheless, we observed that the specificity and precision can be increased using simple filters: the first one retains only variants with a single hit to a known TE consensus, which increases both specificity and precision above 80% (Figure 2, middle). The second filter adds a minimum length of 250 bp for a pME to be retained, and allows all combinations of tools and data type to reach a F1 score above 90% (Figure 2, bottom).

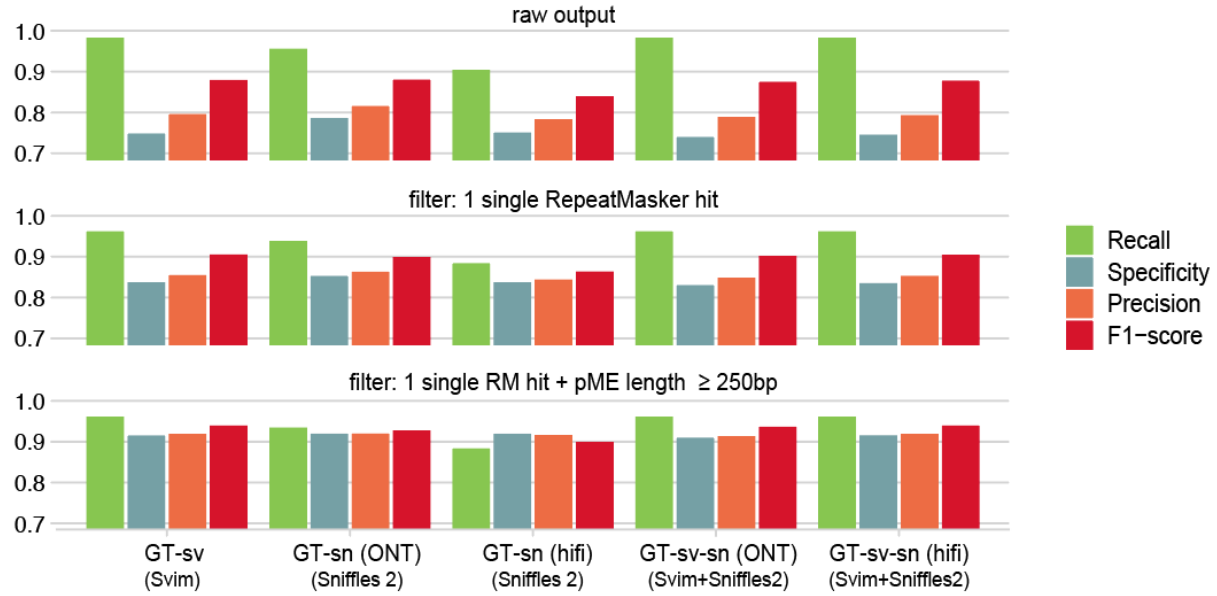

Figure 2. pME detection on simulated human datasets. Recall, Specificity, Precision and F-1 scores for pME detection in a total of 100 simulated human chromosome 22. “all”: unfiltered output VCF from GraffiTE, “1 hit”: pMEs annotated with a single RepeatMasker hit >= 80% of the SV sequence. “1 hit + pMEs >= 250bp”: identical to “1 hit” with pMEs < 250bp filtered out. GraffiTE modes: “GT-sv”: GraffiTE-Svim (SV search from assemblies), “GT-sn”: GraffiTE-Sniffles 2 (from long-read sets), “GT-sv-sn” = GraffiTE Svim-Sniffles 2 (pME search from assemblies and long-read sets, see also Table S1).

GraffiTE maintains high performance when benchmarked against real data

The main limitation of the simulations is that the alternative assemblies (simulated chromosome 22) are identical to the ground truth, which is unrealistic. Thus, we further tested the relative performance of the tools implemented in GraffiTE for pME detection using a benchmark set derived from[64] (see Methods). The set includes high-quality pME variants for Alu, LINE1 and

SVA elements reported for HG002. Additionally, we applied TLDR[28], a method designed to find non-reference pME from long-reads, to the long-read sets generated for HG002, and applied MELT2[6] and MEGAnE[65] to short Illumina read-sets. The best performances were obtained with GraffiTE on long-read sets (GT-sn) or a combination of long-read sets and the diploid assembly of HG002 (GT-sv-sn, Figure 3 A). Using only the genomic assemblies with GraffiTE to detect pME yielded high precision (> 94%) with, however, a sharp decline in recall (< 50%), indicating that assemblies' completion remains a limiting factor for direct pME detection. Compared to TLDR, which searches for pME insertions in long-read data, detection from long reads yielded better performance with GraffiTE, which suggests that Sniffles 2 provides significant improvement for both recall and specificity. Non-surprisingly, short-reads methods for pME detection had the lowest performances, both in terms of recall and precision, as expected when SVs are searched from short-read alignments.

Beyond the detection of pMEs, GraffiTE supports graph-genotyping, i.e. inferring the allele composition for each pME previously detected by mapping short or long reads against a graph genome. We thus tested the performances of Pangenie[58] (suited for short-reads) and GraphAligner[61] (long-reads) as implemented within GraffiTE. Using the same benchmark set a previously, we focus here on the correspondence between the bi-alleleic genotypes provided by the caller and the GIAB genotypes, thus separating heterozygous from homozygous genotypes. We performed graph-genotyping using short or long reads, using the TE-graph-genomes created with either GT-sv (assemblies only), GT-sn (long-read only), or GT-sv-sn (both) (Figure 3 B). We set a fixed read set coverage of 30X for all the tests, as competing methods based on short-reads require to perform both pME detection and genotyping with the same data (MELT2, MEGaNE). The best performances, with precision and recall > 95% were obtained using long-reads and GraphAligner with GraffiTE. As expected, methods using short-reads for detection and genotyping had globally lower performances, though MEGAnE for reference pMEs (DEL), and MELT2 for non-reference pMEs (INS) had a higher recall (MEGAnE = 0.79, GraffiTE = 0.65, MELT2 = 0.68, GraffiTE = 0.66), but lower precision (MELT2 and MEGAnE < 0.91), than GraffiTE when pMEs are initially detected with long-reads and later genotyped with short reads (GT-sn-PG and GT-sv-sn-PG). Similarly to pME detection, genotyping based on GT-sv (detection from assemblies only) had high precision but lower recall, as fewer pMEs could be initially imputed in the TE-graph-genome starting from assemblies only.

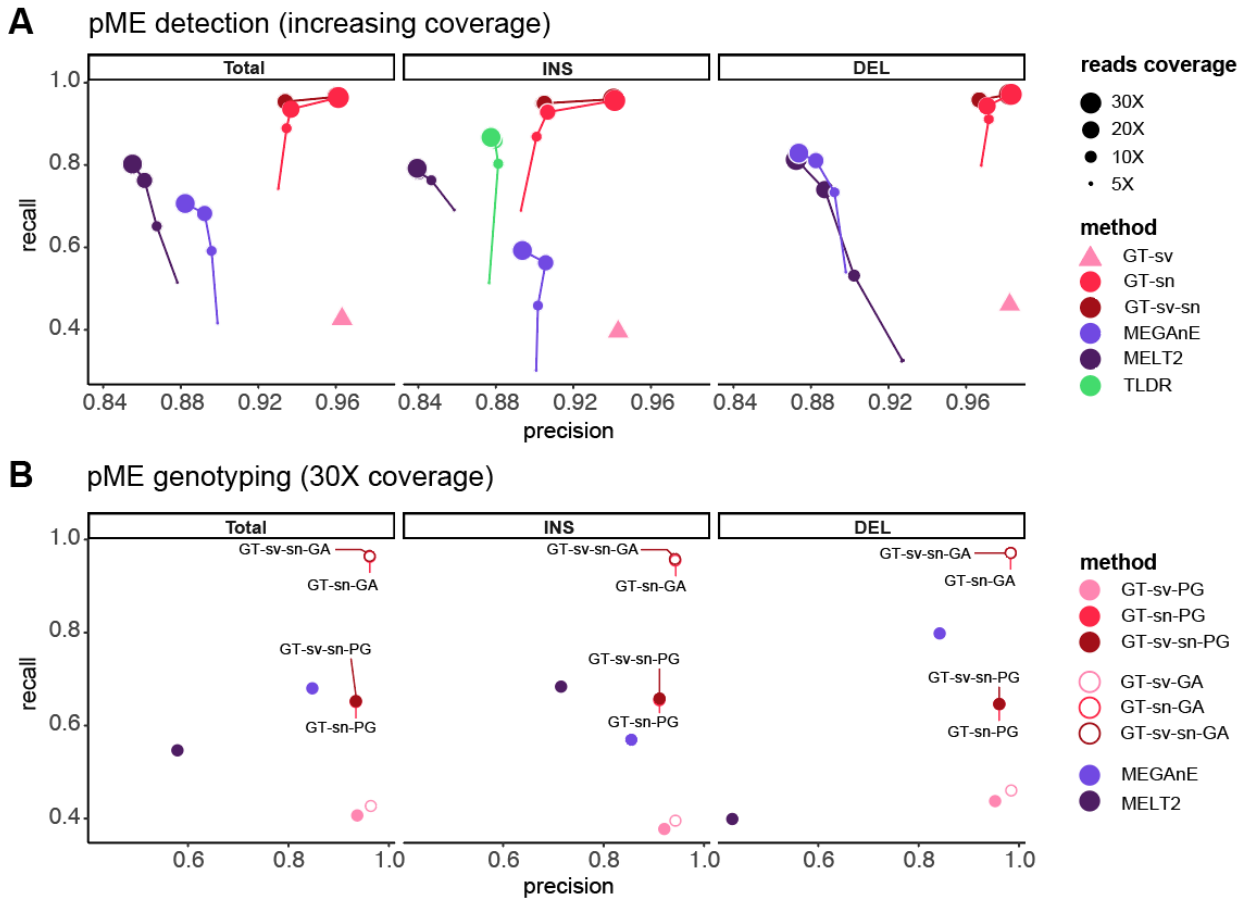

Figure 3 pME detection and genotyping benchmark against HG002/Genome in a Bottle dataset. A. Evaluation of pME detection performance and comparison with existing tools using a benchmark set derived from the GIAB analysis of HG002/NA24385. Methods legend and acronyms according to Table S1. GT-sv (pink triangle) has no coverage information for pME search, since the variants were directly detected from genomic assemblies of the maternal and paternal chromosomes. Connecting dots indicate runs with increased read coverage. Note that the software TLDR only reports non-reference (INS) pMEs. **B.** Evaluation of pME genotyping performances and comparison with short-read linear-mapping methods using 30X reads coverage. The suffix -PG denotes runs genotyped with Pangenie (short Illumina reads), and -GA GraphAligner (long, PacBio HiFi reads). Methods legend and acronyms according to Table S1.

GraffiTE provides a comprehensive picture of pMEs segregating in the human pangenome

As pangenomes are expected to grow larger in sample representation, the storage and handling of hundreds of raw read sets can represent a computational and financial challenge. With the improvement of both sequencing and assembly methods, it is becoming realistic to perform

population-wide study of genetic variation only using assemblies. To illustrate this first use case, we applied GraffiTE to the collection of 47 diploid assemblies recently published by the Human Pangenome Reference Consortium[34] (HPRC), using the hg38 reference. After the first two steps of GraffiTE (1-SV search, TE filtering), and among 94 haplotypes, a total of 14,838 distinct pME loci were identified, including 11,950 Alu; 2,413 LINE-1 and 475 SVA loci. GraffiTE also reported that 18.7% (451/2413) LINE-1 pMEs bear a signature of 5' inversion, compatible with twin priming or similar mechanisms[66]. Per individual (diploid genome), the average number of pMEs identified was 3199±322s.d. (2654±274s.d. Alu, 455±43s.d. L1 [incl. 74±7s.d. 5' inverted] and 90±10s.d. SVA). These estimates are higher than previously reported (2500 to 3000 pMEs per diploid genome) in a recent survey using 30X short-reads datasets from 3,739 individuals from the 1000 Genome Project and BioBank Japan[65]. Additionally, we detected a total of 3100 VNTR polymorphisms located within fixed SVA elements. Using Alu, LINE-1, and SVA pMEs, we show that between 20 and 30 haplotypes are sufficient to recapitulate common segregating variants (haploid genomes frequency > 5%) in the HPRC pangenome (Figure 4 A). Principal Coordinate Analysis (PCoA) using presence/absence call per haplotype recapitulates the characteristic genetic structure of the human population (Figure 4 B). Furthermore, the count of pMEs per subpopulation meets the expectation of higher genetic diversity harbored by individuals of African ancestry (Figure 4 C).

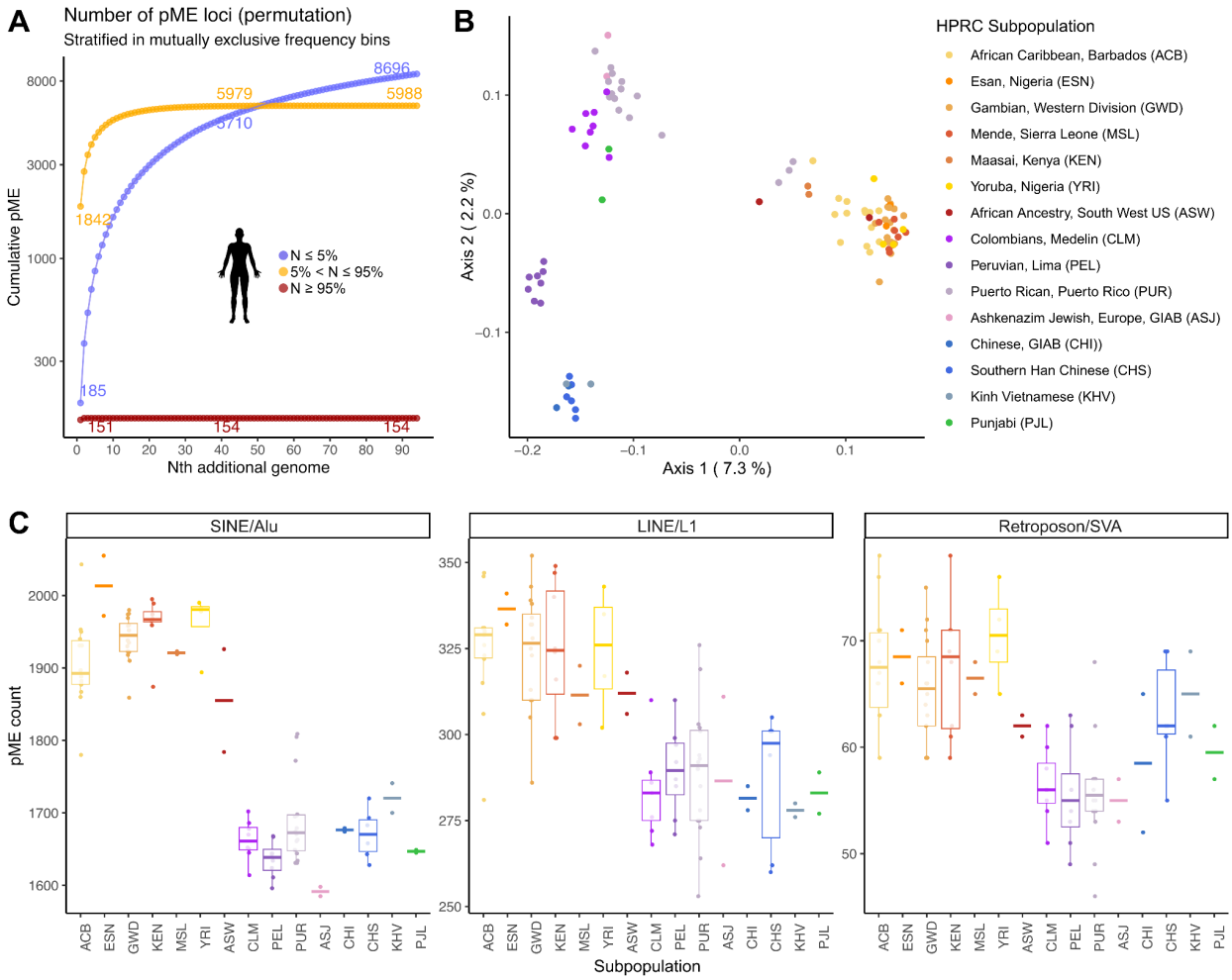

Figure 4. HPRC pME pangenome analysis. A. Discovery curves for pMEs detected with GraffiTE using 47 diploid genome assemblies (GT-sv mode) against the reference genome hg38. To reduce putative false-positives, only pMEs with a single hit on Alu, LINE-1 or SVA and with a size >= 250bp were kept, based on the simulations and benchmark results. For each mutually exclusive frequency bin, the average number of pMEs in N genomes is reported after performing 100 permutations and sampling of the genomes. **B.** PCoA representing the 94 haploid assemblies of the HPRC pangenome using segregating Alu, LINE-1 and SVA insertion polymorphisms. **C.** Count of pMEs for each haploid assembly, broken-down per TE type (Alu, LINE-1, SVA) and subpopulation. Subpopulation codes are referenced according to Figure 4 B.

Analysis of the *Drosophila melanogaster* pangenome confirms a highly dynamic repeatome

D. melanogaster displays a vastly different context for pME detection compared to humans, as the overall genomic TE content is rather small (~20% of the genome) but a much greater number of TE families are mobile, and many do so in a population specific manner[11]. We applied

GraffiTE to a subset of 30 *Drosophila melanogaster* genomes produced and analyzed by Rech et al.[8]. These genomes encompass an identical number of strains originating from 12 distinct geographic locations, chosen based on their shared utilization of ONT long-read sequences. Genome assemblies and raw, long-read sets, were used as input for GraffiTE (GT-sv-sn mode) and graph genotyping was performed with Graphaligner on the long-read (GT-sv-sn-GA mode). This combination of tools showcases an analysis taking advantage of all the data available, and in addition to the analysis of the human pangenome, provides bi-allelic genotype for each pME in each strain represented in the pangenome. The output VCF was further filtered to retain variants with a single hit on the authors' manually curated TE library (MCTE), as our simulations show that this simple filtering improves the method's specificity (Figure 3A). Because of fundamental differences in the reporting of events between GraffiTE and [8] (Supplementary Methods 2.a), we focused on expanding the pME search and genotyping to heterochromatic regions, which were not originally analyzed. After graph-genotyping using ONT long reads, GraffiTE reported a total of 20,353 pME, including 15,250 euchromatic and 5,103 heterochromatic variants (Figure 5, A-B). Discovery curves (Figure 5 C) indicate that approximately 20 genomes are sufficient to discover the most common polymorphisms (pME loci detected in $2 < N < 29$ genomes relative to the reference ISO-1/dm6), matching the previous report in euchromatic regions[8]. This comparison suggests that the discovery rate of GraffiTE in heterochromatic regions might be comparable to the euchromatin in this *Drosophila* pangenome. Nevertheless, the bulk of pME appears to be segregating a very low frequency, 18,422 variants being found in 1 to 2 genomes only (Figure 5 C,D). This result is largely expected given the documented TE variation observed in natural *D. melanogaster* populations[11,67]. In contrast to other pME families, we noticed an uneven distribution of variants for members of the INE-1 (Rolling-circle/Helitron) and LARD elements. Most novel INE-1 variants are found in heterochromatic regions, so does LARD pME which is also found vastly through the euchromatin of the chromosome X. Based on their rather high frequency on the pangenome (Figure 5 D), and the fact that these families are estimated to be much older than the other pME [8], it is likely that these variants are indicative of the lack of completion of the dm6 reference in traditionally hard to access regions. In addition, we notice that LARD pME distribution shows two unusual peaks in the heterochromatic regions of chromosome 3L. Both peaks coincide with elevated density of non-TE repeats (low-complexity, simple repeats and satellites). These may be caused by two non-exclusive factors: (i) the lack of completion of the dm6 reference in these regions, and (ii) possible false positive hot-spot, as repeat-rich regions

remain challenging for genome to genome and long-read to genome SV search methods implemented in GraffiTE[68].

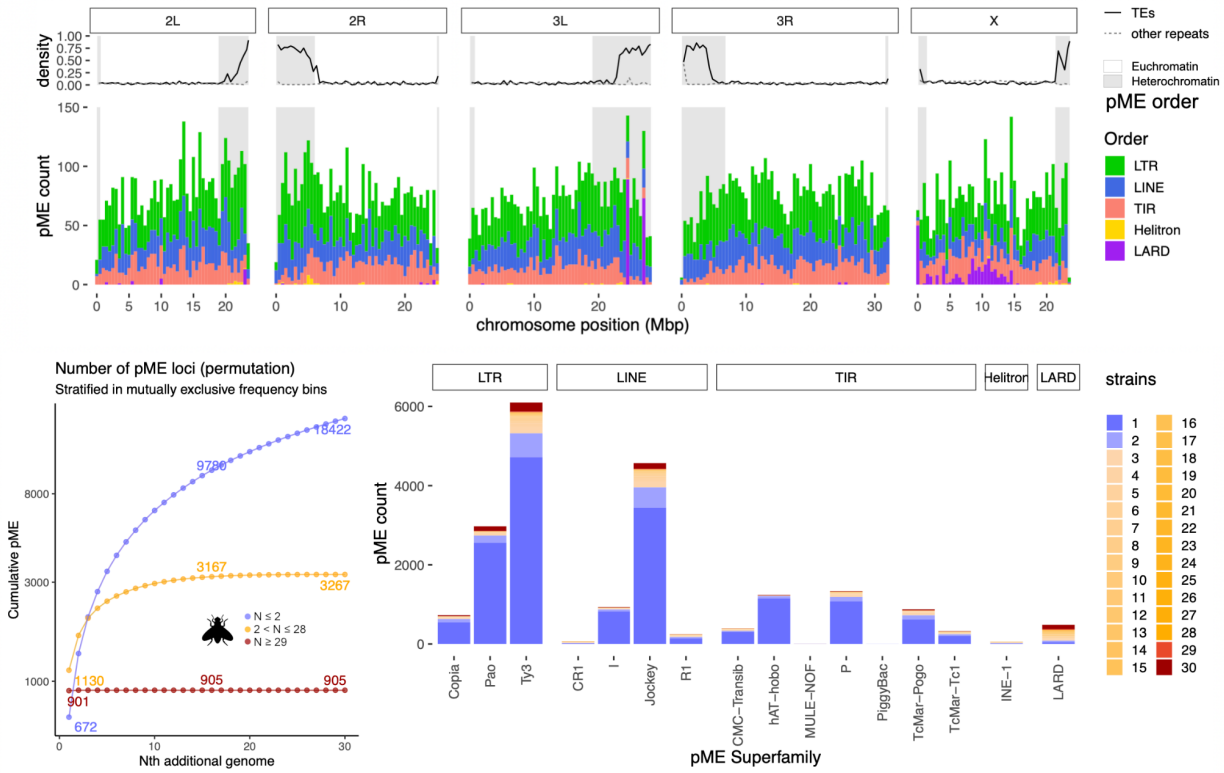

Figure 5 *Drosophila melanogaster* pME pangenome analysis. A. Genomic repeat density for TEs and simple repeats (regrouping RepeatMasker annotations: “low complexity”, “simple repeats”, and “Satellite”) as annotated on the reference genome dm6, per bin of 0.5Mbp. Grey rectangle indicates heterochromatic regions according to Rech et al., 2022 **B.** Distribution of pME detected by GraffiTE using 30 paired genome assemblies and long-read (ONT) sets, further genotyped with the same long-reads and GraphAligner (GT-sv-sn-GA mode). To reduce putative false-positives, pMEs with more than 1 hit on the TE library were filtered out. pME position is relative to the reference genome dm6; pME counts, per TE order and bins of 0.5Mbp. Grey rectangle indicates heterochromatic regions according to Rech et al., 2022 **C.** Discovery curves for pMEs detected with GraffiTE. For each mutually exclusive frequency bin (red: shared by 1-2 genomes, yellow: 3-28, red: 29-30), the average number of pMEs in N genomes is reported after performing 100 permutations and sampling of the genomes. **D.** Count of pMEs detected by GraffiTE according to their superfamily and colored by their frequency in the pangenome.

GraffiTE detects segregating repeat families in a model with limited prior knowledge about TEs

We previously demonstrated two use cases of GraffiTE in model species for which high-confidence, manually curated TE libraries are available. Furthermore, the subset of active TE families in humans and *D. melanogaster* are well described, and provide solid hypotheses for filtering and interpreting the outputs of GraffiTE (expected families, copies' size-range, etc,...). In this last use case, we demonstrated that GraffiTE can also be applied to new models, with little-to-no information available about their TEs. *Cannabis sativa* is an emerging crop of economic and medical importance, and for which genomic resources are accumulating, including several pangenome projects[69]. Moreover, *C.sativa* is an example of TE-rich genomes (>70% of the genome[70]) for which high-level of polymorphism is expected[71]. We first ran RepeatModeler 2[72] on the reference genome Cs10 (GenBank ID: GCA_900626175.2) in order to create a consensus sequences' library of the TEs present in *C. sativa*. This automatic library was combined with a smaller collection of manually curated consensus sequences from Repbase[73] (Figure 6 A and Methods). We collected 9 alternative genome assemblies for *C. sativa*, publicly available through NCBI, and selected them for being based on long-read technologies (Table S3). These 9 assemblies were used as input for GraffiTE (GT-sv mode) using Cs10 as reference. Conversely to human and drosophila models, putative pME annotations were often composed of multiple hits against different consensus sequences from the automatically generated TE library. This is expected for models lacking manual curation of TE catalogs. To circumvent this limitation, we further extracted and clustered together the sequences of all 75,679 variants present in the raw TE-graph-genome generated by GraffiTE (Figure 6 A and Methods). We then retained clusters with N >= 3 sequences (3 distinct loci in the TE-graph-genome) and with a size between 200 and 40,000 bp as probable pMEs and computed the discovery curves (Figure 6 B). We showed that 7 assemblies were required to observe the common pMEs in the *C. sativa* pangenome (pME found in 2 to 8 assemblies), while 208 variants from the reference Cs10 were shared by all 9 alternative genomes. Both common and rare (singleton) pMEs were extremely abundant, for a total of 42,517 loci in total (in contrast, ~15,000 and ~20,000 pMEs were found in 94 human and 30 *D. melanogaster* assemblies, respectively). To further illustrate how GraffiTE can be used to improve TE annotation, we selected the most abundant pME clusters (putative TE families) and measured their abundance in the pangenome (Figure 6 C). Based on copy number, the 100 most abundant pME families represent approximately 25 Mbp of facultative DNA sequence, differentially present or absent between the 9 assemblies and the reference Cs10; the most interspersed TE family (representative sequence: Pink_pepper.svim_asm.INS.10081) being represented by a total of 756 loci in the pangenome. Manual annotation for a subset of these

families highlight the diversity of the (recently) active TEs in *C. sativa*, including Miniature Inverted-repeat Transposable Elements (Class II, MITE), LTR Retrotransposons (Class I) and DNA/TIR of the MULE-MuDR superfamily (Class II).

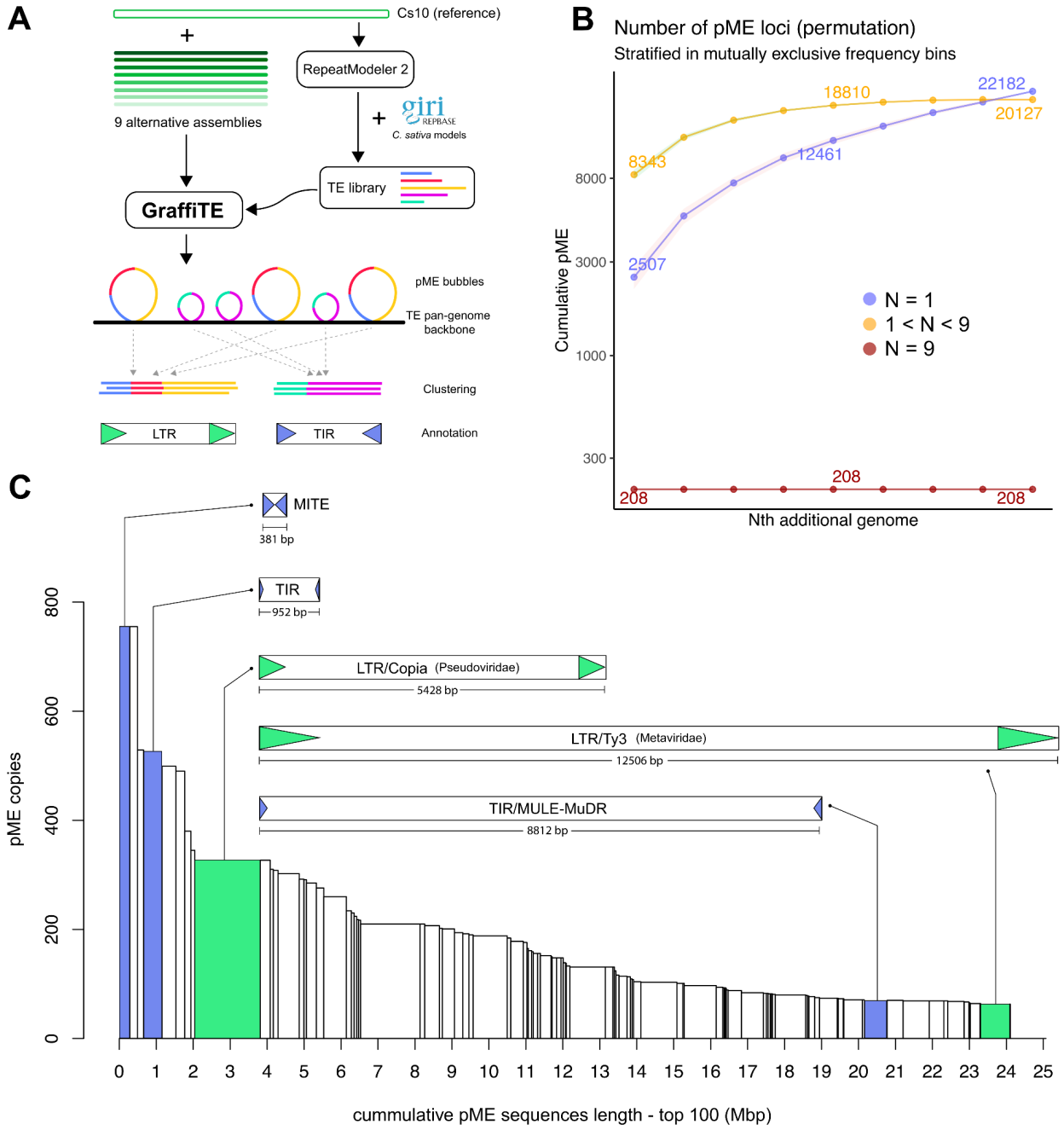

Figure 6. *Cannabis sativa* pME pangenome analysis. A. Overview of the experiment: the reference genome Cs10 was used to generate an automatic TE library with RepeatMolder 2 and subsequently merged with a collection of manually curated consensus sequences from Repbase. Next, 9 alternative assemblies were used as input for GraffiTE (GT-sv mode). Most of the discovered candidate pMEs contained multiple TE annotation from the automatic library; in order to reconstruct TE families, the sequences of 75,679 putative pMEs were clustered with

MMseqs2 and variants belonging to clusters with 3 or more sequences were kept for analysis. **B.** Discovery curves for the selected pMEs in 9 genome assemblies, relative to the reference Cs10. **C.** Distribution of pME clusters size, in total number of loci (y axis) and total sequence length in the pangenome (x axis = pME loci number x pME lengths per cluster). The manually annotated sequence of 5 representative pME families are shown to illustrate the diversity of TE segregating between *C. sativa* cultivars.

Discussion

We created GraffiTE with the goal of facilitating pME analysis in a wide-range of organisms, enabling the exploration of TE dynamics in model species with or without prior information about their mobilome. A particular emphasis was made to include a collection of methods ready for the pangenome era, allowing the direct detection of pMEs from multiple genome assemblies and/or long-read sets. More importantly, GraffiTE introduces for the first time the ability to perform graph-based genotyping of pME loci, by leveraging cutting-edge software such as Pangenie[58] and GraphAligner[61].

A major goal was to provide flexibility and ease of use for a wide audience of researchers. To do so, GraffiTE relies on Nextflow, which provides multiple avantages when designing such pipelines including: (i) the self-containment of the code and its dependencies using a single Appatainer (formerly known as Singularity) image, (ii) the ability to swap or combine different software and modules, and (iii) the scalability of the pipeline, allowing its deployment to a wide array of systems including high performance clusters (HPC) and cloud services. Furthermore, GraffiTE provides detailed annotated outputs, and is released with extensive documentation.

Another strength of GraffiTE is the wide variety of data types to which it can be applied. pMEs can be detected from genome assemblies or any type of long-read data, and genotyping can be performed using short- and long-read sets. This flexibility allows researchers to get the most out of their data; for example, by performing the initial SV search with high-quality – though perhaps less abundant – data, such as chromosome-level assemblies and long-read sequences, while genotyping in larger cohorts or populations using cost-effective short-read sets.

GraffiTE provides additional features not found in existing software, such as the ability to report deletions (pMEs present in the reference genome but absent from a given sample), a feature absent from long-read methods such as TELR[29] or TLDR[28]. While TrEMOLO[74] recently

introduced the ability to detect pMEs from an alternative genome assembly, GraffiTE expands this feature to multiple genomes in a single analysis. Furthermore, GraffiTE uses and produces annotated VCFs as a standard, which allows swift interoperability both between tools within the pipeline but also for the users to share and compare results. Finally, users can annotate and genotype pME from SV call sets generated by other tools, as long as both reference and alternative alleles are documented in an input VCF. Thus, GraffiTE can be directly used on already established collections of SVs, and will allow filtering and annotation of variants to retain and graph-genotype pMEs.

Simulations show that the different modes (pME detection from assemblies only, long-read only, or combined) produce high recall (low false negatives) and that specificity and precision can be greatly improved with simple hypothesis-driven filters, such as pME length and type of annotation. In real data derived from the Genome in a Bottle (GIAB) benchmark[64], the performances are maximized when long-reads are used for pME search (alone, or in combination with assemblies: GT-sn and GT-sv-sn modes). Using only maternal and paternal chromosome assemblies of HG002, and in spite of a reduced recall, precision remains very high (> 90%), in par with long-read modes, and above the scoring of alternative methods such as TLDR (long-reads) or MELT2 and MEGanE (short-reads). This result is expected, as assembly completeness will directly affect the tool ability to detect SVs.

We further evaluated the performance of graph-based genotyping, showing a high concordance with the high-quality bi-alleleic genotypes reported by the GIAB consortium and GraffiTE (Figure 3-B). This result demonstrates that graph-genotyping improves genotype quality of pME variants, which is particularly relevant when pMEs are used in association studies such as GWAS or molecular QTLs[19,75]. Applying GraffiTE with genome assemblies only remains highly informative as we show that 20 to 30 human haploid assemblies are sufficient to discover common polymorphisms in the HPRC pangenome[34], without the need for long-read data. This is particularly relevant, as the storage of raw data from larger pangenomes can be a significant limitation for research teams.

Application to the *Drosophila* pangenome highlights the flexibility of the pipeline to adapt to various models with very diverse TE dynamics. In particular, our analysis shows great consistency with population genetics reports of TE polymorphisms in this species[8,11,67]. Next, we illustrated with the *C. sativa* pangenome how GraffiTE can be used in non- or emerging model

species, for which little-to-no information about TE is available. A simple clustering-based strategy quickly identified the most abundant polymorphic repeats, which could be further manually annotated. Thus, we argue that in complement to automated TE discovery approaches, such as RepeatModeler[72], REPET[76] or EDTA[77], GraffiTE can be a useful addition to tools helping curation of repeat libraries.

As with any automated approach, we can identify current limitations with the proposed software. SVs annotation, in order to filter likely pME, remains particularly challenging. Our current implementation based on homology between the variant sequence and a known TE library, favors generalization but can occasionally provide erroneous or ambiguous results. For example, full-length LTR insertions (proviral LTR) may be reported as three separate hits (5' LTR, internal sequence, 5' LTR) if the library doesn't report the relationship between LTR and internal consensus. Though we attempt to mitigate this behavior using OneCodeToFindThemAll[55], recombination between different LTR families during transposition may lead to similar patterns and cannot be easily detected automatically. In very rare cases, we found that RepeatMasker can report hits much longer than the consensus sequence, which can be incompatible with the adjudicated pME family. Such artifacts can be later discarded by the user using the detailed output of GraffiTE, applying specific knowledge about the concerned TE families.

We present GraffiTE, a powerful and user-oriented tool that facilitates pME analysis in a wide range of organisms. It enables precise analyses in model species as well as exploration of TE dynamics in systems with limited mobilome information. With its graph-based genotyping capabilities and compatibility with multiple data types, including genome assemblies and long-read datasets, GraffiTE offers flexibility and optimal use of researcher's ressources. The software provides unique features such as analyzing multiple assemblies at a time, annotation of structural variants, and interoperability through annotated VCFs. It demonstrates high performance and accuracy in TE detection and genotyping, making it valuable for association studies and pangenome analyses. We aim to provide continuous support for the pipeline, through its Github page (https://github.com/cgroza/GraffiTE) and are looking forward to incorporating user's feedback into updated versions. Future work is planned to tackle complex TE polymorphisms, such as LTR-LTR recombination, pME nesting, and perform gold-standard method evaluations through PCR assays. Thus, we believe that GraffiTE represents a valuable

contribution to TE research, enabling insights into TE biology, genome evolution, and molecular diversity.

Methods

Implementation

SV search. The first step of GraffiTE searches for SV between a reference genome and one or more alternative assemblies for the same species. SV can also be searched from long-read sets (either alternatively or in combination with search from assembled genomes). When using assemblies, each input file is expected to be representing a single haplotype in fasta format (if diploid assemblies are used, each haplotype is analyzed separately). Each queried assembly is first aligned onto the reference with minimap2 (v.2.24-r1155-dirty) using the preset `-x asm5` which is optimized for genome-to-genome alignment below 5% divergence (minimap2 documentation, accessible online at: https://lh3.github.io/minimap2/minimap2.html). This default preset can be modified to accommodate more divergent data, up to 20% divergence (see GraffiTE documentation). Additional parameters `-a --cs -r2k` are also applied, as recommended for the next program SVIM-ASM (v.1.0.3) which calls SVs based on minimap2 outputs (BAM alignments). By default SVIM-ASM is tasked to report insertions and deletions >= 100 bp relative to the reference genome (inversions, duplication and translocation are not reported in order to focus on candidate products of transposition). From long-reads, SVs are searched by first mapping reads to the reference genome with minimap2 (presets `-x ont/pb/hifi` are used according to the input reads, see online documentation). Then, insertion or deletion variants >= 100bp are searched in the alignments using Sniffles 2 (v.2.0.7). Upon SV calling, variants are reported in one VCF file per alternative assembly or read set. If multiple samples are used, individual VCFs are further merged using SURVIVOR (v.1.0.7). SVs are merged into the same loci if they belong to the same type (INS or DEL), their distance is within 10% of their length and each SV >= 100bp (`SURVIVOR merge 0.1 0 1 0 0 100`). GraffiTE allows variants found in only one assembly or read-set to be kept (singletons).

This SV search can be bypassed by directly providing a sequence-resolved VCF (REF and ALT sequence for each variant must be provided) to GraffiTE (Figure S1).

SV filtering. The next step for GraffiTE is to filter the candidate SVs to retain likely pMEs. Insertion and deletion sequences are extracted from the merged VCF (or alternatively a provided sequence-resolved VCF) in FASTA, and scanned using RepeatMasker v. 4.1.4[54] using its RMBlastn engine (modified Blastn distribution). At this step, a user-provided fasta library of TE models to analyze (consensus sequences) is required. The input TE library can be the full collection consensus sequences for the TEs of the target species, or only include TE families of interest. Since TEs can be hitchhiking within SVs caused by other events than transposition, GraffiTE requires that TEs identified by RepeatMasker covers >= 80% of the variant sequence. Using this threshold, the retained variants can include a single or multiple hits against different known TEs. In some cases, multiple hits within the same SV can be due to rearrangement of the newly inserted TE sequence (insertions, deletions or inversion in the TE copy relative to the consensus sequence), others source of multiple TE hits within SV can be caused by distinction in the TE library between long terminal repeat (LTR) and internal part of LTR elements, as well as artifacts (e.g. spurious hits) caused by the RepeatMasker algorithm. To help identify and regroup hits corresponding to the same TE copy (fragments), GraffiTE uses OneCodeToFindThemAll.pl[55], a Perl script dedicated to this task. Note that to take the best advantage of this step, it is preferable that the nomenclature of the TE library follows the RepeatMasker style: TENAME#Order/Superfamily (e.g. L1HS#LINE/L1) and that internal (I) and LTR sequences of the same elements are named such as TENAME_I#LTR/Superfamily and TENAME_LTR#LTR/Superfamily (e.g. ROO_I#LTR/Pao and ROO_LTR#LTR/Pao). Following this step, for each SV retained the number of hits (distinct TE instances), fragments (pieces of the same TE instance) and classification details are reported in a VCF file.

TSD search. We implemented a module to search for target site duplications (TSD) which can be the hallmark of transposition in many TE families. Any TE-containing variant retained following the previous step (SV filtering) and for which a single TE is identified (this can include hits made of multiple fragments) will be processed. The TSD (if present) is often located at the 5’ or 3’ end of the SV sequence, while the second TSD is present at the opposite end, in the flanking region. However, TSDs can sometimes be found overlapping the junction between the flanking and the first base pairs of the SV. To take these cases into consideration, the TSD module of GraffiTE first removes the base pairs masked by RepeatMasker in the SV, leaving (if any reminder) a few base pairs at the 5’ and 3’ end. Then, these reminders (if any) are concatenated to 30 bp of flanking sequence (directly adjacent to the SV breakpoints) at each respective end (Figure S2 A). These reconstructed flanking sequences are then compared to

each other using blastn (v. 2.9.0+), with a seed of 4bp (thus the minimal TSD size reported is 4bp). Though the best match (longest hit with the least mismatches) for each comparison (each pME) is reported in an output table, we implemented an empirical filter to reduce false positives. The filter returns a PASS flag (and report the TSD sequence in the output VCF) if the TSD initiates or ends within 5bp of the defined TE end (according to the consolidated RepeatMasker outputs) and if the number of mismatches is below 1 or [TSD-length]xDIV whichever the larger, with DIV the divergence of the TE copy relative to its consensus, as reported by RepeatMasker (i.e. older TE insertion tolerates more mismatches). An additional subroutine makes sure that the TSD are located outside any poly-A (or T) tail (Figure S2 B).

Additional annotation filters. Initial testing on human data revealed that the automated annotation of GraffiTE was, as-is, likely to mislabel specific types of TE-associated polymorphisms. In particular, we noted that 5' LINE-1 inversions (due to twin-priming and associated mechanisms [66,78] would be reported as two distinct L1 insertions on opposite orientations and won't be stitched together by OneCodeToFindThemAll. Accordingly, we implemented a filter to recognize such configuration and re-label insertions as a single "twin-primed" L1 (Figure S3). In addition, GraffiTE is able to detect variation in the number of tandem repeats (VNTR) present within SINE-VNTR-Alu elements (SVA), a human lineage-specific type of TE. In order to distinguish insertion polymorphism of new SVA copies (presence/absence of the SVA) from VNTR-only variation (SVA present in both haplotypes, but with VNTR length variation between alleles) the variant's sequence (being either an insertion or a deletion relative to the reference genome) is blasted against the consensus sequence of each of the 6 known SVA models (SVA_A, B, C, D, E and F). If the variant fully maps within the VNTR region of the consensus, it will be labeled "VNTR-only" in the INFO field of the output VCF, while canonical SVA (i.e. pME) insertion will not harbor this tag.

Variant validation and genotyping. Following the SV detection and annotation to retain likely pMEs, GraffiTE gives the possibility to validate and genotype (i.e. predicts the alleles of) each variant by mapping reads from individual samples onto a genome graph where TE-containing haplotypes (each insertion or deletion sequence in the annotated VCF) are represented as bubbles and TE-absent haplotypes are represented by short paths (skipping the TE bubbles). In order to accommodate different sequencing technology, we included Pangenie (short-reads[58]), Giraffe (short-reads[60]) and GraphAligner (long-reads, including low-fidelity ONT[61]).

Evaluation of GraffiTE performance

Simulations. In order to evaluate the ability of the tools implemented to retrieve pMEs from either assemblies or long-read sets, we simulated pMEs and random (background) SVs on the human chromosome 22. For each simulation, 20 pMEs were randomly sampled among members of the AluY, L1HS/L1PA2 and SVA_E,F superfamilies present in the reference genome GRCh38.p14 (hg38) while 20 background SVs were sampled from random intervals in the genome (Supplementary Methods 1A). The VCF file was then passed to SimuG, a perl script tasked to create artificial genomes based on the variants described in the VCF[79]. Once generated, each simulated genome was artificially sequenced at 10X depth with PBSIM3[80].

GIAB Benchmark. An ideal benchmark dataset would include a set of TE insertion polymorphisms for which the presence or absence is known at the allelic level with high-confidence (ideally PCR-based genotyping) in multiple individuals. Such an optimal dataset does not directly exist for pMEs, however we created a reference pME catalog from the high-confidence structural variants sets reported by the genome in a bottle consortium (GIAB), which offer an exhaustive representation of the SVs > 50bp found in the genome of HG002/NA24385 compared to the reference hs37h5 (GRCh37/hg19, accessible at: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references/GRCh37/). The GIAB/HG002 variant calls rely on the combination of 4 sequencing technologies (Illumina 250bp paired-end, Illumina mate-pair, 10X genomics linked-reads, and PacBio HiFi) and 19 SV callers (accessible at: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/). To retain SVs representing pMEs, we applied the GraffiTE annotation step on the HG002_SVs_Tier1_v0.6.vcf.gz VCF file using the --vcf input option. We collected the complete list of human TE consensus sequences from DFAM v.3.6[81] in fasta format using the FamDB toolkit (available online at: https://github.com/Dfam-consortium/FamDB). We then retained pMEs with a single RepeatMasker hit along more then 80% of the sequence and annotated as Alu, LINE-1 or SVA, with a minimum size of 250 bp.

To test GraffiTE and compare its outputs with other existing tools (see below), we collected the alternative assemblies HG002.mat.cur.20211005 (GenBank: GCA_021950905.1) and HG002.pat.cur.20211005 (GenBank: GCA_021951015.1) representing the maternal and paternal haplotypes of HG002. We also gathered ~50X of Illumina 2x250bp short-reads

available for HG002 (140528_D00360_0018_AH8VC6ADXX and 140528_D00360_0019_BH8VDAADXX/ accessible at https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/HG002_HiSeq300x_fastq/) and ~32X of PacBio high-fidelity (HiFi) long-reads (accessible at: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_SequelII_CCS_11kb/reads/). The read sets (long and short) were subsequently re-sampled to 5, 10, 20 and 30X coverage using Rasusa[82].

pME detection evaluation

In order to assess the performance of GraffiTE on pME detection, we ran the pipeline with different combinations of tools and data (Table S1). First, we ran GraffiTE using only the maternal and paternal assemblies as input; in this condition (further labeled GT-sv), SVIM-asm is the only software used to detect variants. Alternatively, we ran GraffiTE on long-read data only, using either 5, 10, 20 or 30X coverage. This analysis implied the use of Sniffles 2[51] and is further referred to as GT-sn. Finally, we also run GraffiTE with a combination of alternative assemblies and increasing coverage of long-reads (GT-sv-sn). GraffiTE was run with the flag --mammal, which identifies L1 with 5' inversion as a single pME and annotates SVA VNTR polymorphism (SVA-VNTR polymorphisms were further discarded for the benchmark). We filtered the GraffiTE output VCFs using the same criteria that for the GIAB reference call: we retained only variants with a single RepeatMasker hit to an Alu, L1 or SVA element and a variant length >= 250bp. According to their input specification, we also ran a collection of representative methods either using long-reads (TLDR[28]) or short-reads (MELT2[6], MEGAnE[65]). Details about the parameters used and output conversion in the VCF format can be found in Supplementary Methods 1b. To asses pME search performances, comparison between the different methods and the constructed GIAB pME reference set were carried out using the R package sveval[83] with defaults parameters and provided the file HG002_SVs_Tier1_v0.6.bed (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.bed) which "*should contain close to 100 % of true insertions and deletions >=50bp*" according to the GIAB documentation.

pME genotyping evaluation.

We further seeked to evaluate the performance of the graph-genotyping tools implemented in GraffiTE, specifically Pangenie (for short-read data) and GraphAligner (for long-read data). We selected the pME annotated VCFs from GraffiTE for the runs GT-sv, GT-sn and GT-sv-sn and performed graph-genotyping with short- or long-reads at 30X coverage (Table S1). We also used the results from MELT2 and MEGAnE (at 30X coverage) for evaluation, as these tools provide bi-allelic genotypes for the detected pMEs. Genotyping performance was evaluated using sveval with the additional parameters: geno.eval=TRUE, method="bipartite", stitch.hets=TRUE, merge.hets=FALSE.

Application Examples

Human pangenome. We ran GraffiTE on the 47 HPRC diploid assemblies [34] (94 inputs as haploid genomes) using the human transposable element models from DFAM 3.6[81] and the hg38 reference genome[84]. We then calculated discovery curves, separating polymorphisms into rare (<=5% frequency), common (between 5% and 95%) and “fixed” (>95%) categories, permuting the order of genomes 100 times. We recapitulated the expected population structure by running principal coordinate analysis on pME presence/absence calls present in the GraffiTE pangenome.vcf output (no graph-genotyping was performed as only assemblies were used) with the ade4 R package[85] (distance = coefficient S5 of [86]) using pMEs with a pangenome frequency between 5% and 95%. Samples’ metadata were obtained from[34].

***Drosophila melanogaster* pangenome.** We obtained genome assemblies and raw long read sets for 30 genomes resequenced by[8] and using ONT long-reads technology (details can be found on table S2 of [8], for genome with “technology: ONT”). We then applied GraffiTE, using for each sample both the assembly and long-reads (ONT) set for SV search, and GraphAligner to genotype pMEs in the TE pangenome with long-reads (GT-sv-sn-GA mode, see table S1-B). To annotate pMEs, we used the new “MCTE” library produced by Rech et al. 2022. The final GraffiTE VCF was filtered to retain variants with a single TE hit (n_hits=1) and fixed variants relative to dm6 after genotyping were removed. We calculated discovery curves for pMEs detected in the *D. melanogaster* pangenome for low frequency (N = 1/30 to 2/30 genomes), common (3/30 ≤ N ≤ 28/30 genomes) and fixed (N = 29/30 to 30/30 genomes), permuting and sampling genomes 100 times for each frequency bin.

Cannabis sativa *pangenome.* We used the cs10 reference genome (GenBank: GCA_900626175.2) and 9 publicly available alternative assemblies (Table S3) to build a TE-graph-genome with GraffiTE. First, we created a *de-novo* TE library using RepeatModeler2[72] and clustered 1737 models automatically generated with 128 models present in Repbase v.27.06[87] to reduce redundancy and remove the models already present in Repbase from the automatic library (Supplementary Methods 2.b). We then applied GraffiTE to the 9 alternative assemblies, using cs10 as reference and the generated TE library. Conversely to *D. melanogaster* and human samples, pMEs for C. sativa were previously undocumented. After running GraffiTE, we noticed that several reported pMEs had annotations composed of a patchwork of hits against different consensus sequences of the library (which is common with non-curated libraries). To circumvent this, and demonstrate the use of GraffiTE with non-model organisms, we further clustered the sequences of all candidate pMEs (variants reported in the pangenome.vcf output) among the 9 genomes using MMseqs2 (parameters easy-cluster --cov-mode 0 -c 0.8 --min-seq-id 0.8 -s 100 --threads 8 --exact-kmer-matching 1, see also Figure 6A). This clustering allowed to group together pME loci which share at least 80% of their sequence with a minimum identity of 80%. We then retained clusters with a minimum of 3 loci as putative TE families and with a size of 200 to 40,000 bp. The top 100 (in number of pME loci per cluster/family across the *C. sativa* pangenome) were analyzed and representative sequences of 4 families were manually annotated to illustrate the diversity of pME families in *C. sativa*.

References

1. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).
2. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
3. Chandler, M., Gellert, M., Lambowitz, A. M., Rice, P. A. & Sandmeyer, S. B. *Mobile DNA III*. (John Wiley & Sons, 2020).
4. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).

5. Bourgeois, Y. & Boissinot, S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes* **10**, 419 (2019).
6. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
7. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
8. Rech, G. E. *et al.* Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. *Nat. Commun.* **13**, 1948 (2022).
9. Watkins, W. S. *et al.* The Simons Genome Diversity Project: A global analysis of mobile element diversity. *Genome Biol. Evol.* **12**, 779–794 (2020).
10. Goubert, C. *et al.* High-throughput sequencing of transposable element insertions suggests adaptive evolution of the invasive Asian tiger mosquito towards temperate environments. *Mol. Ecol.* **26**, 3968–3981 (2017).
11. Lerat, E. *et al.* Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol. Ecol.* **28**, 1506–1522 (2019).
12. Li, Z.-W. *et al.* Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana. *Genome Biol. Evol.* **10**, 2140–2150 (2018).
13. Rech, G. E. *et al.* Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLoS Genet.* **15**, e1007900 (2019).
14. Van't Hof, A. E. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).
15. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nat. Rev. Genet.* **20**, 760–772 (2019).
16. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2012).

17. Goubert, C., Zevallos, N. A. & Feschotte, C. Contribution of unfixed transposable element insertions to human regulatory variation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190331 (2020).
18. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
19. Chen, X., Bourque, G. & Goubert, C. Genotyping of Transposable Element Insertions Segregating in Human Populations Using Short-Read Realignments. *Methods Mol. Biol.* **2607**, 63–83 (2023).
20. Rajaby, R. & Sung, W.-K. TranSurVeyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic Acids Res.* **46**, e122 (2018).
21. Chen, X. & Li, D. ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. *Bioinformatics* **35**, 3913–3922 (2019).
22. Kojima, S. *et al.* Mobile elements in human population-specific genome and phenotype divergence. *bioRxiv* 2022.03.25.485726 (2022) doi:10.1101/2022.03.25.485726.
23. Bogaerts-Márquez, M. *et al.* T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics* **36**, 1191–1197 (2020).
24. Website. https://doi.org/10.1101/2023.02.13.528343 doi:10.1101/2023.02.13.528343.
25. Yu, T. *et al.* A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. *Nucleic Acids Res.* **49**, e44 (2021).
26. Rahman, R. *et al.* Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Res.* **43**, 10655–10672 (2015).
27. Kofler, R., Gómez-Sánchez, D. & Schlötterer, C. PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol. Biol. Evol.* **33**, 2759–2764

(2016).

28. Ewing, A. D. *et al.* Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol. Cell* **80**, 915–928.e5 (2020).
29. Han, S. *et al.* Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. *Nucleic Acids Res.* **50**, e124 (2022).
30. Mohamed, M. *et al.* A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells* **9**, (2020).
31. Zhou, W. *et al.* Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020).
32. Chu, C. *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* **12**, 3836 (2021).
33. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
34. Liao, W.-W. *et al.* A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321 (2022) doi:10.1101/2022.07.09.499321.
35. Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390 (2016).
36. Shang, L. *et al.* A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
37. Ruggieri, A. A. *et al.* A butterfly pan-genome reveals that a large amount of structural variation underlies the evolution of chromatin accessibility. *Genome Res.* **32**, 1862–1875 (2022).
38. *The Pangenome*. (Springer International Publishing).
39. Groza, C. *et al.* Pangenome graphs improve the analysis of rare genetic diseases. *bioRxiv* (2023) doi:10.1101/2023.05.31.23290808.
40. Li, R. *et al.* A sheep pangenome reveals the spectrum of structural variations and their

effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023).

41. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).
42. Gupta, P. K. GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers. *Bioessays* **43**, e2100109 (2021).
43. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
44. Qi, W. *et al.* The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *Gigascience* **11**, (2022).
45. Groza, C. *et al.* Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. *Cell Genom* **3**, 100294 (2023).
46. Groza, C., Kwan, T., Soranzo, N., Pastinen, T. & Bourque, G. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol.* **21**, 124 (2020).
47. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
48. Kurtzer, G. M., Bauer, M., Kaneshiro, I., Trudgian, D. & Godlove, D. hpcng/singularity: Singularity 3.7.3. (2021) doi:10.5281/zenodo.4667718.
49. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
50. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).
51. Smolka, M. *et al.* Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *bioRxiv* 2022.04.04.487055 (2022) doi:10.1101/2022.04.04.487055.
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

53. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 1–11 (2017).
54. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0. 2013-2015*.
55. Bailly-Bechet, M., Haudry, A. & Lerat, E. 'One code to find them all': a perl tool to conveniently parse RepeatMasker output files. *Mob. DNA* **5**, 1–15 (2014).
56. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
59. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
60. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
61. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
62. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
63. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
64. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *bioRxiv* (2019) doi:10.1101/664623.
65. Kojima, S. *et al.* Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. *Nat. Genet.* (2023) doi:10.1038/s41588-023-01390-2.
66. Meyer, T. J., Srikanta, D., Conlin, E. M. & Batzer, M. A. Heads or tails: L1

insertion-associated 5' homopolymeric sequences. *Mob. DNA* **1**, 7 (2010).

67. Kapun, M. *et al.* Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Mol. Biol. Evol.* **37**, 2661–2678 (2020).
68. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
69. Sirangelo, T. M., Ludlow, R. A. & Spadafora, N. D. Multi-Omics Approaches to Study Molecular Mechanisms in Cannabis sativa. *Plants* **11**, (2022).
70. Gao, S. *et al.* A high-quality reference genome of wild Cannabis sativa. *Hortic Res* **7**, 73 (2020).
71. Pisupati, R., Vergara, D. & Kane, N. C. Diversity and evolution of the repetitive genomic content in Cannabis sativa. *BMC Genomics* **19**, 156 (2018).
72. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
73. Repbase Reports - 2022, Volume 22, Issue 4. https://www.girinst.org/2022/vol22/issue4/.
74. Mohamed, M. *et al.* TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. *Genome Biol.* **24**, 63 (2023).
75. Billingsley, K., Thomas, J. & Goubert, C. Transposable Element Structural Variants in Parkinson's Disease: Focusing on Genotyping Alu Transposable Element Insertions with TypeTE. in *Neuromethods* 43–62 (Springer US, 2022).
76. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
77. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
78. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation

of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).

79. Yue, J.-X. & Liti, G. simuG: a general-purpose genome simulator. *Bioinformatics* **35**, 4442–4444 (2019).
80. [No title]. https://academic.oup.com/nargab/article/4/4/lqac092/6855700.
81. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
82. Hall, M. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* **7**, 3941 (2022).
83. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
84. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
85. Thioulouse, J. *et al. Multivariate Analysis of Ecological Data with ade4*. (Springer, 2018).
86. Gower, J. C. & Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classification* **3**, 5–48 (1986).
87. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).