Chen et al. Genome Biology (2019) 20:291
https://doi.org/10.1186/s13059-019-1909-7

Genome Biology

METHOD Open Access

Paragraph: a graph-based structural variant genotyper for short-read sequence data

Sai Chen[1†], Peter Krusche[2,3†], Egor Dolzhenko[1], Rachel M. Sherman[4], Roman Petrovski[2], Felix Schlesinger[1], Melanie Kirsche[4], David R. Bentley[2], Michael C. Schatz[4,5], Fritz J. Sedlazeck[6†] and Michael A. Eberle[1*†]

Abstract

Accurate detection and genotyping of structural variations (SVs) from short-read data is a long-standing area of development in genomics research and clinical sequencing pipelines. We introduce Paragraph, an accurate genotyper that models SVs using sequence graphs and SV annotations. We demonstrate the accuracy of Paragraph on whole-genome sequence data from three samples using long-read SV calls as the truth set, and then apply Paragraph at scale to a cohort of 100 short-read sequenced samples of diverse ancestry. Our analysis shows that Paragraph has better accuracy than other existing genotypers and can be applied to population-scale studies.

Keywords: Sequence graphs, Targeted variant calling, Structural variation, Population studies

Background

Structural variants (SVs) contribute to a large fraction of genomic variation and have long been implicated in phenotypic diversity and human disease [1–3]. Whole-genome sequencing (WGS) is a common approach to profile genomic variation, but compared to small variants, accurate detection and genotyping of SVs still remains a challenge [4, 5]. This is especially problematic for a large number of SVs that are longer than the read lengths of short-read (100–150 bp) high-throughput sequence data, as a significant fraction of SVs have complex structures that can cause artifacts in read mapping and make it difficult to reconstruct the alternative haplotypes [6, 7].

Recent advances in long-read sequencing technologies (e.g., Pacific Biosciences and Oxford Nanopore Technologies) have made it easier to detect SVs, including those in low-complexity and non-unique regions of the genome. This is chiefly because, compared to short reads, long (10–50 kbp) reads can be more reliably mapped to such regions and are more likely to span entire SVs [8–10]. These technologies combined with data generated by population studies using multiple sequencing platforms are leading to a rapid and ongoing expansion of the reference SV databases in a variety of species [11–13].

Currently, most SV algorithms analyze each sample independent of any prior information about the variation landscape. The increasing availability and completeness of a reference database of known SVs, established through long-read sequencing and deep coverage short-read sequencing, makes it possible to develop methods that use prior knowledge to genotype these variants. Furthermore, if the sequence data remain available, they can be re-genotyped using new information as the reference databases are updated. Though the discovery of de novo germline or somatic variants will not be amenable to a genotyping approach, population studies that involve detection of common or other previously known variants will be greatly enhanced by genotyping using a reference database that is continually updated with newly discovered variants.

Targeted genotyping of SVs using short-read sequencing data still remains an open problem [14]. Most targeted methods for genotyping are integrated with particular discovery algorithms and require the input SVs to be originally discovered by the designated SV caller [15–17], require a complete genome-wide realignment [18, 19], or need to be optimized on a set of training samples [12, 20]. In addition, insertions are generally more difficult to detect than deletions using short-read technology and thus are usually genotyped with lower

* Correspondence: meberle@illumina.com
†Sai Chen and Peter Krusche contributed equally to this work.
†Fritz J. Sedlazeck and Michael A. Eberle contributed equally to this work.
[1]Illumina Inc, 5200 Illumina Way, San Diego, CA, USA
Full list of author information is available at the end of the article

accuracy or are completely excluded by these methods [21–23]. Finally, consistently genotyping SVs across many individuals is difficult because most existing genotypers only support single-sample SV calling.

Here, we present a graph-based genotyper, Paragraph, that is capable of genotyping SVs in a large population of samples sequenced with short reads. The use of a graph for each variant makes it possible to systematically evaluate how reads align across breakpoints of the candidate variant. Paragraph can be universally applied to genotype insertions and deletions represented in a variant call format (VCF) file, independent of how they were initially discovered. This is in contrast to many existing genotypers that require the input SV to have a specific format or to include additional information produced by a specific de novo caller [14]. Furthermore, compared to alternate linear reference-based methods, the sequence graph approach minimizes the reference allele bias and enables the representation of pan-genome reference structures (e.g., small variants in the vicinity of an SV) so that variants can be accurate even when variants are clustered together [24–28].

We compare Paragraph to five popular SV detection and genotyping methods and show that the performance of Paragraph is an improvement in accuracy over the other methods tested. Our test set includes 20,108 SVs (9238 deletions and 10,870 insertions) across 3 human samples for a total of 60,324 genotypes (38,239 alternative and 22,085 homozygous reference genotypes). Against this test set, Paragraph achieves a recall of 0.86 and a precision of 0.91. By comparison, the most comprehensive alternative genotyping method we tested achieved 0.76 recall and 0.85 precision across deletions only. In addition, the only discovery-based SV caller we tested that could identify both insertions and deletions had a recall of 0.35 for insertions compared to 0.88 for Paragraph. Finally, we showcase the capability of Paragraph to genotype on a population-scale using 100 deep-coverage WGS samples, from which we detected signatures of purifying selection of SVs in functional genomic elements. Combined with a growing and improving catalog of population-level SVs, Paragraph will deliver more complete SV calls and also allow researchers to revisit and improve the SV calls on historical sequence data.

Result

Graph-based genotyping of structural variations

For each SV defined in an input VCF file, Paragraph constructs a directed acyclic graph containing paths representing the reference sequence and possible alternative alleles (Fig. 1) for each region where a variant is reported. Each node represents a sequence that is at least one nucleotide long. Directed edges define how the node sequences can be connected to form complete haplotypes. The sequence for each node can be specified explicitly or retrieved from the reference genome. In the sequence graph, a branch is equivalent to a variant breakpoint in a linear reference. In Paragraph, these breakpoints are genotyped independently and the genotype of the variant can be inferred from genotypes of individual breakpoints (see the "Methods" section). Besides genotypes, several graph alignment summary statistics, such as coverage and mismatch rate, are also computed which are used to assess quality, filter, and combine breakpoint genotypes into the final variant genotype. Genotyping details are described in the "Methods" section.

Construction of a long read-based ground truth

To estimate the performance of Paragraph and other existing methods, we built a long-read ground truth (LRGT) from SVs called in three samples included in the Genome in a Bottle (GIAB) [11, 29] project data: NA12878 (HG001), NA24385 (HG002), and NA24631 (HG005). Long-read data from these three individuals was generated on a Pacific Biosciences (PacBio) Sequel system using the Circular Consensus Sequencing (CCS) technology (sometimes called "HiFi" reads) [30]. Each sample was sequenced to an average of 30 fold depth and ~11,100 bp read length. Previous evaluations showed high recall (0.91) and precision (0.94) for SVs called from PacBio CCS NA24385 with similar coverage levels against the GIAB benchmark dataset in confident regions [11, 30], thus indicating SVs called from CCS data can be effectively used as ground truth to evaluate the performance of SV genotypers and callers.

For each sample, we called SVs (50 bp+) as described in the "Methods" section and identified a total of 65,108 SV calls (an average 21,702 SVs per sample) representing 38,709 unique autosomal SVs. In addition, we parsed out SV loci according to regions with a single SV across the samples and those with multiple different SVs and identified that 38,239 (59%) of our SV calls occur as single, unique events in the respective region and the rest 26,869 (41%) occur in regions with one or more nearby SVs (Additional file 1: Figure S1). Recent evidence suggests that a significant fraction of novel SVs could be tandem repeats with variable lengths across the population [31, 32], and we found that 49% of the singleton unique SVs are completely within the UCSC Genome Browser Tandem Repeat (TR) tracks while 93% of the clustered unique SVs are within TR tracks. Because regions with multiple variants will pose additional complexities for SV genotyping that are beyond the scope of the current version of Paragraph, we limited our LRGT to the 9238 deletions and 10,870 insertions that are not confounded by the presence of a different nearby or overlapping SV (see the "Methods" section). Considering

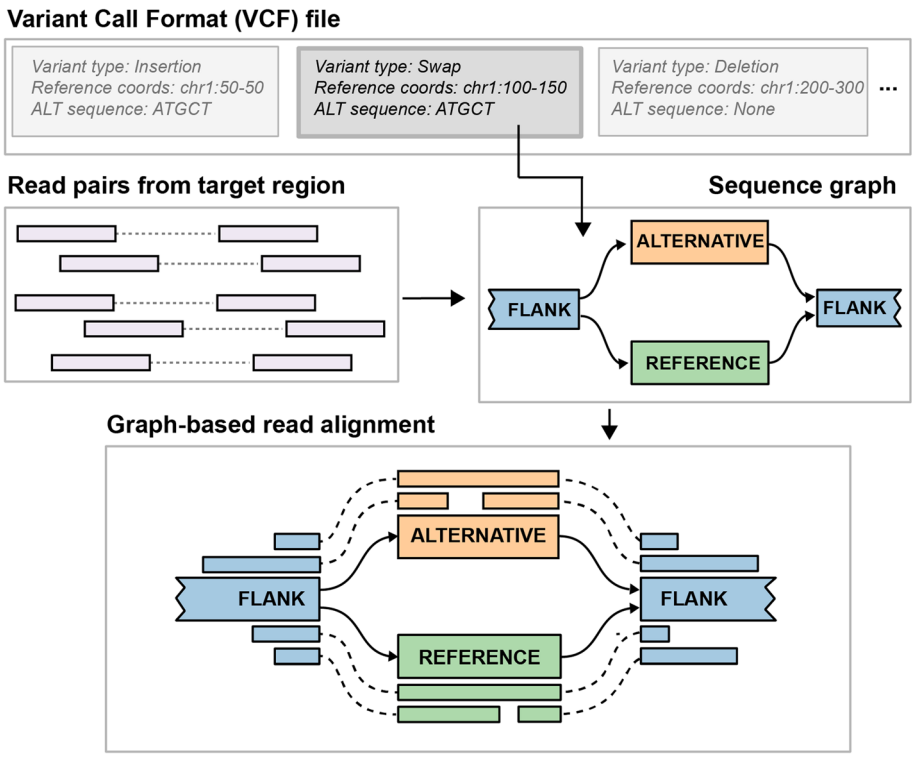

Fig. 1 Overview of the SV genotyping workflow implemented in Paragraph. The illustration shows the process to genotype a blockwise sequence swap. Starting from an entry in a VCF file that specifies the SV breakpoints and alternative allele sequences, Paragraph constructs a sequence graph containing all alleles as paths of the graph. Colored rectangles labeled FLANK, ALTERNATIVE, and REFERENCE are nodes with actual sequences, and solid arrows connecting these nodes are edges of the graph. All reads from the original, linear alignments that aligned near or across the breakpoints are then realigned to the constructed graph. Based on alignments of these reads, the SV is genotyped as described in the "Methods" section

all three samples, there are (1) 4260/4439 deletions/insertions that occurred in just 1 sample, (2) 2258/2429 deletions/insertions that occurred in 2 samples, and (3) 2720/4002 deletions/insertions that occurred in all 3 samples. With short-read sequencing also available for these three samples, we are able to test any SV genotyping method and can estimate recall and precision using the long-read genotypes as the ground truth.

Test for recall and precision

To evaluate the performance of different methods, we genotyped the LRGT SVs on short-read data of NA12878 (63×), NA24385 (35×), and NA24631 (40×) using Paragraph and two widely used SV genotypers, SVTyper [16] and Delly Genotyper [17]. Additionally, we ran three methods that independently discover SVs (i.e., de novo callers), Manta [21], Lumpy [33], and Delly [17]. Because the genotyping accuracy of classifying homozygous versus heterozygous alleles may vary for the short- and long-read methods used here, we focus our test on the presence/absence of variants and not genotyping concordance. Thus, we define a variant as a true positive (TP) if LRGT also has a call in the same sample and a false positive (FP) if LRGT did not call a variant in that sample. We have 38,239 individual alternative genotypes in LRGT to calculate TPs and 22,085 individual reference genotypes in LRGT to calculate FPs. Since some of the methods are not able to call certain sizes or types of SVs, we only tested these methods on a subset of the SVs when calculating recall and precision.

Paragraph has the highest recall: 0.84 for deletions and 0.88 for insertions (Table 1) among all the genotypers and de novo callers tested. Of the genotypers, Paragraph had the highest genotype concordance compared to the LRGT genotypes (Additional file 1: Table S1). The precision of Paragraph is estimated as 0.92 for deletions, which is 7% higher than Delly Genotyper (0.85), and 0.89 for insertions. Though SVTyper had the highest precision (0.98) of all the methods tested, it achieved that by sacrificing recall (0.70). Furthermore, SVTyper is limited to deletions longer than 100 bp. When measuring precision only on 100 bp+ deletions, Paragraph has a slightly lower precision (0.93) than SVTyper (0.98) but the recall is 12% higher (0.82 vs. SVTyper 0.70). Combining recall and precision, Paragraph has the highest *F*-score among all genotypers also for this subset of 100

Table 1 Performance of different genotypers and de novo callers, measured against 50 bp or longer SV from our LRGT

Type	Deletion						Insertion	
	Paragraph	Delly Genotyper	SVTyper (100+ bp)	Manta	Delly	Lumpy (100+ bp)	Paragraph	Manta
#Tested TPs	16,936	16,936	11,160	16,936	16,936	11,160	21,303	21,303
Recall	0.84	0.76	0.70	0.62	0.61	0.64	0.88	0.35
#Tested FPs	10,778	10,778	6960	–	–	–	11,307	–
Precision	0.92	0.85	0.98	–	–	–	0.89	–
F-score	0.88	0.80	0.82	–	–	–	0.88	–

Genotyping/calling was evaluated on short-read data of the three samples sequenced with 150 bp paired-end reads on Illumina platforms. As SVTyper and Lumpy are limited to deletions longer than 100 bp, they have fewer tested SVs than other methods

bp+ deletions (0.88 vs. 0.80 for Delly Genotyper and 0.82 for SVTyper). In addition, we tested another short-read genotyper, BayesTyper, a kmer-based method, and estimated a recall of 0.47 and precision of 0.94 across all of the LRGT SVs. The low recall of BayesTyper is because it produced no genotype call for 56% of the LRGT SVs. We speculate that this may be largely caused by sequencing errors that would have a greater impact on methods that require exact matches of kmers.

Since genotyping performance is often associated with SV length (e.g., depth-based genotypers usually perform better on larger SVs than smaller ones), and some of the tested methods only work for SVs above certain deletion/insertion sizes, we partitioned the LRGT SVs by length and further examined the recall of each method (Fig. 2). In general, for deletions between 50 bp and ~ 1000 bp, the genotypers (Paragraph, SVTyper, and Delly Genotyper) have better recall than the de novo callers (Manta, Lumpy, and Delly). SVTyper and Paragraph have comparable recall for larger (> 300 bp) deletions, and in that size range, Delly Genotyper has lower recall than these two. For smaller deletions (50–300 bp), the recall for Paragraph (0.83) remains high while we observe a slight drop in the recall of Delly Genotyper (0.75) and a larger drop in the recall of SVTyper (0.43). We speculate that this is because SVTyper mainly relies on paired-end (PE) and read-depth (RD) information and will therefore be less sensitive for smaller events. Only Paragraph and Manta were able to call insertions, and while Paragraph (0.88) has consistently high recall across all insertion lengths, Manta (0.35) has a

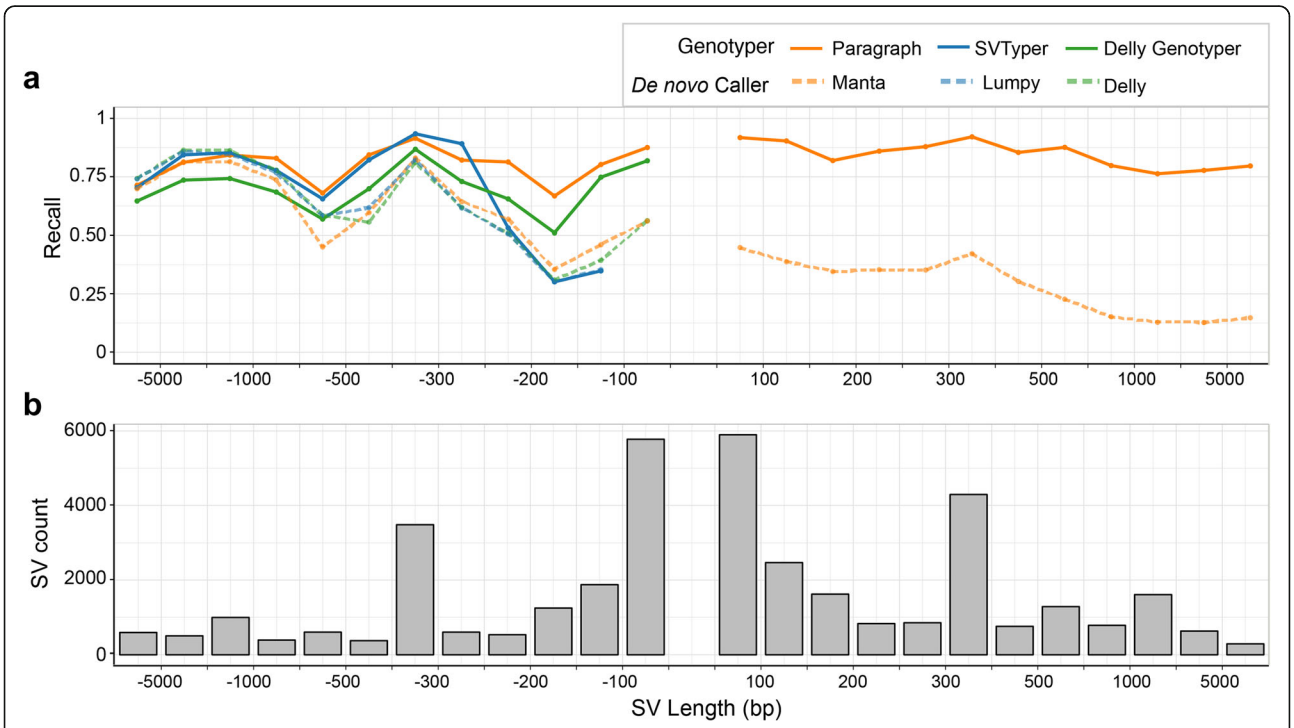

Fig. 2 Estimated recall of different methods, partitioned by SV length. Recall was estimated on the three samples using LRGT as the truth set. A negative SV length indicates a deletion, and a positive SV length indicates an insertion. Colored lines in **a** show recall of different methods; solid gray bars in **b** represent the count of SVs in each size range in LRGT. The center of the plot is empty since SVs must be at least 50 bp in length

much lower recall which drops further for larger insertions.

We additionally partitioned the precision of each genotyper by SV length (Additional file 1: Figure S1). The result suggests that false positives are more likely to occur in small SVs than in large ones. Paragraph has a consistent precision for deletions and insertions, while the only comparable method in genotyping very small deletions (50–100 bp), Delly Genotyper, has a precision drop in this range (Additional file 1: Figure S2). We further examined Paragraph FPs in one of the tested samples, NA24385, and found nearly all of the FP deletions (91%) and the FP insertions (90%) are completely within TR regions. We performed a visual inspection of the 21 FP deletions and 83 FP insertions that are outside of TRs: 12% (12) have 2 or more supporting reads for an SV but were not called by the long-read caller in LRGT, 40% (42) have 1 or more large indels (longer than 10 bp) in the target region, and 48% (50) have no evidence of variants in the long-read alignments in the target region, and thus, these FPs are likely to come from short-read alignment artifacts.

So far, we tested the recall using high depth data (> 35×) with 150 bp reads but some studies may use shorter reads and/or lower read depths. To quantify how either shorter reads or lower depth will impact genotyping performance, we evaluated data of different read lengths and depths by downsampling and trimming reads from our short-read data of NA24385. Generally, shorter read lengths are detrimental to recall; reductions in depth have less of a deleterious effect until the depth is below ~ 20× (Additional file 1: Figure S3).

Genotyping with breakpoint deviations

The LRGT data we used here will be both costly and time-consuming to generate in the near term because generating long-read CCS data is still a relatively slow and expensive process. An alternative approach to build up a reference SV catalog would be to sequence many samples (possibly at lower depth) using PacBio contiguous long reads (CLR) or Oxford Nanopore long reads rather than CCS technology and derive consensus calls across multiple samples. The high error rates (~ 10–15%) of these long reads may result in errors in SV descriptions especially in low-complexity regions where just a few errors in the reads could alter how the reads align to the reference. Since Paragraph realigns reads to a sequence graph using stringent parameters, inaccuracies in the breakpoints may result in a decreased recall.

To understand how the genotypers perform with input SVs that have imprecise breakpoints, we called SVs from CLR data of NA24385 that were generated on a PacBio RS II platform. 9534 out of the total 12,776 NA24385 SVs in LRGT closely match those generated from the CLR data (see the "Methods" section for matching details). Of these, 658 (17%) deletions and 806 (14%) insertions have identical breakpoints in the CLR and CCS SV calls. The remaining 3306 deletions and 4763 insertions, although in approximately similar locations, have differences in representations (breakpoints and/or insertion sequences). Assuming breakpoints found using the CCS data within the LRGT SVs are correct, we consider deviations in the CLR breakpoints as errors in this sample. For the matching deletions between LRGT and CLR calls but with deviating breakpoints, Paragraph recall decreased from 0.97 to 0.83 when genotyped the CLR-defined deletions. Overall, there is a negative correlation between Paragraph recall and breakpoint deviations: the larger the deviation, the less likely the variant can be genotyped correctly (Fig. 3). While deviations of a few base pairs can generally be tolerated without issue, deviations of 20 bp or more reduce recall to around 0.44. For insertions with differences in breakpoints and/or insertion sequences, Paragraph recall decreased from 0.88 to 0.66 when genotyped the CLR-defined insertions. We also investigated how inaccurate breakpoints impact insertion genotyping, but found no clear trend between recall and base-pair deviation in breakpoints.

On the same set of CLR calls, we estimated the impact of breakpoint deviation on SVTyper and Delly Genotyper (Additional file 1: Figure S4). Similar to Paragraph, the split-read genotyper, Delly Genotyper, shows the same negative relationship between its recall and breakpoint deviations. As a contrast, SVTyper, which genotypes SVs mostly using information from read depth and pair-read insert size distribution, does not depend much on breakpoint accuracy and is not significantly affected by deviations in breakpoints.

Genotyping in tandem repeats

We identified that most of the SVs having breakpoint deviations between the CLR calls and LRGT are in low-complexity regions: of the 8069 matching SVs with breakpoint deviations, 3217 (77%) are within TRs. SVs within TRs have larger breakpoint deviations in CLR calls from the true breakpoints than those not in TRs: 35% of the SVs with smaller (≤ 10 bp) deviations are within TRs while 66% of the SVs with larger breakpoint deviations (> 20 bp) are within TRs. Additionally, we found that 59% of the FNs and 77% of the FPs in NA24385 occur in SVs that are completely within TRs. To further understand the impact of TRs on the performance of Paragraph, we grouped LRGT SVs according to whether they are in TRs and plotted Paragraph recall binned by SV lengths. Paragraph has a better recall in SVs that are outside of TRs (0.89 for deletions and 0.90 for insertions), compared to its recall in SVs that are within TRs (0.74 for deletions and 0.83 for

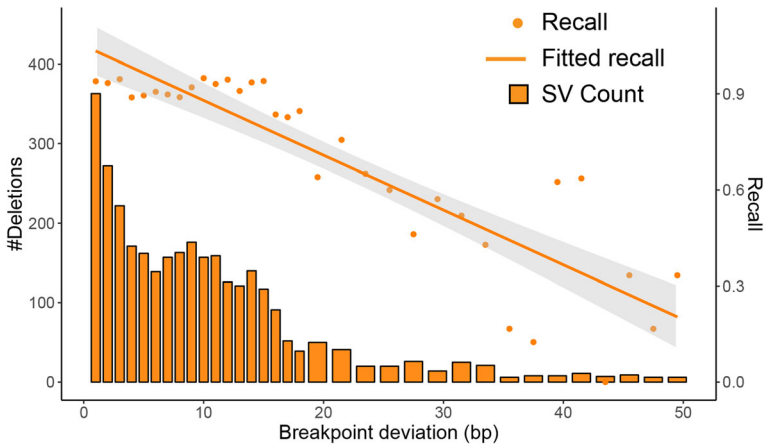

Fig. 3 Demonstration of the impact of recall when tested SVs include errors in their breakpoints. Breakpoint deviations measure the differences in positions between matching deletions in the CLR calls and in LRGT. Paragraph recall was estimated using CLR calls as genotyping input and TPs in LRGT as the ground truth. Breakpoint deviations were binned at 1 bp for deviations less than 18 bp and at 2 bp for deviations larger or equal to 19 bp. Solid bars show the number of deletions in each size range (left axis). Points and the solid line show the recall for individual size and the overall regression curve (right axis)

insertions) (Fig. 4a). Small (< 200 bp) SVs are much more likely to be within TRs (~ 75%) than large (> 1000 bp) SVs (~ 35%) (Fig. 4b), and that matches our earlier observation that Paragraph and other genotypers have decreased recall and precision, in small SVs.

When building our LRGT, we excluded SVs with other nearby SVs in one or more samples (named as clustered SVs in the "Construction of long read-based ground truth" section). The majority of these SVs (93%) are within TRs; therefore, benchmarking against these clustered SVs could be informative to quantify the impact of TRs in SV genotyping. As none of the tested methods could model each SV cluster as a whole without an appropriate annotation, we instead model each of the SVs in the clusters as a single SV and evaluated the performance of Paragraph and other methods on the same three samples using long-read genotypes of these clustered SVs as the underlying truth (Additional file 1: Table S2). All methods have a lower recall and precision in the clustered SVs than in LRGT highlighted by their

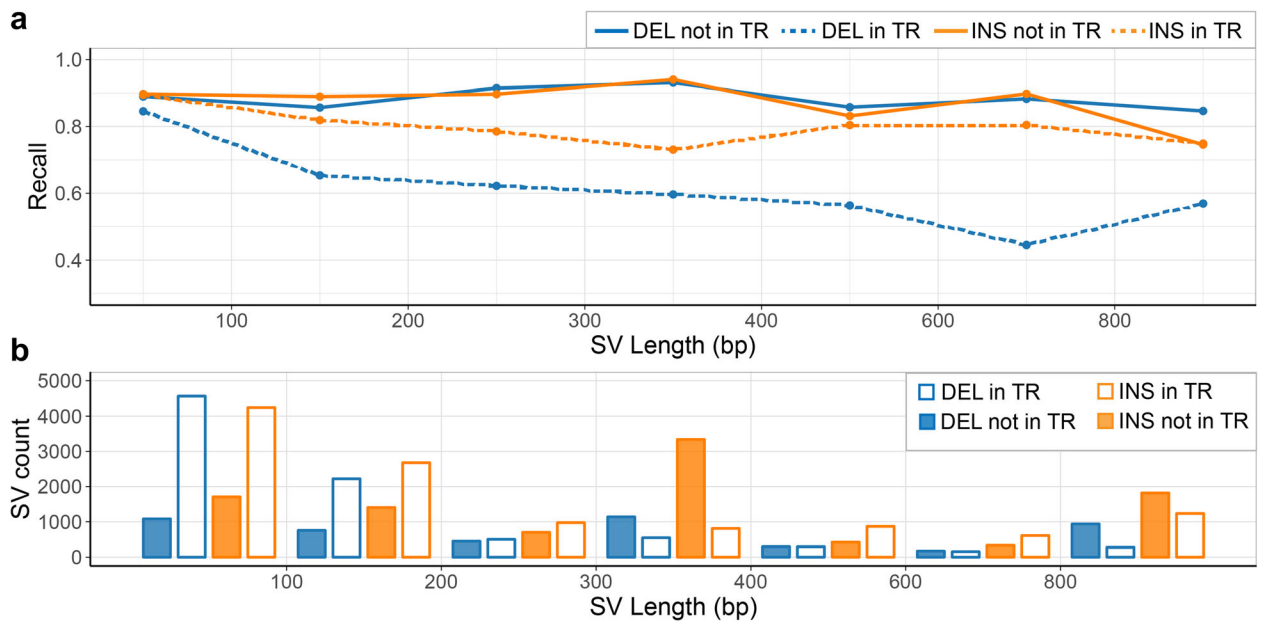

Fig. 4 The impact of TRs on SV recall. **a** Estimated Paragraph recall from LRGT, partitioned by SV length and grouped by their positioning with TRs. **b** LRGT SV count partitioned by length and grouped by their positionings with TRs

reduced *F*-scores: Paragraph (0.64 vs. 0.88), Delly Genotyper (0.58 vs. 0.80), and SVTyper (0.42 vs. 0.82). The three de novo callers have a deletion recall of 0.15–0.20 in the clustered SVs, much lower than their recall of 0.61–0.64 in LRGT.

Population-scale genotyping across 100 diverse human genomes

A likely use case for Paragraph will be to genotype SVs from a reference catalog for more accurate assessment in a population or association studies. To further test and demonstrate Paragraph in this application, we genotyped our LRGT SVs in 100 unrelated individuals (not including NA24385, NA12878, or NA24631) from the publicly available Polaris sequencing resource (https://github.com/Illumina/Polaris). This resource consists of a mixed population of 46 Africans (AFR), 34 East Asians (EAS), and 20 Europeans (EUR). All of these samples were sequenced on Illumina HiSeq X platforms with 150 bp paired-end reads to at least 30-fold depth per sample.

Most deletions occur at a low alternative allele frequency (AF) in the population, whereas there is a gradually decreasing number of deletions at progressively higher AF. Over half of the insertions also occur at a low AF, but there is a sizeable number of insertions with very high AF or even fixated (AF = 1) in the population. As been reported previously [12], these high AF insertions are likely to represent defects and/or rare alleles in the reference human genome. Based on the Hardy-Weinberg Equilibrium (HWE) test, we removed 2868 (14%) SVs that are inconsistent with population genetics expectations. The removed SVs chiefly come from the unexpected AF peak at 0.5 (dashed lines in Fig. 5a). Seventy-nine percent of these HWE-failed SVs are within TRs, which are likely to have higher mutation rates and be more variable in the population [34, 35]. SVs that showed more genotyping errors in the discovery samples were more likely to fail the HWE test (Additional file 1: Table S3). For example, while just 9% of the SVs with no genotyping errors failed our HWE test, 40% of the SVs with two genotyping errors in our discovery samples failed our HWE test.

Because these samples are derived from different populations, our HWE test can be overly conservative, although only 962 (5%) of LRGT SVs have significantly different AFs between populations as measured by the test of their Fixation Index (F_{st}) [36]. In the principal component analysis (PCA) of the HWE-passing SVs, the samples are clearly clustered by populations (Fig. 5b). Interestingly, in PCA of the HWE-failed SVs, the samples also cluster by population (Additional file 1: Figure

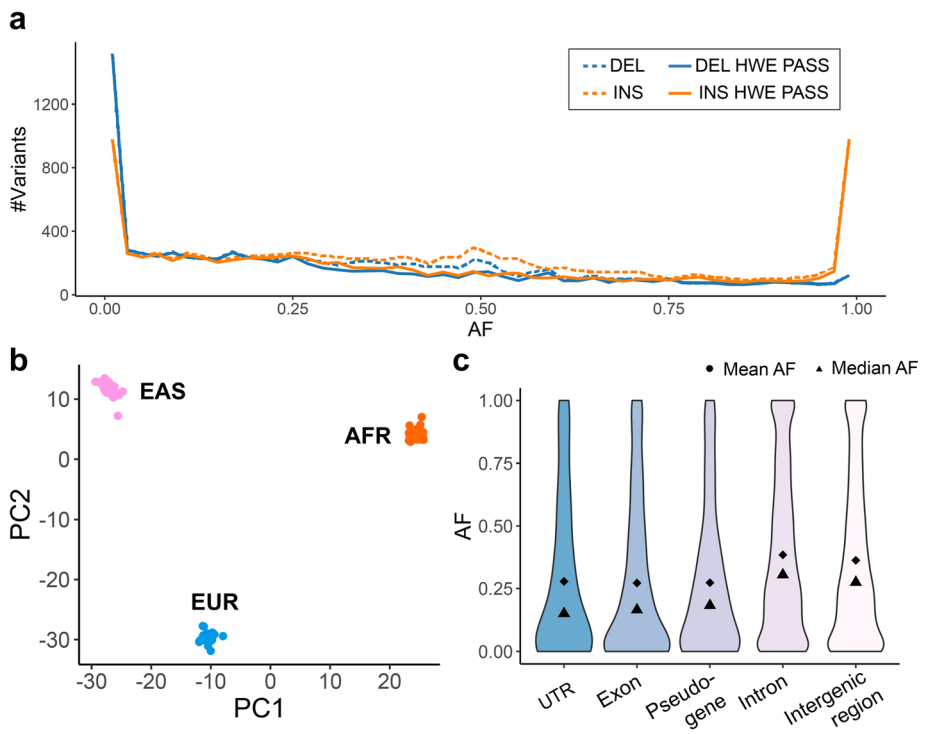

Fig. 5 Population-scale genotyping and function annotation of LRGT SVs. **a** The AF distribution of LRGT SVs in the Polaris 100-individual population. **b** PCA biplot of individuals in the population, based on genotypes of HWE-passing SVs. **c** The AF distribution of HWE-passing SVs in different functional elements. SV count: 191 in UTRs, 554 in exons, 420 in pseudogenes, 9542 in introns, and 6603 in intergenic regions

S5), indicating that some SVs could fail our HWE test because of population substructure rather than poor genotyping performance. Genotyping more samples in each of the three populations will allow better assessment of the genotyping accuracy without the confounding factor of subpopulations that could lead to erroneous HWE deviations.

The population AF can reveal information about the potential functional impact of SVs on the basis of signals of selective pressure. By checking the AFs for SVs in different genomic elements, we found that SVs within exons, pseudogenes, and untranslated regions (UTRs) of coding sequences, in general, have lower AFs than those in intronic and intergenic regions. SVs in introns and intergenic regions have more uniform AF distributions compared to the more extreme AFs in functional elements (UTRs, exons) (Fig. 5c). All these suggest a purifying selection against SVs with potentially functional consequences [25]. Common SVs are more depleted in functional regions than rare SVs, although we do see a few common SVs within exons of genes including *TP73* (AF = 0.09, tumor suppressor gene), *FAM110D* (AF = 0.60, functions to be clarified, possibly related with cell cycle), and *OVGP1* (AF = 0.18, related to fertilization and early embryo development). As the three discovery samples are likely healthy individuals, and these SVs are found at a high frequency in the population, and we expect unlikely to have functional significance.

We also observed 17 exonic insertions fixated (AF = 1) in the population (Additional file 1: Table S4). Since these insertions are present and homozygous in all 100 genotyped individuals, the reference sequence reflects either rare deletion or errors in GRCh38 [37]. Specifically, the 1638-bp exonic insertion in *UBE2QL1* was also reported at high frequency in two previous studies [38, 39]. Particularly, a recent study by TOPMed [39] reported this insertion in all 53,581 sequenced individuals from mixed ancestries. Applying Paragraph to population-scale data will give us a better understanding of common, population-specific, and rare variations and aid in efforts to build a better reference genome.

Discussion

Here, we introduce Paragraph, an accurate graph-based SV genotyper for short-read sequencing data. Using SVs discovered from high-quality long-read sequencing data of three individuals, we demonstrate that Paragraph achieves substantially higher recall (0.84 for deletions and 0.88 for insertions) compared to three commonly used genotyping methods (highest recall at 0.76 for deletions across the genome) and three commonly used de novo SV callers (highest recall of 0.64 for deletions). Of particular note, Paragraph and Manta were the only two methods that worked for both deletions and insertions, and based on our test data, Paragraph achieved substantially higher recall for insertions compared to Manta (0.88 vs. 0.35).

As highlighted above, a particular strength of Paragraph is the ability to genotype both deletions and insertions genome-wide, including those within complicated regions. While we expect that there are as many insertions as there are deletions in the human population, the majority of the commonly used methods either do not work for insertions or perform poorly with the inserted sequence. In particular, insertions are poorly called by de novo variant callers from short reads. Currently, the most effective method to identify insertions is through discovery with long reads. Once a reference database of insertions is constructed, they can then be genotyped with high accuracy in the population using Paragraph. We expect this will be especially helpful to genotype clinically relevant variants as well as to assess variants of unknown significance (VUS) by accurately calculating AFs in healthy and diseased individuals.

Existing population reference databases for SVs may include many variants that are incorrectly represented. Since errors in the breakpoints may be a limitation for population-scaled SV genotyping, we have quantified the genotyping performance of Paragraph and its correlation with breakpoint accuracy (Fig. 3). Our analysis shows that Paragraph can generally tolerate breakpoint deviation of up to 10 bp in most genomic contexts, although the performance suffers as the breakpoints deviate by more bases. Undoubtedly, recent advances in long-read accuracy will lead to more accurate SV reference databases and thus better performance for Paragraph as a population genotyper.

Paragraph works by aligning and genotyping reads on a local sequence graph constructed for each targeted SV. This approach is different from other proposed and most existing graph methods that create a single whole-genome graph and align all reads to this large graph [18, 40]. A whole-genome graph may be able to rescue reads from novel insertions that are misaligned to other parts of the genome in the original linear reference; however, the computational cost of building such a graph and performing alignment against this graph is very high. Adding variants to a whole-genome graph is also a very involved process that typically requires all reads to be realigned. Conversely, the local graph approach applied in Paragraph is not computationally intensive and can easily be adapted into existing secondary analysis pipelines. The local graph approach utilized by Paragraph also scales well to population-level studies where large sets of variants identified from different resources can be genotyped rapidly (e.g., 1000 SVs can be genotyped in 1 sample in 15 min with a single thread) and accurately in many samples.

In this study, we demonstrated that Paragraph can accurately genotype single SVs that are not confounded by the presence of nearby SVs (Table 1, Additional file 1: Table S2). Though, of the SVs identified in these three samples, almost half (48%) occurred in the presence of one or more different SVs. The current version of Paragraph only genotypes one SV per locus though we are actively working on the algorithm to consider and test the ability to annotate overlapping SVs and genotype them simultaneously. In addition, it will be equally important to create a more complete catalog of SVs in these highly variable loci so that the entire complexity can be encoded into the graph.

The primary use case for Paragraph will be to allow investigators to genotype previously identified variants with high accuracy. This could be applied to genotype known, medically relevant SVs in precision medicine initiatives or to genotype SVs from a reference catalog for more accurate assessment in a population or association study. Importantly, the catalog of both medically important SVs and population-discovered SVs will continue to evolve over time and Paragraph will allow scientists to genotype these newly identified variants in historical sequence data. Certainly, the variant calls for both small (single sample) and large (population-level) sequencing studies can continue to improve as our knowledge of population-wide variation becomes more comprehensive and accurate.

Conclusions

Paragraph is an accurate SV genotyper for short-read sequencing data that scales to hundreds or thousands of samples. Paragraph implements a unified genotyper that works for both insertions and deletions, independent of the method by which the SVs were discovered. Thus, Paragraph is a powerful tool for studying the SV landscape in populations, human or otherwise, in addition to analyzing SVs for clinical genomic sequencing applications.

Methods

Graph construction

In a sequence graph, each node represents a sequence that is at least one nucleotide long and directed edges define how the node sequences can be connected together to form complete haplotypes. Labels on edges are used to identify individual alleles or haplotypes through the graph. Each path represents an allele, either the reference allele, or one of the alternative alleles. Paragraph currently supports three types of SV graphs: deletion, insertion, and blockwise sequence swaps. Since we are only interested in read support around SV breakpoints, any node corresponding to a very long nucleotide sequence (typically longer than two times the average read length) is replaced with two shorter nodes with sequences around the breakpoints.

Graph alignment

Paragraph extracts reads, as well as their mates (for paired-end reads), from the flanking region of each targeted SV in a Binary Alignment Map (BAM) or CRAM file. The default target region is one read length upstream of the variant starting position to one read length downstream of the variant ending position, although this can be adjusted at runtime. The extracted reads are realigned to the pre-constructed sequence graph using a graph-aware version of a Farrar's Striped Smith-Waterman alignment algorithm implemented in GSSW library [41] v0.1.4. In the current implementation, read pair information is not used in alignment or genotyping. The algorithm extends the recurrence relation and the corresponding dynamic programming score matrices across junctions in the graph. For each node, edge, and graph path, alignment statistics such as mismatch rates and graph alignment scores are generated.

Only uniquely mapped reads, meaning reads aligned to only one graph location with the best alignment score, are used to genotype breakpoints. Reads used in genotyping must also contain at least one kmer that is unique in the graph. Paragraph considers a read as supporting a node if its alignment overlaps the node with a minimum number of bases (by default 10% of the read length or the length of the node, whichever is smaller). Similarly, for a read to support an edge between a pair of nodes means its alignment path contains the edge and supports both nodes under the above criteria.

Breakpoint genotyping

A breakpoint occurs in the sequence graph when a node has more than one connected edges. Considering a breakpoint with a set of reads with a total read count R and two connecting edges representing haplotype h_1 and h_2, we define the read count of haplotype h_1 as R_{h1} and haplotype h_2 as R_{h2}. The remaining reads in R that are mapped to neither haplotype are denoted as $R_{\neq h1,\ h2}$.

The likelihood of observing the given set of reads with the underlying breakpoint genotype $G_{h1/h2}$ can be represented as:

$$p(R \mid G_{h1/h2}) = p(R_{h1}, R_{h2} \mid G_{h1/h2}) \times p(R_{\neq h1,h2} \mid G_{h1/h2}) \quad (1)$$

We assume that the count of the reads for a breakpoint on the sequence graph follows a Poisson distribution with parameter λ. With an average read length l, an average sequencing depth d, and the minimal overlap of m bases (default: 10% of the read length l) for the

criteria of a read supporting a node, the Poisson parameter can be estimated as:

$$\lambda = d \times (l - m)/l \quad (2)$$

When assuming the haplotype fractions (expected fraction of reads for each haplotype when the underlying genotype is heterozygous) of h_1 and h_2 are μ_{h1} and μ_{h2}, the likelihood under a certain genotype, $p(R_{h1}, R_{h2} \mid G_{h1/h2})$, or the first term in Eq. (1), can be estimated from the density function *dpois*() of the underlying Poisson distribution:

$$p(R \mid G_{h1/h2}) = dpois(R_{h1}, \lambda \times \mu_{h1}) \times dpois(R_{h2}, \lambda \times \mu_{h2}) \quad (3)$$

If h_1 and h_2 are the same haplotypes, the likelihood calculation is simplified as:

$$p(R \mid G_{h1/h1}) = dpois(R_{h1}, \lambda(1-\varepsilon)) \quad (4)$$

where ε is the error rate of observing reads supporting neither h_1 nor h_2 given the underlying genotype $G_{h1/h2}$. Similarly, the error likelihood, $p(R_{\neq h1,\ h2} \mid G_{h1/h2})$, or the second term in eq. (1), can be calculated as:

$$p(R_{\neq h1,h2} \mid G_{h1/h2}) = dpois(R_{\neq h1,h2}, \lambda \times \varepsilon) \quad (5)$$

Finally, the likelihood of observing genotype $G_{h1/h2}$ under the observed reads R can be estimated under a Bayesian framework:

$$p(G_{h1/h2} \mid R) \sim p(G_{h1/h2}) \times p(R \mid G_{h1/h2}) \quad (6)$$

The prior $P(G_{h1/h2})$ can be pre-defined or calculated using a helper script in Paragraph repository that uses the expectation-maximization algorithm to estimate genotype likelihood-based allele frequencies under the Hardy-Weinberg Equilibrium across a population [42].

SV genotyping

We perform a series of tests for the confidence of breakpoint genotypes. For a breakpoint to be labeled as "passing," it must meet all of the following criteria:

1. It has more than one read aligned, regardless of which allele the reads were aligned to.
2. The breakpoint depth is not significantly high or low compared to the genomic average (p value is at least 0.01 on a two-sided Z test).
3. The Phred-scaled score of its genotyping quality (derived from genotype likelihoods) is at least 10.
4. Based on the reads aligned to the breakpoint, regardless of alleles, the Phred-scaled p value from FisherStrand [43] test is at least 30.

If a breakpoint fails one or more of the above tests, it will be labeled as a "failing" breakpoint. Based on the test results of the two breakpoints, we then derive the SV genotype using the following decision tree:

1. If two breakpoints are passing:
   (a) If they have the same genotype, use this genotype as the SV genotype.
   (b) If they have different genotypes, pool reads from these two breakpoints and perform the steps in the "Breakpoint genotyping" section again using the pooled reads. Use the genotype calculated from the pooled reads as the SV genotype.
2. If one breakpoint is passing and the other one is failing:
   (a) Use the genotype from the passing breakpoint as the SV genotype.
3. If two breakpoints are failing:
   (a) If the two breakpoints have the same genotype, use this genotype as the SV genotype
   (b) If two breakpoints have different genotypes, follow the steps in 1b.

Note that for 1b and 2b, as we pool reads from two breakpoints together, the depth parameter d in Eq. (2) needs to be doubled, and reads that span two breakpoints will be counted twice. We also set a filter label for the SV after this decision tree, and this filter will be labeled as passing only when the SV is genotyped through decision tree 1a. SVs that fail the passing criteria 1 and 2 for any one of its breakpoints were considered as reference genotypes in the evaluation of Paragraph in the main text.

Sequence data

The CCS data for NA12878 (HG001), NA24385 (HG002), and NA24631 (HG005) are available at the GiaB FTP (ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/). These samples were sequenced to an approximate 30× depth with an average read length of 11 kb on the PacBio Sequel system. We realigned reads to the most recent human genome assembly, GRCh38, using pbmm2 v1.0.0 (https://github.com/PacificBiosciences/pbmm2). Pacbio CLR data of NA24385 [11] were sequenced to 50× coverage on a PacBio RS II platform, and reads were aligned to GRCh38 using NGMLR [10] v0.2.7.

To test the performance of the methods on short-read data, we utilized three matching samples that were sequenced using TruSeq PCR-free protocol on Illumina platforms with 150 bp paired-end reads: 35× (NA24385)

on HiSeq X, 64× (NA12878), and 48× (NA24631) on NovaSeq 6000. Reads were mapped to GRCh38 using the Issac aligner [44]. To estimate the recall of Paragraph in samples of lower depth, we downsampled the 35× NA24385 data to different depths using SAMtools [45]. To estimate the recall of Paragraph in 100 bp and 75 bp reads, we trimmed the 150-bp reads from their 3′ end in the downsampled NA24385 data.

Long-read ground truth and performance evaluation

SVs were called from the CCS long-read data of the three samples using PBSV v2.0.2 (https://github.com/PacificBiosciences/pbsv). When merging SVs across samples, we define deletions as "different" if their deleted sequences have less than 80% reciprocal overlap; we define insertions as "different" if their breakpoints are more than 150 bp apart, or their insertion sequences have less than 80% of matching bases when aligning against each other using the Smith-Waterman algorithm. After merging, we obtained 41,186 unique SVs. From these unique SVs, we excluded 1944 from chromosome X or Y, 53 SVs that had a failed genotype in 1 or more samples, and 480 SVs where a nearby duplication was reported in at least 1 sample. In the remaining 38,709 unique SVs, 20,108 have no nearby SVs within 150 bp upstream and downstream and these SVs were used as LRGT to test the performance of Paragraph and other methods.

For each method, we define a variant as a true positive (TP) if the LRGT data also has a call in the same sample and a false positive (FP) if the LRGT did not call a variant in that sample. For each genotyper, we estimate its recall as the count of its TPs divided by the count of alternative genotypes in LRGT. We calculate the precision of each method as its TPs divided by its TPs plus FPs. Variants identified by the de novo methods (Manta, Lumpy, and Delly) may not have the same reference coordinates or insertion sequences as the SVs in LRGT. To account for this, we matched variants from de novo callers and SVs in LRGT using Illumina's large-variant benchmarking tool, Wittyer (v0.3.1). Wittyer matches variants using centered-reciprocal overlap criteria, similar to Truvari (https://github.com/spiralgenetics/truvari) but has better support for different variant types and allows stratification for variant sizes. We set parameters in Wittyter as "--em simpleCounting --bpd 500 --pd 0.2, " which means for two matching variants, their breakpoint needs to be no more than 500 bp apart from each other, and if they are deletions, their deleted sequences must have no less than 80% reciprocal overlap.

Estimation of breakpoint deviation

From CLR NA24385, SVs were called using the long-read SV caller, Sniffles [10], with parameters "--report-seq -n -1" to report all supporting read names and insertion sequences. Additional default parameters require 10 or more supporting reads to report a call, and require variants to be at least 50 bp in length. Insertion calls were refined using the insertion refinement module of CrossStitch (https://github.com/schatzlab/crossstitch), which uses FalconSense, an open-source method originally developed for the Falcon assembler [46] and is also used as the consensus module for Canu [47].

We used a customized script to match calls between the CLR and LRGT SVs of NA24385. A deletion from the CLR data is considered to match a deletion in LRGT if their breakpoints are no more than 500 bp apart and their reciprocal overlap length is no less than 60% of their union length. An insertion from the CLR data is considered to match an insertion in LRGT if their breakpoints are no more than 500 bp apart. Base pair deviations between insertion sequences were calculated from the pairwise alignment method implemented the python module biopython [48].

Population genotyping and annotation

The 100 unrelated individuals from the Polaris sequencing resource (https://github.com/Illumina/Polaris) were sequenced using TruSeq PCR-free protocol on Illumina HiSeq X platforms with 150 bp paired-end reads. Each sample was sequenced at an approximate 30-fold coverage. We genotyped the LRGT SVs in each individual using Paragraph with default parameters.

For each SV, we used Fisher's exact test to calculate its Hardy-Weinberg p values [49]. SVs with p value less than 0.0001 were considered as HWE-failed. We used dosage of HWE-passing SVs to run PCA, which means 0 for homozygous reference genotypes and missing genotypes, 1 for heterozygotes, and 2 for homozygous alternative genotypes.

We used the annotation tracks from the UCSC Genome Browser to annotate SVs in LRGT. We define an SV as "within TR" if its reference sequence is completely within one or more TRF tracks. We categorized an SV as functional if it overlaps with one or more functional tracks. We used the ENCODE Exon and PseudoGene SupportV28 track for exons, IntronEst for introns, and ENCFF824ZKD for UTRs. SVs that overlap with any functional track SVs that do not overlap with any of these tracks were annotated as intergenic.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1909-7.

Additional file 1: Figure S1. Examples of singleton SVs and clustered SVs. **Figure S2.** Estimated precision of different genotypers, partitioned by SV length. **Figure S3.** Recall under different read lengths and depths.

Figure S4. The impact of recall when tested SVs include errors in their breakpoints. **Figure S5.** PCA biplot of the Polaris population using only HWE-failed SVs. **Table S1.** Estimated genotype concordance of genotypers. **Table S2.** Performance of different methods in clustered SVs. **Table S3.** SVs with genotyping errors in the three discovery samples. **Table S4.** List of fixated exonic SVs.

Additional file 2. Review history.

Acknowledgements
We thank Mitch Bekritsky and Camilla Colombo for their support in the sequence data, and Chi Kent Ho and Yinan Wan for their help in evaluating the genotyping performance of different methods.

Review history
The review history is available as Additional file 2.

Peer review information
Andrew Cosgrove was the primary editor for this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
PK, SC, ED, RP, and FS developed the algorithms and wrote the source code. MAE, SC, FJS, and MCS designed the manuscript framework. SC performed the analysis, prepared the figures, and wrote the manuscript. FJS, RMS, and MK prepared the test set on long-read data. MAE, DRB, FJS, MCS, RMS, PK, and ED edited the manuscript. All authors read and approved the manuscript.

Funding
This work was supported in part by the National Science Foundation (DBI-1350041 to MCS) and the US National Institutes of Health (R01-HG006677 to MCS and UM1 HG008898 to FJS).

Availability of data and materials
Paragraph software is publicly available on GitHub under Apache 2.0 license [50]. All analysis in the manuscript was performed under version 2.3 [51]. For NA12878, NA24385, and NA24631, the PacBio Sequel II data is deposited in SRA under PRJNA540705 [52], PRJNA529679 [53], and PRJNA540706 [54], and the Illumina data is deposited in ENA under PRJEB35491 [55]. The Illumina Polaris sequencing data is deposited in ENA under PRJEB20654 [56].

Ethics approval and consent to participate
Not applicable.

Competing interests
SC, PK, ED, RP, DRB, and MAE are or were employees of Illumina, Inc., a public company that develops and markets systems for genetic analysis. FJS has sponsored travel granted from Pacific Biosciences and Oxford Nanopore and is a receiver of the SMRT Grant from Pacific Biosciences in 2018. The other authors declare that they have no competing interests.

Author details
[1]Illumina Inc, 5200 Illumina Way, San Diego, CA, USA. [2]Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, UK. [3]Novartis Pharma AG, Basel, Switzerland. [4]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [6]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA.

Received: 28 May 2019 Accepted: 2 December 2019
Published online: 19 December 2019

References
1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97.
3. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. Expert Rev Mol Med. 2010;12:e8.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
5. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016;17:507–22.
6. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19:1117–23.
7. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20:265–72.
8. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010;20:1165–73.
9. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 2013;23:121–8.
10. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
11. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025
12. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. Cell; 2019; Available from: https://doi.org/10.1016/j.cell.2018.12.019
13. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. bioRxiv. 2018 [cited 2019 Jan 25]. p. 193144. Available from: https://www.biorxiv.org/content/early/2018/06/13/193144.abstract
14. Chander V, Gibbs RA, Sedlazeck FJ. Evaluation of computational genotyping of structural variation for clinical diagnoses [Internet]. GigaScience. 2019; Available from: https://doi.org/10.1093/gigascience/giz110.
15. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27:677–85.
16. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966–8.
17. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
18. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9.
19. Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet; 2019; Available from: https://doi.org/10.1038/s41588-018-0316-4.
20. Antaki D, Brandler WM, Sebat J. SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics. 2018;34:1774–7.
21. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
22. Abel HJ, Duncavage EJ. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet. 2013;206:432–40.
23. Sibbesen JA, Maretty L, Danish Pan-Genome Consortium, Krogh A. Accurate genotyping across variant classes and lengths using variant graphs. Nat Genet. 2018;50:1054–9.
24. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. G3. 2015;5:931–41.
25. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75.
26. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009;25:3207–12.
27. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27:665–76.
28. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019; Available from: https://doi.org/10.1093/bioinformatics/btz431.

29. Zook J, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, et al. Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. Available from: https://doi.org/10.1101/281006. Accessed 10 Dec 2019.
30. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. bioRxiv. 2019 [cited 2019 Jan 25]. p. 519025. Available from: https://www.biorxiv.org/content/early/2019/01/13/519025.abstract
31. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Available from: https://doi.org/10.1101/159228
32. Bakhtiari M, Shleizer-Burko S, Gymrek M, Bansal V, Bafna V. Targeted genotyping of variable number tandem repeats with adVNTR. Genome Res. 2018;28:1709–19.
33. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
34. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27:1895–903.
35. Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res, 904. 2014;24:1894.
36. Weir BS, Ott J. Genetic data analysis II. Trends Genet. 1997;13:379.
37. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.
38. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30–5.
39. Taliun D, Harris DN, Kessler MD, Carlson J. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv; 2019; Available from: https://www.biorxiv.org/content/10.1101/563866v1.abstract
40. Hickey G, Heller D, Monlong J, Sibbesen JA, Siren J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. bioRxiv. 2019 [cited 2019 Sep 10]. p. 654566. Available from: https://www.biorxiv.org/content/10.1101/654566v1.abstract
41. Zhao M, Lee W-P, Garrison EP, Marth GT. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. PLoS One. 2013;8:e82138.
42. Garthwaite PH, Jolliffe IT, Jolliffe IT, Jones B. Statistical inference. Oxford: Oxford University Press; 2002.
43. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
44. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29:2041–3.
45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
46. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13:1050–4.
47. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation . Genome Research. 2017. 722–36. Available from: https://doi.org/10.1101/gr.215087.116
48. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
49. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76:887–93.
50. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, Eberle MA. Paragraph: a suite of graph-based genotyping tools. Github. 2019; https://github.com/Illumina/paragraph.
51. Chen S, et al. Paragraph v2.3. Zenodo. 2019. https://doi.org/10.5281/zenodo.3440238. Accessed 10 Dec 2019.
52. Pacific Biosciences. WGS of HG001/NA12878 with PacBio CCS on the Sequel II System. 2019; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA540705. Accessed 10 Dec 2019.
53. Wenger AM, et al. Highly-accurate long-read sequencing of HG002/NA24385. 2019; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA529679. Accessed 10 Dec 2019.
54. Pacific Biosciences. WGS of HG005/NA24631 with PacBio CCS on the Sequel II System. 2019; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA540706. Accessed 10 Dec 2019.
55. Illumina Inc. WGS for Paragraph SV assessment. 2019; https://www.ebi.ac.uk/ena/data/view/PRJEB35491. Accessed 10 Dec 2019.
56. Illumina Inc. Polaris HiSeq X Diversity Cohort. 2019; https://www.ebi.ac.uk/ena/data/view/PRJEB20654. Accessed 10 Dec 2019.

Publisher's Note