bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Sai Chen[1,7] (schen6@illumina.com)

Peter Krusche[2,3,7](pkrusche@gmail.com)

Egor Dolzhenko[1] (edolzhenko@illumina.com)

Rachel M. Sherman[4] (rsherman@jhu.edu)

Roman Petrovski[2] (RPetrovski@illumina.com)

Felix Schlesinger[1] (fschlesinger@illumina.com)

Michael Kirsche[4] (mkirsche@jhu.edu)

David R. Bentley[2] (DBentley@illumina.com)

Michael C. Schatz[4,5] (mschatz@cs.jhu.edu)

Fritz J. Sedlazeck[6,8] (fritz.sedlazeck@bcm.edu)

Michael A. Eberle[1,8,*] (meberle@illumina.com)

[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

[7] These authors contributed equally to this work

[8] These authors contributed equally to this work

* Correspondence: Michael A. Eberle (meberle@illumina.com)

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of abbreviations

SV: structural variation **bp**: base pair **TR**: tandem repeat

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, a fast and accurate genotyper that models SVs using sequence graphs and SV annotations produced by a range of methods and technologies. We demonstrate the accuracy of Paragraph on whole genome sequence data from a control sample with both short and long read sequencing data available, and then apply it at scale to a cohort of 100 samples of diverse ancestry sequenced with short-reads. Comparative analyses indicate that Paragraph has better accuracy than other existing genotypers. The Paragraph software is open-source and available at https://github.com/Illumina/paragraph

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

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (10-50kbp) 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 for the samples remains available they can be re-analyzed 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 the improved detection of the variants that are added to these reference databases.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Targeted genotyping of SVs using short-read sequencing data still remains an open problem. 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[14–16]. In addition, insertions are generally more difficult to detect than deletions using short-read technology, and thus are usually genotyped with lower accuracy or are completely excluded by these methods[17,18]. Finally, consistently genotyping SVs across many individuals is difficult because most existing genotypers only support single-sample SV calling.

Here, we present a fast 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 evaluate systematically how reads align across the breakpoints of the candidate variant relative to the reference sequence. Unlike 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, Paragraph can be universally applied to different methods and the different sequence data types used for constructing the input SV set. 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 the SV) so that variants can be accurate even when variants are clustered together[19–22].

We compare Paragraph to several popular SV detection and genotyping methods and show that the performance of Paragraph is an improvement over the other methods tested, in terms of accuracy. Paragraph is able to genotype over 17,000 SVs in a deep coverage (~35x) short-read whole genome dataset, containing approximately equal numbers of deletions and insertions,

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

while achieving recall over 0.80 and precision over 0.95. By comparison the best genotyping method we tested achieved just 0.68 recall and 0.92 precision and only genotyped deletions. The only discovery-based SV caller we tested that could identify both insertions and deletions, had a recall of 0.39. Finally, we showcase the capability to genotype on a population-scale using 100 deep-coverage WGS samples, from which we detect 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.

Results

Graph-based genotyping of structural variations

For each SV defined in an input variant call format (VCF) file, Paragraph constructs a directed acyclic graph containing paths representing the reference sequence and possible alternative alleles (**Figure 1**). Each node represents a sequence that is at least one nucleotide long. Directed edges define how the node sequences can be connected together 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 a graph-based approach, these breakpoints are genotyped independently and the genotype of the SV can be inferred from the genotypes of the individual breakpoints (see **Methods**). 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 SV genotype. Genotyping details are described in the **Methods** section.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

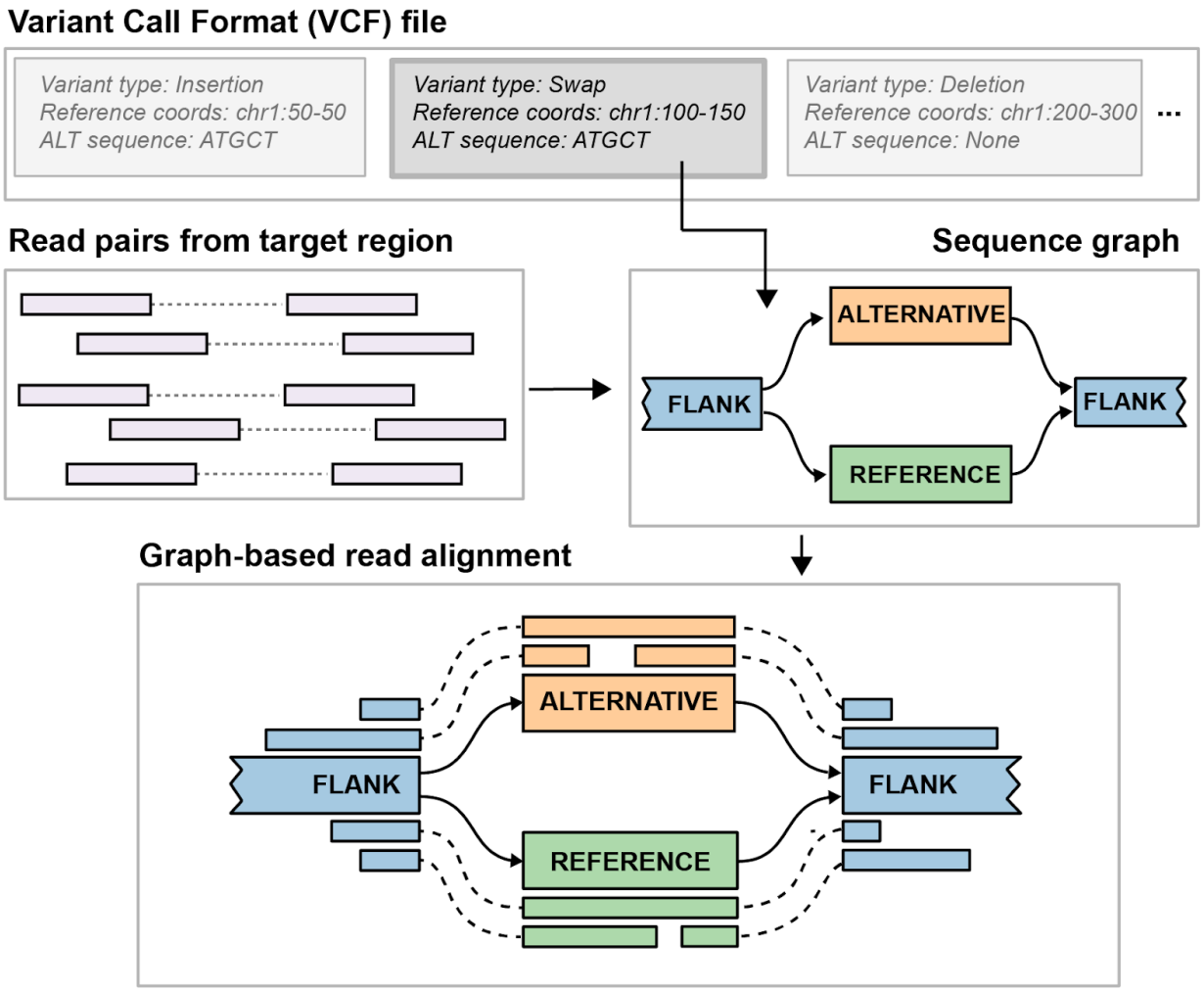

Figure 1. Overview of the SV genotyping workflow implemented in Paragraph. The illustration shows the process to genotype a blockwise sequence swap. Starting from an entry in a VCF file that specifies the SV breakpoints and alternative allele sequences, Paragraph constructs a sequence graph containing all of the paths in the graph (here the reference bases can be replaced with an alternative sequence). Sequence nodes of the graph are shown as the colored rectangles labeled FLANK, ALTERNATIVE and REFERENCE and the edges are shown as the solid arrows that connect the nodes. All reads from the original, linear alignments that aligned near or across the breakpoints are then realigned to the constructed graph. From these alignments, the SV is genotyped as described in the **Methods**.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recall of Insertions and Deletions

First, we calculated the genotyping performance of Paragraph using sequencing data and SVs from the individual HG002 (also known as NA24385) from Genome in a Bottle (GIAB)[11,23]. We used the short-read sequence data to run Paragraph as well as other methods, and used SVs from long read sequence data as the ground truth. The short-read data were generated on an Illumina HiSeqX system to 34.5-fold depth using 150bp paired-end reads. The long read data were generated on a Pacific Biosciences (PacBio) Sequel system using the Circular Consensus Sequencing (CCS) technology[24], to 28-fold coverage with an average read length of 13,500 bp. Previous evaluations showed high recall (0.91) and precision (0.94) for SVs called from PacBio CCS HG002 against the GIAB benchmark dataset[11,24], indicating SVs called from CCS data can be effectively used as ground truth to evaluate the performance of SV genotypers and callers. This long read ground truth (LRGT) set of SVs, all of which are expected to be present in HG002, includes 8,355 deletions and 8,956 insertions. Using this LRGT set, we estimated the performance of Paragraph and two widely-used SV genotypers, SVTyper[15] and Delly Genotyper[16], as well as three methods that independently discover SVs (i.e. *de novo* callers), Manta[17], Lumpy[25] and Delly[16].

Type	**Deletion**						**Insertion**	
	Paragraph	**Delly Genotyper**	**SVTyper (100+ bp)**	**Manta**	**Lumpy (100+ bp)**	**Delly (100+ bp)**	**Paragraph**	**Manta**
#Tested SVs	8,355	8,355	5,372	8,355	5,372	5,372	8,956	8,956
Recall	0.82	0.68	0.35	0.45	0.36	0.21	0.82	0.33

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Recall for different genotypers and *de novo* callers measured against HG002 LRGT. Genotyping/calling was evaluated using a dataset of HG002 with 150 bp paired-end reads sequenced to 34.5-fold depth on an Illumina HiSeqX. Note that SVTyper, Lumpy, and Delly are limited to deletions 100bp or larger so have fewer tested SVs than the other methods.

Paragraph has the highest recall when measured using the LRGT calls: 0.82 for deletions and 0.82 for insertions (**Table 1**). Since recall 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 methods only work for SVs above certain deletion/insertion sizes, we partitioned the SVs by length and further examined the recall performance (**Figure 2**). Excluding the largest deletions (>2,500bp), 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 (>300bp) deletions, and Delly Genotyper has lower recall than these two. For smaller deletions (<300 bp), the recall for Paragraph (0.85) remains high while we observe a large drop in recall for SVTyper (0.12). 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 has consistently high recall across all SV lengths Manta has a much lower recall which drops further for larger insertions.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

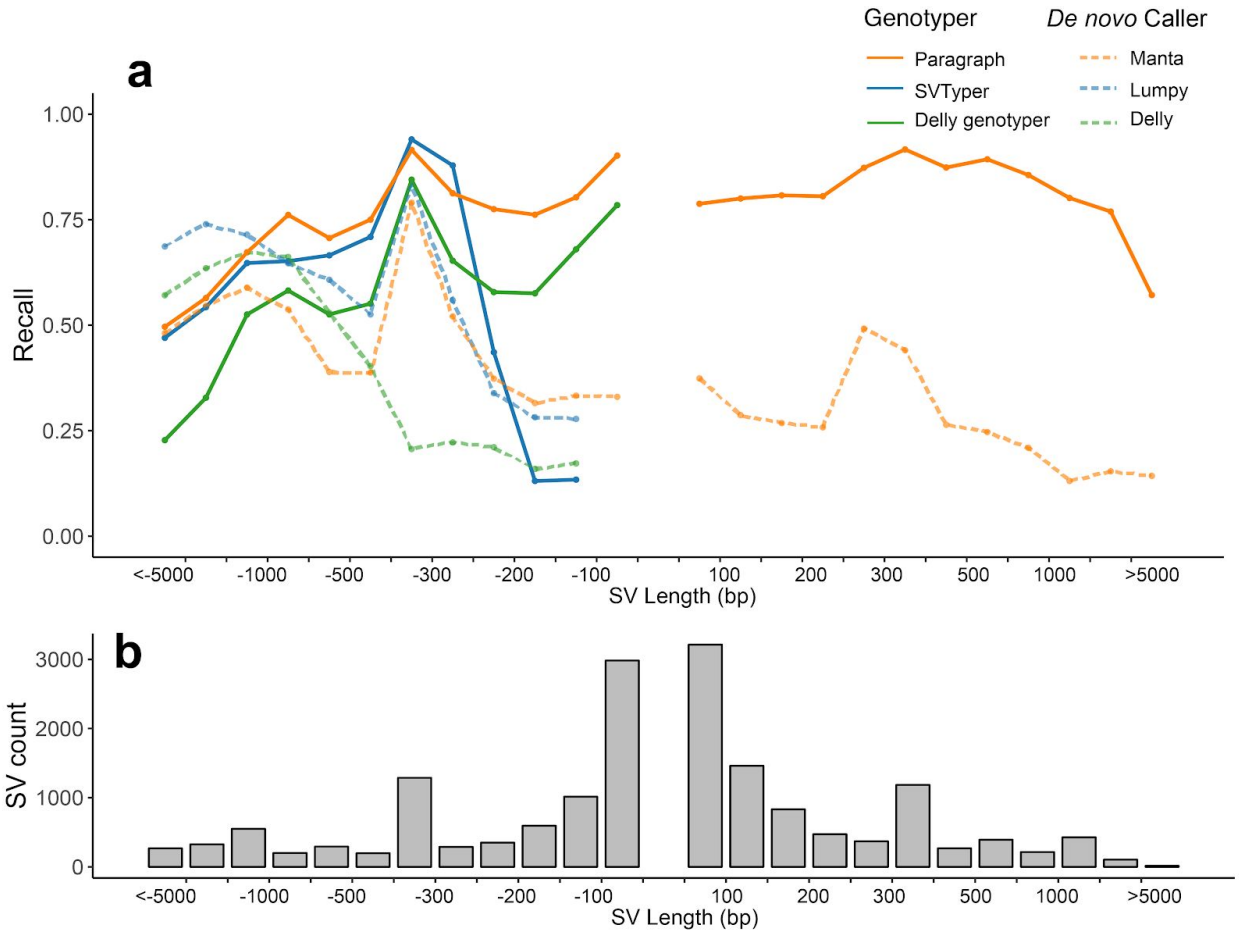

Figure 2. Comparison of recall, partitioned by SV length. HG002 LRGT was used as the input for genotypers and also as the ground truth for estimating recall for all methods. A negative SV length indicates a deletion. Colored lines in (a) show recall of different methods evaluated; Solid grey bars in (b) represent the count of SVs in each size range. The center of the plot is empty since SVs must be at least 50 bp in size.

We tested the recall using high depth data (>30x) with 150bp reads but some studies may use shorter reads and/or lower depths. To quantify how either shorter reads or lower depth will impact genotyping performance, we simulated data of different read lengths and depths by downsampling and trimming reads from the short-read data of HG002. Generally, shorter read lengths are detrimental to recall; reductions in depth have less of a deleterious effect until the depth is below ~20x (**Supplementary Figure 2**).

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genotyping with breakpoint deviations

The LRGT data used above 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 for SV discovery to build up a reference catalog would be to sequence many samples (possibly at lower depth) using contiguous long reads (CLR) rather than CCS technology and derive consensus calls across multiple samples. The high error rates (~10-15%) of CLR may result in errors in the accuracy of the 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 the reads to a sequence graph using stringent parameters, inaccuracies in the breakpoints may decrease recall.

To understand how Paragraph performs with prior SVs that have incorrect breakpoints, we constructed SV calls on HG002 using CLR data that was generated on the PacBio RS II platform. Most (13,058) of the SVs in our LRGT data closely match those generated from the CLR data (see **Methods**). Of these, 29% (3,801) SVs were called identically in both CCS and CLR data but the remaining 2,461 deletions and 6,796 insertions, although in approximately correct locations, had different representations (breakpoints and/or insertion sequence). We made the assumption that CCS breakpoints of our LRGT data are correct, and defined any deviations in the SV descriptions as errors in the CLR breakpoints. For the deletions with different breakpoints, the recall decreased from 0.81 to 0.57 using the breakpoints defined by CCS or CLR breakpoints, respectively. Overall, there is a clear negative trend between recall and the degree of breakpoint deviation: the larger the deviation, the less likely the variant can be genotyped correctly. While deviations of a few base pairs can generally be tolerated without

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

issue, deviations of 20 bp or more reduce recall to around 0.50 (**Figure 3**). For the insertions with differences in breakpoints and/or insertion sequences, the recall decreased from 0.85 to 0.60 using the breakpoints and insertion sequences defined by CCS and CLR data, respectively. We also investigated how inaccurate breakpoints impact insertion genotyping, but there was no clear trend between recall and base-pair deviation in breakpoints.

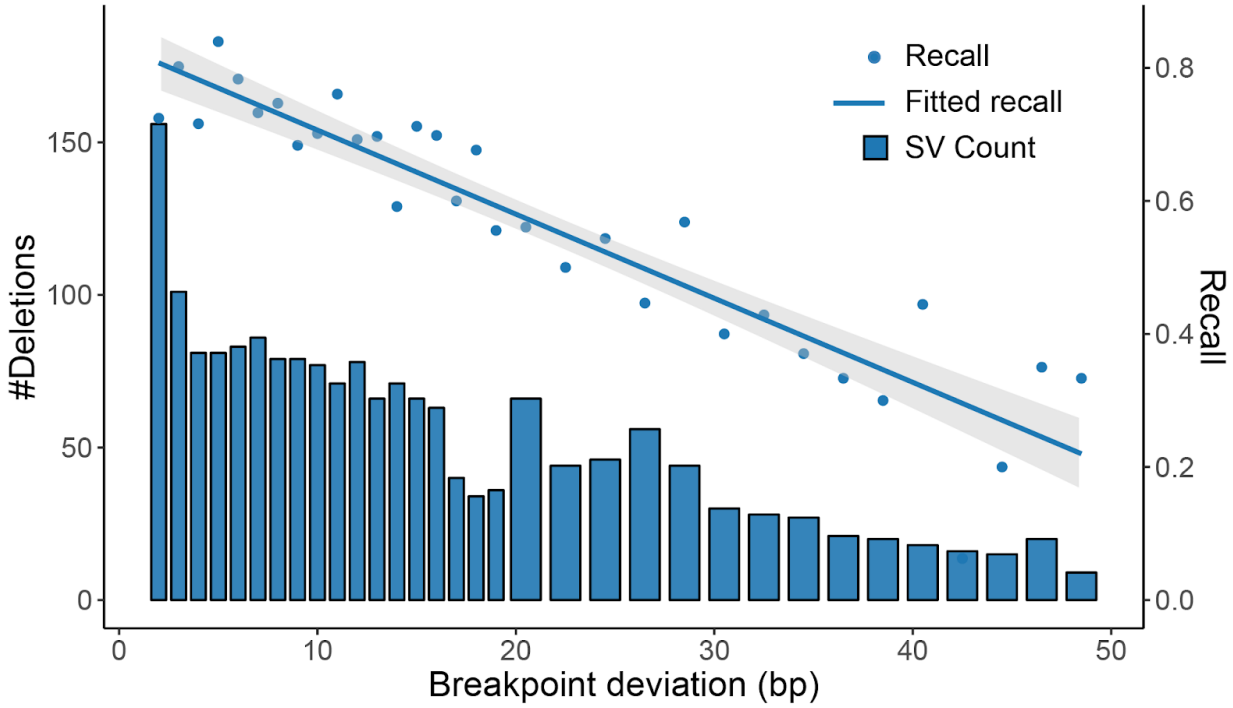

Figure 3. Demonstration of the impact of recall when the SVs include errors in their breakpoints. Breakpoint deviation measures the difference in position between the CLR and CCS based deletion calls. The recall is calculated using the deletions defined by the CLR data. For clarity, breakpoint deviations were binned at 1bp for deviations less than 20bp and at 2bp for deviations larger or equal to 20bp. 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).

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genotyping in tandem repeats

We identified that most of the SVs showing breakpoint deviations between the CCS and CLR calls were within tandem repeats: of the 2,530 deletions with breakpoint deviations, 77% (1,966) were in tandem repeats (TRs). Additionally, the percentage of the deviant variants in TRs increases with increasing deviation in breakpoints: 47% of the SVs with smaller (<=10 bp) deviations are in TRs while 95% of the SVs with larger breakpoint deviations (>20 bp) are in TRs.

To further understand the impact that TRs have on the overall performance we categorized SVs from the CCS data according to repeat context (**Figure 4a**) and found that 76% of Paragraph's false negative deletions and 78% of false negative insertions occur in TRs. Additionally, shorter (<200bp) SVs are much more likely to be TR-originated (>70%) than larger (>1,000bp) SVs (<20%). Separating SVs inside and outside of TRs, we found that the recall is improved for non-TR SVs: 0.89 for deletions and 0.90 for insertions, compared to SVs in TRs that had 0.77 recall for both deletions and insertions.

We then analyzed the TRs by grouping them according to their average in-repeat mismatch rate, which measures the differences between the repeat units within each TR. A lower in-repeat mismatch rate means the TR is composed of highly similar repeated units and indicates a lower sequence complexity, e.g. 0% mismatch rate indicates the repeat is identical through the entire TR. Binned by the in-repeat mismatch rate, we examined the call for deletions and insertions for Paragraph and the *de novo* caller, Manta (**Figure 4b,c**). For deletions, the in-repeat mismatch rate does not have a significant impact on recall for Paragraph but has a

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

negative effect on the recall for Manta (**Figure 4b**). For insertions, as the in-repeat mismatch rate becomes lower, meaning the individual repeat units become more similar to each other, the recall of Paragraph decreases from 0.82 to 0.56 (**Figure 4c**).

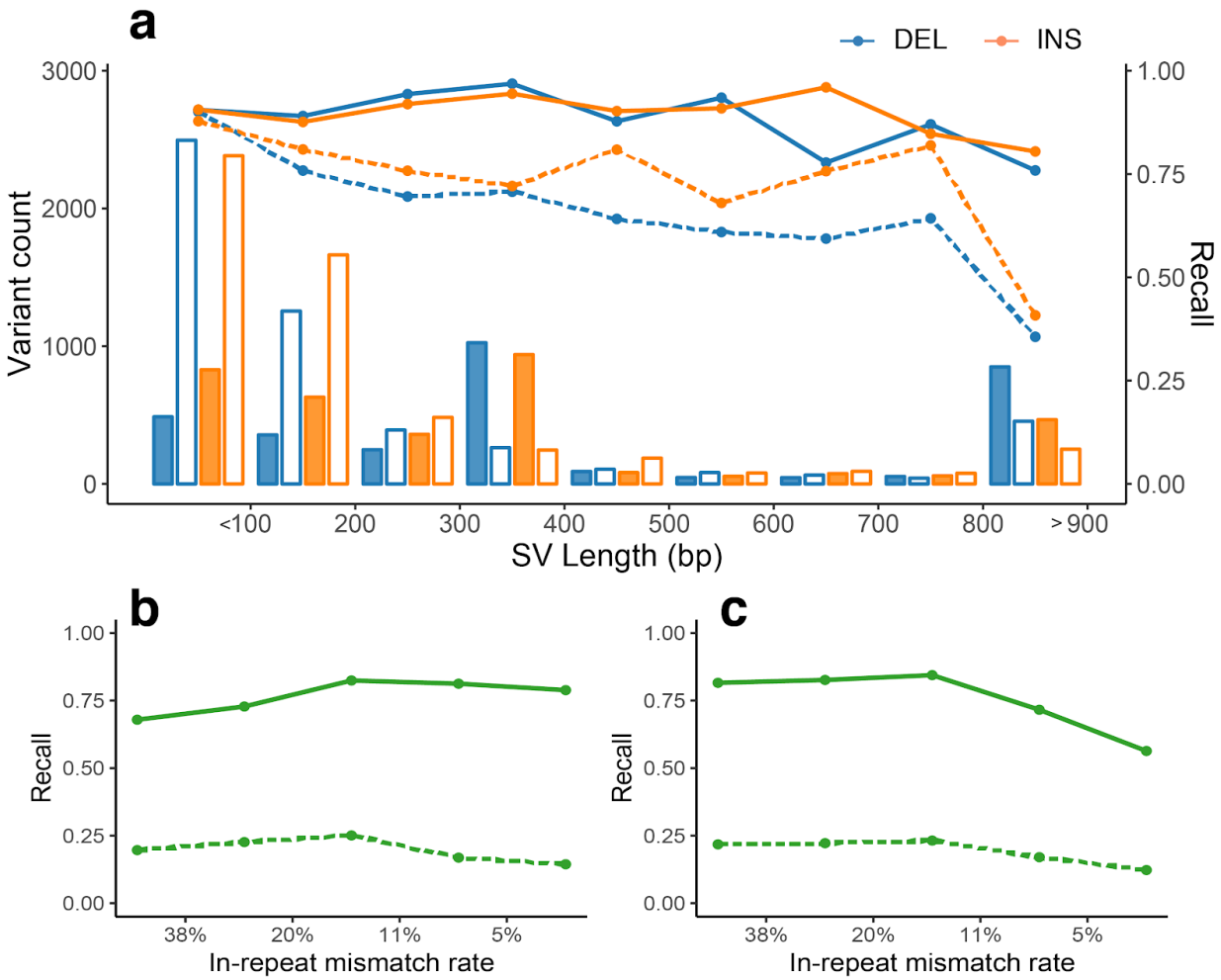

Figure 4. The impact of TRs on SV genotyping performance. (a) SV counts (bars, left axis) and recall (lines, right axis) from HG002 LRGT, binned by SV length. Bars are colored according to SV type (blue for deletions and orange for insertions) and TR status (solid and outlined bars indicate SVs outside of TRs and inside of TRs, respectively). Lines are colored by SV type (blue for deletions and orange for insertions) and TR status (solid and dashed lines indicate SVs outside of TRs and inside of TRs, respectively). Lower panels show comparisons of recall for Paragraph (solid line) and Manta (dashed line) for (b) deletions and (c) insertions within TRs binned according to in-repeat mismatch rate.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Calculating precision

So far, we have measured recall against our LRGT data in the test sample, and investigated the factors that affect genotyping performance. However, applying Paragraph to a sample using SVs identified from a large population will also include genotyping variants that are not present in the test sample. To define such positions, we considered SVs identified from a second sample, ENC002, that was sequenced on PacBio RS II platform as part of the ENCODE project[26]. SVs that were called in ENC002 but are not in our HG002 LRGT data were defined as confident reference positions (**Supplementary Figure S1**). The precision of Paragraph and other genotypers was estimated by genotyping these confident reference positions in the short-read HG002.

Incorporating the 2,366 deletions and 2,855 insertions that occur in ENC002 but not HG002, the precision for Paragraph was estimated as 0.92 for deletions and 0.90 for insertions (**Supplementary Table 1**). Notably, the precision for deletions was 0.10 higher than that of Delly Genotyper (0.80). SVTyper is limited to deletions longer than 100bp, and when estimating precision just on the deletions longer than 100bp, Paragraph has a slightly lower precision (0.96) than SVTyper (0.98) though the recall is much higher for Paragraph (0.89 vs 0.35). Combining recall and precision, Paragraph has the highest F-score for deletions in all of the tested genotypers (0.89 vs 0.78 for Delly Genotyper and 0.52 for SVTyper), and also has a high F-score for insertions (0.89). Nearly all (97%) of false positive (FP) deletions and the majority (78%) of FP insertions are completely within TRs. Of the 48 FP insertions that are outside of TRs: 32 have one or more indels (longer than 10 bp) in the target region; 9 have two or more supporting reads for the insertion in HG002 CCS data and those could be false negatives in SV

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calling from the CCS data; 7 have no evidence of variants in the CCS alignments in the target region, and these FPs are likely to come from alignment artifacts in short-read mapping.

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 study. To further test and demonstrate Paragraph in this application, we genotyped our LRGT variants in 100 unrelated individuals (not including HG002 or ENC002) from the publicly-available Polaris sequencing resource (https://github.com/Illumina/Polaris). This resource consists of a mixed population of 46 individuals from Africa (AFR), 34 from East Asia (EAS) and 20 from Europe (EUR). All of these samples were sequenced with Illumina HiSeq X 150 bp paired-end reads to at least 30x depth.

The population allele frequencies (AFs) for deletions show that most of them occur at a low AF in this sample set, whereas there is a gradually decreasing number of deletions at progressively higher AF. For insertions, most are also found at low AF, but some appear to be fixed in the population. These high-AF insertions are likely to represent defects and/or rare alleles in the reference human genome, as has been reported previously[12]. An interesting feature of the frequencies for both insertions and deletions is a peak around 50% AF that is inconsistent with population genetics expectations (dashed lines in **Figure 5a**). This peak likely indicates genotyping errors such as additional mutations in the population and/or false positives in variable-length TRs that were not detected in our single sample analysis. Filtering the population calls based on Hardy-Weinberg Equilibrium (HWE) p-values, we removed 4,429 (26%) variants, chiefly from the unexpected peak. Interestingly, 79% of these HWE-failed variants lie in TRs, which are likely to have higher mutation rates and be more variable in the

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population[27,28]. Because these samples are derived from different populations, the HWE test can be overly conservative, although only 8% (358) of the HWE-failed SVs have significantly different allele frequencies between populations as measured by their Fixation Index (F_{st})[29]. In contrast, 5% of the HWE-passing SVs had significant F_{st} scores. Genotyping more samples in each of the three populations will allow better filtering of the data without the confounding factor of subpopulations that could lead to erroneous HWE deviations. After filtering out the HWE-failed variants, the samples clearly cluster by population when we used the remaining SVs in a principal component analysis (PCA) (**Figure 5b**). Interestingly, using just the HWE-failed variants, the samples also cluster by population (**Supplementary Figure S3**) indicating that we are filtering some variants based on population substructure rather than poor genotyping performance.

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 the coding sequences, in general, have lower AFs than those in intronic and intergenic regions (**Figure 5c**). The set of SVs in introns and intergenic regions have more uniform AF distributions, while a larger fraction of SVs in functional elements (UTRs, exons) have AF of zero or one. This suggests a purifying selection against SVs with potentially functional consequences, as has been seen in other studies[20]. Common SVs are more depleted in functional regions than rare SVs, although we do see a few common SVs within the 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 HG002 is an apparently healthy individual, and these variants are

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

found at a high frequency in the population, we expect that these variants are unlikely to have functional significance.

We also observed 7 exonic insertions fixated (AF=1) in the population in the following genes: *FOX06*, *UBE2QL1*, *KMT5A*, *FURIN*, *ZNF523*, *SAMD1*, *EDEM2* and *RTN4R*. Since these insertions are present and homozygous in all 100 genotyped individuals, the reference sequence reflects either rare deletion or errors in GRCh38[30]. Specifically, the 1,638 bp exonic insertion in *UBE2QL1* was also reported at high frequency in two previous studies [31,32]. Particularly, a recent study by TOPMed [32] 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.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

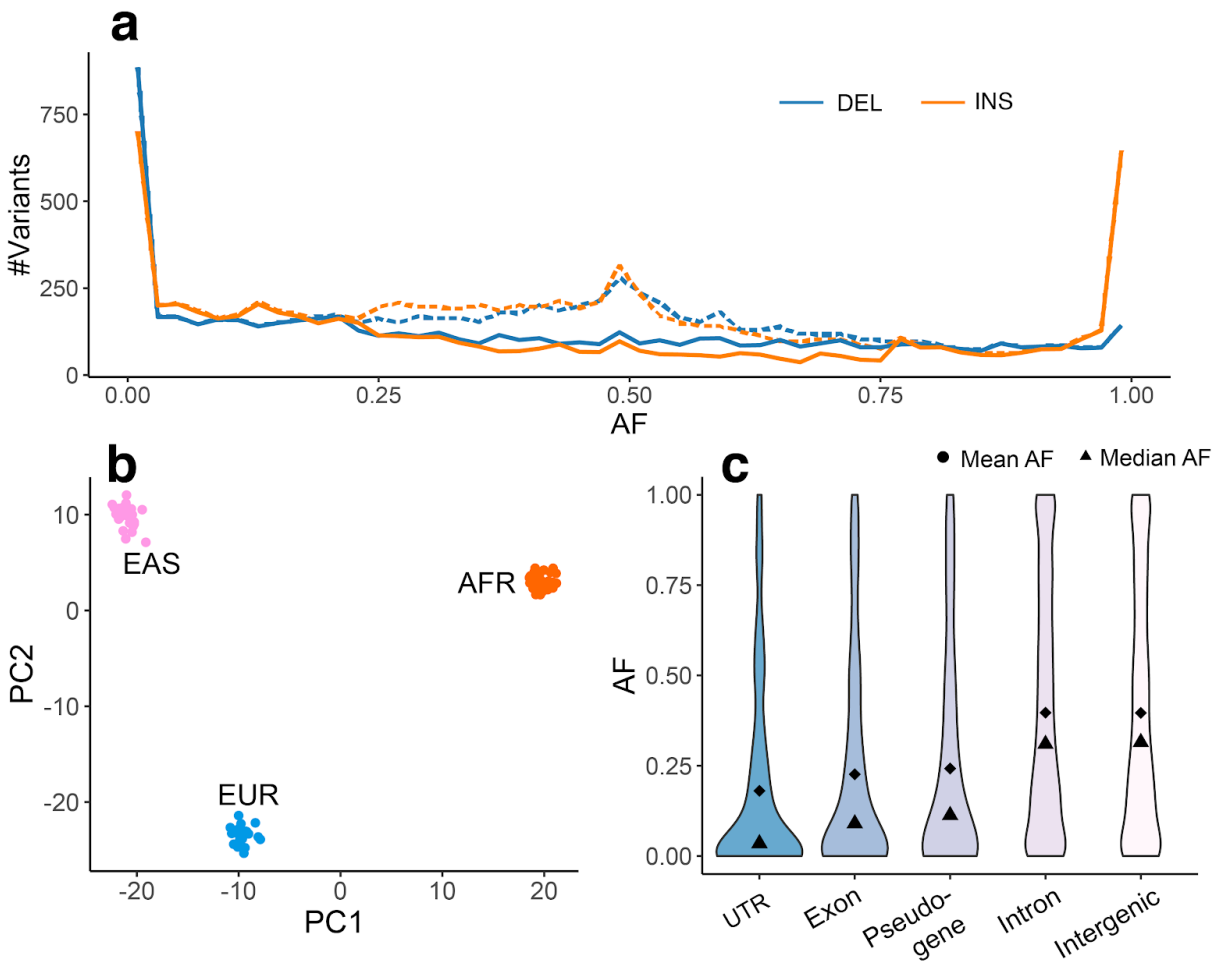

Figure 5. Population information and function annotation of SVs in HG002 LRGT. (a) The AF distribution of these SVs in the 100-individual population, dashed lines for all variants, and solid lines for variants that pass the HWE filter. (b) PCA biplot of individuals in the population, based on genotypes of HWE-passing SVs. (c) The AF distributions of HWE-passing SVs in different functional elements.

Discussion

Here we introduce Paragraph, a fast and accurate graph-based SV genotyper for short-read sequencing data. Using variants discovered from high-quality long read sequencing data, we demonstrate that Paragraph achieves substantially higher recall (0.81) compared to two other commonly used genotyping methods and three commonly used *de novo* SV callers (0.21 and

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

0.68). 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 (0.92 vs 0.49).

As highlighted above, a particular strength of Paragraph is the ability to genotype both deletions and insertions genome-wide, including those within complicated regions such as tandem repeats. 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, because insertions are poorly called by *de novo* variant calling methods, currently the most effective method to identify them 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 allele frequencies 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 relationship with breakpoint accuracy (**Figure 3**). Our analysis shows that Paragraph can generally tolerate breakpoint deviation of up to 10 base pairs in most genomic contexts, although the performance suffers as the breakpoints deviate by larger amounts. 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.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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[33]. 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 and accurately in many samples.

Paragraph does not solve the problem of SV calling across all applications so continued development and improvement of *de novo* methods is essential. For example, cancer genomes include mostly *de novo* SVs that will not be included in any variant catalog. 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 date. Thus, the variant calls for both small (single sample) and large (population-level) sequencing studies can

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

continue to improve as our knowledge of population-wide variation becomes more comprehensive and accurate.

Conclusions

Paragraph is a fast and accurate SV genotyper for short-read sequencing data that scales to hundreds or thousands of samples. Moreover, Paragraph implements a unified genotyper that works for both insertions and deletions, and as 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.

Online 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 graphs for SVs: 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 read length) is replaced with two shorter nodes with sequences around the breakpoints.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Graph alignment

For each SV, Paragraph extracts reads, as well as their mates (for paired-end reads), from the flanking region of the 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. 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, the count of aligned reads, mismatch rate, gap rate, and graph mapping score are computed.

Only uniquely mapped reads, meaning reads aligned to only one graph location with the best 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

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 equation (1), can be estimated from the density function $dpois()$ of the underlying Poisson distribution:

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

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

, 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 equation (1), can be calculated as

$$p(R_{\neq h1,h2} \mid G_{h1/h2}) = dpois(R_{\neq h1,h2}, \lambda(1-\varepsilon)) \tag{5}$$

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

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

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 Hardy-Weinberg Equilibrium across a population[34].

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

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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[35] 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 **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 steps in 1b.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Note that for 1b and 2b, as we pool reads from two breakpoints together, the depth parameter d in equation (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 have filter labels other than "passing" were not excluded from the evaluation of Paragraph in the main text.

Sequence data

PacBio CCS reads for HG002 (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/) were sequenced to an approximate 30x depth with an average read length of 13.5 kb. We re-aligned the reads to the most recent human genome assembly, GRCh38 using NGMLR[10]. Pacbio CLR data from two samples, an Ashkenazi Jewish male, HG002, from Genome in a Bottle[11], and a Caucasian male, ENC002, from the ENCODE project (also identified as ENCDO451RUA; BAM available: http://labshare.cshl.edu/shares/schatzlab/www-data/encode/diploid/2017.10.26/enc002/ENC-002_all_ngm2.7.bam), were sequenced to 50x and 57x coverage, respectively, on an RS II platform. CLR data was aligned to GRCh38 using NGMLR as well.

To test the performance of the methods on short-read data, we utilized an HG002 sample that was sequenced on an Ilumina HiSeq X platform to an average depth of 34.5x, with 150 bp paired-end reads. Reads were mapped to GRCh38 using the Issac aligner[36]. To assess lower sequence depths, we downsampled the short-read data to coverages of 5,10,15,20,25 and 30 using samtools[37]. To obtain the recall of Paragraph in 100bp and 75bp reads, besides downsampling of depths, we trimmed the 150bp reads from their 3' end.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ground truth SVs and confident reference positions

SVs were called from the PacBio CCS and CLR data 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 variant supporting reads to report a call, and require variants be at least 50 bp in length. Insertion calls were refined using the insertion refinement module of CrossStitch (https://github.com/schatzlab/crossstitch). This module uses FalconSense, an open-source method originally developed for the Falcon assembler[38] and is also used as the consensus module for Canu[39].

To estimate breakpoint deviation, we used a customized script to match calls between CLR and CCS data. A deletion from CLR data is considered to match a deletion from CCS data if their breakpoints are no more than 500 bp away and their reciprocal overlap length is no less than 60% of their union length. An insertion from CLR data is considered to match an insertion from CCS data if their breakpoint on the reference sequence is no more than 500 bp away. Base pair deviations between insertion sequences were calculated from the pairwise alignment using the python module biopython pairwise2.

Confident reference positions were defined using SVs from CLR ENC002 and CCS HG002. If a deletion is only observed in ENC002 and no deletion is observed 500 bp upstream or downstream in HG002 with at least 3 supporting reads, this deletion is defined as a confident reference position in HG002. Similarly, if an insertion is only observed in ENC002 and there is no insertion observed in upstream or downstream 200 bp regions in HG002 with at least 3

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

supporting reads, or there is an insertion in HG002 within 200 bp but their insertion sequences are than 25% concordant, this insertion is defined as a confident reference position in HG002.

Calculations of recall and precision

For each genotypers, the recall was calculated as the fraction of SVs in LRGT that were genotyped as non-reference, or the fraction of true positions (TP). The precision was estimated as the fraction of confident reference positions that were genotyped as reference genotypes. Naturally, the confident reference positions that were genotyped as non-reference are the false positions (FP). Thus, the precision is estimated as TP/(TP+FP).

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 or confident reference positions. To account for this we matched variants called by *de novo* methods and those in LRGT or confident reference positions 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 allow 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 distance needs to be no more than 500 bp and if they are deletions, their deleted sequences should have no more than 20% base pair discrepancies.

Population-scale genotyping, filtering and annotation

The 100 unrelated individuals from the Polaris sequencing resource were sequenced using Illumina HiSeq X platform with 150 bp paired-end reads. Each sample was sequenced at an

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

approximate 30-fold coverage. Using SV descriptions from the HG002 LRGT, we genotyped each individual using Paragraph with default parameters.

For each SV, we used Fisher's exact test to calculate its Hardy-Weinberg p-values[40]. SVs with p-value less than 0.001 were excluded from downstream analysis. To run PCA for individual samples, we used the dosage of the SVs, which means 0 for homozygous reference genotypes and missing genotypes, 1 for heterozygotes and 2 for homozygous alternative genotypes.

To identify each SV's overlapping status with TR and their functional impact, we used annotation tracks from the UCSC Genome Browser. We used Encode Exon and PseudoGene SupportV28 track for exons, IntronEst for introns and ENCFF824ZKD for UTRs. SVs that do not overlap with any of these tracks were classified as intergenic. To identify if an SV lies in TRs, we used the Tandem Repeat Finder (TRF) track and we define an SV as TR-originated only when its reference sequence overlaps 100% with one or more TRs.

Acknowledgments

We thank Mitch Bekritsky for his support in obtaining the sequence data, and Chi Kent Ho and Yinan Wan for help calculating performance.

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).

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability of data and materials

Paragraph software is publicly available on https://github.com/Illumina/paragraph. All analysis in the manuscript was performed under version 2.2. The Illumina HiSeq X sequenced HG002 will be available soon. The HG002 SV calls from PacBio CCS data, and the individual genotypes in Polaris cohort, are available in Paragraph GitHub repository.

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 experiments, 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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
2. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
3. Lee, C. & Scherer, S. W. The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.* **12**, e8 (2010).
4. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
5. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
6. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
7. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
8. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
9. Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).
10. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
11. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**, 160025 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* (2019). doi:10.1016/j.cell.2018.12.019
13. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2018). doi:10.1101/193144
14. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
15. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
16. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
17. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
18. Abel, H. J. & Duncavage, E. J. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genet.* **206**, 432–440 (2013).
19. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5**, 931–941 (2015).
20. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75 (2015).
21. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
22. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).
23. Zook, J. *et al.* Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

materials. doi:10.1101/281006

24. Wenger, A. M. *et al.* Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *bioRxiv* 519025 (2019). doi:10.1101/519025
25. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
26. The ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
27. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
28. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
29. Weir, B. S. & Ott, J. Genetic data analysis II. *Trends Genet.* **13**, 379 (1997).
30. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
31. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
32. Taliun, D., Harris, D. N., Kessler, M. D. & Carlson, J. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv* (2019).
33. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
34. Garthwaite, P. H., Jolliffe, I. T., Jolliffe, I. T. & Jones, B. *Statistical Inference*. (Oxford University Press, 2002).
35. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using

bioRxiv preprint doi: https://doi.org/10.1101/635011; this version posted May 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

36. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
39. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017).
40. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).