Mini review

Population-scale genotyping of structural variation in the era of long-read sequencing

Cheng Quan a, Hao Lu a, Yiming Lu a,d,* , Gangqiao Zhou a,b,c,d,*

a Department of Genetics & Integrative Omics, State Key Laboratory of Proteomics, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, Beijing 100850, PR China
b Collaborative Innovation Center for Personalized Cancer Medicine, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu Province 211166, PR China
c Medical College of Guizhou University, Guiyang, Guizhou Province 550025, PR China
d Hebei University, Baoding, Hebei Province 071002, PR China

ABSTRACT

Population-scale studies of structural variation (SV) are growing rapidly worldwide with the development of long-read sequencing technology, yielding a considerable number of novel SVs and complete gap-closed genome assemblies. Herein, we highlight recent studies using a hybrid sequencing strategy and present the challenges toward large-scale genotyping for SVs due to the reference bias. Genotyping SVs at a population scale remains challenging, which severely impacts genotype-based population genetic studies or genome-wide association studies of complex diseases. We summarize academic efforts to improve genotype quality through linear or graph representations of reference and alternative alleles. Graph-based genotypers capable of integrating diverse genetic information are effectively applied to large and diverse cohorts, contributing to unbiased downstream analysis. Meanwhile, there is still an urgent need in this field for efficient tools to construct complex graphs and perform sequence-to-graph alignments.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction ..................................................................................................... 2640
2. Population-scale structural variation studies using a hybrid sequencing strategy ......................................................... 2640
3. Linear representation of the alternative allele ................................................................. 2643
4. Genotyping structural variation in pan-genome graphs ................................................................. 2643
4.1. Pan-genome graph construction. ........................................................................... 2643
4.2. Sequence-to-graph alignment and genotyping models ......................................................... 2644
5. Summary and outlook .......................................................................................... 2645
CRediT authorship contribution statement ............................................................................ 2645
Declaration of Competing Interest ...................................................................................... 2645
Acknowledgements .................................................................................................... 2645
References ................................................................................................................. 2645

* Corresponding authors at: Department of Genetics & Integrative Omics, State Key Laboratory of Proteomics, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, PR China (G. Zhou). Department of Genetics & Integrative Omics, State Key Laboratory of Proteomics, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing, 100850, PR China (Y. Lu).
E-mail addresses: ylu.phd@gmail.com (Y. Lu), zhougq114@126.com (G. Zhou).

https://doi.org/10.1016/j.csbj.2022.05.047
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Structural variation (SV) is arbitrarily defined as chromosomal genomic rearrangements greater than 50 bp, including insertions, duplications, deletions, inversions, and translocations [1–3]. Population-scale studies like Human Genome Structural Variation Consortium (HGSVC) [4], GenomeAD-SV [5], and Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium [6] have found that basic SVs are often nested together to form more complex SVs [7,8]. Compared to the ubiquitous single nucleotide variation (SNV) and small indels, SVs are numerically fewer but larger in size, therefore having a greater impact on DNA sequences and correspondingly on gene expression and protein functionality [9–11].

As a result, SV has received extensive attention in recent studies on genome evolution [10], population diversity [12], demographic history [13], and genetic adaptation [14], bringing new insights into population genetics. In addition, SV can act as a genetic factor underlying disease risk [15] and has already been reported to be involved in the tumorigenesis of various cancers [6,16,17], especially nested SVs in complex genetic backgrounds [18,19].

However, during the rapid development phase of short-read sequencing (SRS) with high base accuracy and relatively low cost, the study of SVs lagged far behind SNVs or indels [2,3]. In contrast to these small variants, SV trends to occur in highly repetitive and polymorphic regions [12], making it more challenging to detect. Most SRS-based methods extract information of discordant read pairs (RP), split-reads (SR), and read-depth (RD) from alignment with the reference genome to infer the existence of breakpoints [20]. Nevertheless, the short readscannot span the entire repetitive sequences, leading to low-quality alignment and false-positive identification of SVs [23,21]. Even in the case of non-repetitive regions, insertions longer than short reads are easily missed because they cannot align correctly with the reference genome [12,22]. Fortunately, the third-generation long-read sequencing (LRS) technologies, such as nanopore sequencing by Oxford Nanopore Technologies and single-molecule real-time sequencing (SMRT) by PacBio, as well as other non-sequence-based long-range technologies [23], such as optical mapping (OM) by Bionano Genomics, have developed rapidly in recent years and revitalized SV studies. LRS is best characterized by a much longer read-length with an average size of 10 kb [22,23], far exceeding the 100–500 bp read-length of SRS. Standard long reads generated via nanopore sequencing (R9.4.1 or R10.3 flow cell) can reach lengths of even 10–100 kb, but their accuracy (87–98%) is highly dependent on the base-calling algorithm and is inferior to that of high-fidelity (HiFi) sequencing reads, which represent the latest data type developed by PacBio with high performance (accuracy >99%, length >10 kb) [22,23]. The advent of LRS makes it possible to span complete repetitive DNA sequences, which enables more accurate measurement of long-distance repeat elements [24,25], resolves complex rearrangements [26,27], and simplifies the computational complexity of de novo genome assembly [28,29]. Taking advantage of LRS technology, researchers have successfully conducted large-scale SV studies in diverse populations worldwide [30–32], yielding a considerable number of novel SVs and complete gap-closed genome assemblies. In this review, we are concerned with technologies that produce continuous reads and do not involve optical mapping.

Current large-scale population studies generally use a hybrid approach combining long-read and short-read sequencing technologies to better utilize sequence-resolved SV collections [22,33] (Fig. 1). In brief, a relatively small number of deep sequenced LRS samples are used for genome assembly and variant detection, followed by a large number of SRS samples for genotyping. On the one hand, this strategy can take advantage of LRS to accurately detect as many variant loci as possible at an economical cost. On the other hand, many large-scale whole-genome sequencing (WGS) datasets collected from valuable clinical samples and other specific populations have been established in the SRS era [34–38]. Detection of novel SVs identified by LRS in these SRS samples allows better estimation of the allele frequency in local populations, facilitating other genotype-based downstream analyses [12,14,22,33]. As a result, genotyping of SVs in cumulative SRS samples remains a critical issue [33,39], although large-scale SV studies using LRS samples exclusively are emerging [40,41]. Similar to the detection strategy, traditional mapping-based genotypers extract SV signatures around the known breakpoints and determine the presence of alternative or reference alleles [39]. However, alignment against a single reference genome is biased towards the reference allele [22,42–45], so sequences containing large deviated alternative alleles are prone to mismatches or multiple alignments [46]. As reported in several population-scale SV studies, many of the SVs identified by LRS were classified as homozygous references in SRS samples [12,14,32], which severely impacts genotype-based population genetic studies or genome-wide association studies of complex diseases. Therefore, it is of paramount importance to explore new methods to eliminate reference bias to improve genotyping accuracy in population-scale SV studies.

The mapping-based tools for SV genotyping are biased towards the reference allele mainly because reads are aligned only to the reference genome [22]. Thus, various strategies have been tried to complement the comparison between sequencing reads and the alternative allele, two of which have been widely implemented in published genotyping tools. The first strategy still utilizes a linear reference genome, realigning short reads to a complete reference library that combines primary contigs and alternative allele sequences [12,47]. Considering that alt-aware mapping is still a non-trivial task, these studies endeavor to filter representative sequences from the original alignment [48]. Another approach is to build a graph-based pan-genome by integrating the reference and alternative alleles and searching for the path that best matches the genetic information of the target haplotype [39]. The graphical representation can describe all nested variants in the sequence more accurately than the linear structure [49]. However, appropriate tools are required to construct the variation graph and perform sequence-to-graph alignment and genotyping [42–44]. Finally, methods based on both strategies calculate support counts for the reference or alternative allele and then estimate the genotype through probabilistic [50] or machine learning models [12]. Some of these genotypers have already been applied in population-scale SV studies and demonstrate their potential in addressing reference bias [4,12].

Although several articles have reviewed SV calling algorithms based on LRS [2,3,19,20,22,23,51–53], little information is available on genotyping for SVs. In this review, we first examined population-scale SV studies using a hybrid sequencing strategy, pointing out the genotyping methods they used and the problems encountered. We are interested in the ability to genotype population-scale SVs, so we did not discuss studies analyzing a small number of target SVs [54,55]. Then we summarized current academic efforts to resolve the reference bias, including linear and graphic representations of the alternative allele. At the end of this review, we list pan-genomic tools available for genotyping of SVs, including graph construction and sequence-to-graph alignment, in the hope of helping to develop more efficient and accurate genotypers.
2. Population-scale structural variation studies using a hybrid sequencing strategy

LRS technology has dramatically improved the sensitivity and accuracy of SV detection, facilitating large-scale SV studies in various populations worldwide [22] (Table 1). In a pioneering study of applying LRS for analyzing SVs in 2017, Huddleston et al. generated deep SMRT sequencing data from two haploid human genomes [47]. Interestingly, although nearly 90% of SVs identified by LRS were missed in the 1000 Genomes Project reports [56], 61% of these sequence-resolved SVs can be successfully genotyped by short-read sequencing data [47]. This imbalance suggests that decoupling SV genotyping from discovery allows for genotyping the majority of previously missed SVs in the human genome [47]. Therefore, this team from the University of Washington School of Medicine has carried out a subsequent intensive series of work on population-scale SV detection and genotyping using the hybrid sequencing strategy [4,21,54,55,57]. In a landmark comprehensive study in 2019, Audano et al. sequenced eleven samples from diverse populations using SMRT sequencing [12]. Combined with four additional published resources [30,47,58], 99,604 non-redundant SVs were identified, 15% of which were shared in more than half of the samples [12], suggesting that the current reference genome either represents minor alleles or contains assembly errors [59]. In addition to characterizing the enrichment of SVs in tandem repeat sequences and closing gaps in the reference genome, they utilized Illumina WGS data collected from 440 samples to genotype sequence-resolved insertions and deletions, finding that 55% of SVs were successfully genotyped with a missing rate <5% [12]. During this period, many researchers followed the pipeline described in the abovementioned study but still used traditional genotypers based on the analysis of alignments only against the reference genome, such as SVTyper [61] and CNVnator [62], which are biased towards the reference allele. Although the SMRT-SV v2 genotyper developed by Audano et al. can represent both reference and alternative alleles [12], it is not scalable to larger populations due to the limitation of time-consuming alt-aware mapping [42].

With the advantages of LRS technology, high-quality and complete genome assemblies and an extensive collection of genetic variants have been accumulated rapidly in multiple populations, making it possible to construct a nonlinear pan-genomic model [22,42]. The pan-genome concept was initially proposed in microbiology to describe comprehensive genetic information, including
the core genome present in all strains and the dispensable genome present in specific strains [66]. Not surprisingly, researchers began introducing graph-based pan-genomic models to eliminate the reference bias and successfully scaled up the population size for genotyping from a few hundred to thousands of samples [4,32,63–65]. In one of the most extensive scale studies, Beyter et al. generated high-confidence SV sets discovered from 3,622 Icelanders by nanopore sequencing [32]. Combining 133,886 sequence-resolved SVs with previously discovered SNPs and indels [67], they constructed an augmented graph with a reference genome backbone using GraphTyper [68], and the resulting genotypes were utilized to explore SVs’ impact on diseases and other traits [32]. Using the same WGS dataset (median coverage 36.9x) [69], Eggertsson et al. reported that genotyping of 543,939 SVs by GraphTyper required 4.15 million CPU hours for 49,962 individuals or 483 CPU-hours per sample on average [68]. In contrast to adding genetic variations to the reference genome [67], Siret et al. demonstrated an alternative strategy to genotype a new sample by tracing haplotype paths through the sequence graph [64]. They developed Giraffe, a haplotype-aware pangenome mapper that prioritizes alignments under supervision from known haplotypes to avoid search-space explosion caused by combinations of biologically unlikely alleles [43,44,49,64,70]. On average, it took about 194 CPU hours to genotype a sample with a median coverage of 20× by the combination of toil-vg [71] and Giraffe [64]. A similar strategy was used by Ebert et al. to integrate information from k-mer tables and genetic variation across the input panel haplotypes, which bypasses the time-consuming sequence-to-graph alignment and only took about 30 CPU hours per sample on the tested coverage of 30× [4,72].

These studies indicate that graph-based genotyping can be effectively applied to large and diverse cohorts and promises to

Table 1
An overview of population-scale structural variation studies using a hybrid sequencing strategy.

| Study | Discovery sample size | Genotyping sample size | Genotyper | Genotyping rate | Recall rate |
|-------|------------------------|------------------------|-----------|----------------|-------------|
| Lu et al. (2022) [63] | 35 (20–40×) | 35 (HGSCV) | dangingtk v1.3 | – | – |
| Beyter et al. (2021) [32] | 3,622 (17×) | 10,000 (deCODE, 34×) | GraphTyper v2.6 | – | 36% |
| Ebert et al. (2021) [4] | 35 (20–40×) | 3,202 (1KG, 34×) | Paragaph v2.4 | 79% | 74% |
| Quan et al. (2021) [14] | 25 (10–20×) | 276 (40x) | Paragaph v2.4 | 69% | 54% |
| Sirén et al. (2021) [64] | 16 (>50×) | 2000 (MESA, 20×) | toil-vg | – | – |
| Yan et al. (2021) [65] | 15 (>50×) | 2504 (1KG, 30×) | Paragaph v2.2 | 86% | 73% |
| Ouzhuluobu et al. (2020) [31] | ZF1 (70×) | 77 (30x) | CNNator | – | – |
| Soto et al. (2020) [60] | 2 nonhumans | 8 (SGDP, 42×) | SVTyper v0.7 | 96% | 45% |
| Audano et al. (2019) [12] | 15 (>50×) | 174 (1KG, 25×) | SMRT-SV v2 | 55% | 97% |
| Chaisson et al. (2019) [21] | 9 (>50×) | 150 (Polaris, 18×) | SMRT-SV v2 | > 92% | > 96% |
| Kronenberg et al. (2018) [57] | CHM13 (>65×) | 24 (1KG, 39×) | SVTyper | – | – |
| | YRI19240 (>65×) | 16 (SGDP) | SVTyper v0.1 | – | – |
| Huddleston et al. (2017) [47] | CHM1 (62×) | 30 (1KG, 30×) | SMRT-SV v2 | 79% | 93% |

Genotyping rate, the proportion of SVs successfully genotyped, is usually determined by a missing rate threshold and the Hardy-Weinberg hypothesis. Recall rate, the proportion of the alternative allele presented in at least one haplotype. HGSVC, Human Genome Structural Variation Consortium. GTeX, the Genotype-Tissue Expression project. 1KG, 1000 Genomes Project. SGDP, Simons Genome Diversity Project. MESA, Multi-Ethnic Study of Atherosclerosis cohort. a Pacbio long-read sequencing. b Nanopore sequencing. – indicates that the information was not addressed in the paper.

Table 2
An overview of structural variation genotypers based on the linear reference genome.

| Tools | Input | Feature extraction | Genotyping model | Supported SV types |
|-------|-------|---------------------|-----------------|------------------|
|       |       |                     | Genotyping rate | INS  | DEL  | DUP  | INV  | TRA  |
| STIX [77] | SRS | RP, SR | – | × | | | | |
| muCNV [78] | SRS | RP, SR, RD | GMM | × | | | | |
| NPSV [79] | SRS | Realignment features | SVM/RF | – | | | | |
| Nebula [50] | SRS | Unique and affected k-mers | GMM | – | | | | |
| CNVJACC [76] | SRS | RP, SR, RD, and other sequence features | RP | – | | | | |
| SMRT-SV [12] | SRS | Realignment features | SVM | – | | | | |
| SV [74] | SRS | RP, SR, RD, HAR | SVM | × | | | | |
| Genome STRIP [80] | SRS | RP, SP, RD | GMM | – | | | | |
| SVTyper [61] | SRS | RP, SR | BM | × | | | | |
| Delllyv2 [73] | SRS | RP, SR, RD | BM | × | | | | |
| CNNator [62] | SRS | RP | GMM | – | | | | |
| SVJedi [48] | LRS | Realignment features | BM | – | | | | |
| Sniffles [81,82] | LRS | SR, alignment events | BST | – | | | | |
| svvsv2 [83] | LRS | Realignment MAPQ | BM | – | | | | |

RP, Read pair. SR, split-read. RD, read-depth. HAR, heterozygous allele ration. MAPQ, mapping quality. GMM, Gaussian mixture models. SVM, Support Vector Machine. RF, random forest. MLE, maximum likelihood estimation. GCM, single Gaussian models. BM, Bayesian model. BST, Binary Search Tree. SRS, short-read sequencing. LRS, long-read sequencing.
make an essential contribution to downstream analysis. Taking our own study as an example, we used Paragraph [39] to genotype a collection of 38,216 sequence-resolved SVs with a short-read sequencing dataset comprising 276 Chinese Tibetan and Han samples [14]. A considerable number of Tibetan-Han stratified SVs and candidate adaptive genes were inferred from unbiased genotypes, highlighting the important role of SVs in the evolutionary processes of adaptation to the Qinghai-Tibet Plateau [14]. In addition to the genotypers already applied to population-scale studies described above, researchers have made great efforts into developing algorithms to eliminate the reference bias. In the following sections, we discussed other approaches that employ linear or graph representations of the alternative allele.

3. Linear representation of the alternative allele

Traditional mapping-based genotypers, such as Delly [73], SVTyper [61], and SV [74], shared similar strategies with the pipeline for SV discovery [75]. These tools extract signatures of breakpoints from alignments only against the reference genome, generally including information on RP, SR, and RD, leading to a bias in favor of the reference allele [43] (Table 2). In addition to the above alignment features, some researchers extracted more sequence features near breakpoints to train a Support Vector Machine (SVM)- or Random Forest (RF)-based classifier [74,76], expecting to improve the genotyping performance. However, these tools are not only likely to yield biased genotypes [39] but also incapable of estimating insertions [75] and are therefore not suitable for comprehensive SV studies [22].

In order to minimize the reference allele bias, researchers tried to perform local realignment around known breakpoints against the alternative allele sequence [12,48,79,80,83]. Among genotypers designed for short-read sequencing data, Handsaker et al. proposed an enhanced version of Genome STRiP back in 2015, a population-based framework for genotyping SVs by aligning reads against a library containing alternative alleles [80,84]. Genome STRiP analyzes the distribution of read-depth by fitting Gaussian mixture models (GMM) corresponding to the homozygous reference allele, the homozygous or the heterozygous alternative allele [80]. The most likely genotype is finally determined by estimating copy number likelihoods [80]. Notably, Genome STRiP is limited to genotyping of deletions and duplications. In recent studies, another two representative tools have implemented this strategy, SMRT-SV genotyper [12] and NPSV [79]. SMRT-SV is an assembly-based approach with a linear representation of both the reference and alternative alleles for each SV [21]. This genotyping method aligns all short reads against the primary contig together with assembled alternative sequences per each variant [12], using an alt-aware manner by BWA-MEM [85]. An SVM-based classifier is trained on 15 features extracted from the alignment and then used to estimate all possible genotypes [12]. In a recent study in 2021, Linder-man et al. proposed NPSV, a simulation-driven approach to genotyping SVs by automatically creating sample- or variant-specific classifiers [79]. Instead of using actual data to train a genotype classifier as SMRT-SV, NPSV first generates synthetic short-read data using an SRS simulator [86] and then locally realigns these reads to the reference and alternate sequences [79]. This strategy helps generate representative training data for any putative SVs with all possible genotypes, avoiding the reference bias at the data source [79]. However, both SMART and NPSV are limited to SV genotyping of insertions and deletions, and they are not scalable to larger populations due to time-consuming alt-aware mapping.

4. Genotyping structural variation in pan-genome graphs

As discussed in the sector of population-scale studies, genotyping SVs using pan-genome graphs is still at a nascent but promising stage (Table 3). The main advantage of pan-genomic approaches is that they can more accurately represent the complex variability of the genome [22] and improve genotyping of nested SVs in complex genetic backgrounds [4,64]. However, there is still an urgent need for efficient tools to construct complex graphs and perform sequence-to-graph alignments [42–44]. In the following sections, we summarize the characteristics of graph-based genotypers. Although Cortex [87] is an early attempt at genotyping SVs using de Bruijn graphs, it was not discussed in our review because it was mainly applied to genotyping of small variants. Pan-genomic tools for graph construction and sequence-to-graph alignment are listed in Table 4, and these tools can be helpful in combination with genotypers, as reported by Sirén et al. [64].

4.1. Pan-genome graph construction

Most graph-based genotypers construct pan-genome graphs based on the directed acyclic graph (DAG). DAG is usually ordered along the reference genome and represents variants with a bubble composed of different branches between two vertices [43]. Therefore, each path in the DAG represents a possible haplotype. Paragraph [39] and GraphTyper2 [68] are two widely used genotypers constructing DAGs from a reference genome and sequence-resolved variants. Both tools extract short reads from original alignments at breakpoints and perform local mapping to the variation-aware graph [43], which helps reduce bias toward the reference genome and improves genotype quality [39,68]. Paragraph enables the representation of clustered SVs in the sequence graph and supports custom graph structures for genotyping more complicated events [39]. In addition, GraphTyper2 can also jointly genotype both small variants and SVs at a population scale by simultaneously encoding SNPs and indels into the pan-genome graph [68]. Nevertheless, these joint genotyping models

| Tools               | Graph construction | Graph Indexing strategy | Sequence-to-Graph alignment strategy | Genotyping algorithm |
|---------------------|--------------------|-------------------------|-------------------------------------|----------------------|
| Gramtools [49]      | NDAG               | vBWT                    | Variation-aware backward search     | Coverage model       |
| Minos [88]          | NDAG               | vBWT                    | Variation-aware backward search     | Coverage model       |
| toil-vg [71]        | VG                 | GCSA2, GBWT, XG, snarl   | SEMM seeds                          | Coverage model       |
| PanGenie [72]       | DAG                | k-mer hash table         | –                                   | HMM                  |
| GraphTyper2 [68]    | DAG                | k-mer hash table         | Matching k-mers as seeds            | Coverage model       |
| Paragraph [39]      | DAG                | Path families            | GSSW                                | Coverage model       |
| BayeTyper [89]      | VG                 | Variant cluster groups   | Heuristic search                    | Generative Model     |

DAG, directed acyclic graph. NDAG, nested DAG. VC, variation graph. BWT, Burrows–Wheeler transform. vBWT, variation BWT. SEMM, super-maximal exact match. HMM, Hidden Markov Model. GSSW, graph SMEM Smith–Waterman algorithm.
have limitations as they cannot represent nested variants like complex SVs [68].

To genotype complex SVs in variant-dense regions containing a large number of combinations of all possible alleles, Letcher et al. applied an algorithm called recursive collapse and cluster (RCC) implemented by Gramtools and generated a nested DAG consisting of a succession of locally hierarchical subgraphs [49]. Taking advantage of the nested data structure, Gramtools helps discover previously unknown recombination patterns between genetic variants from diverged backgrounds [49,88]. Gramtools also outputs a JSON variant call format (jVCF) to address the limitation of storing densely clustered variants in the standard VCF. Another idea about implementing paths [39]. Finally, Paragraph uses an expectation–maximization algorithm to estimate genotype likelihood-based allele frequencies based on the realignment coverage of each allele [39].

Other genotypers usually use a heuristic seed-and-extend paradigm pioneered by BLAST [92]. This paradigm first finds short seed hits, usually based on practical indexing tools, and then extends these hits to obtain complete alignments [95]. A pair of matching k-mers often acts as the seed hit for graph-based genotypers [68,72,89]. For example, GraphTyper2 constructs a k-mer hash table by indexing the full text of DAG and then searches for exact matches with k-mers from the read [68]. The final graph alignment is obtained by extending the longest seed through paths in the genome graph [68]. The genotype call also relies on a likelihood maximizing approach that aggregates both the original and the realignment coverage of each allele [39,68]. Considering that graph-based whole-genome alignment is time-consuming, both Paragraph and GraphTyper2 restrict the mapping operation to local variant clusters. However, this strategy is based on the realignment of reads to local graphs and requires information from original alignments, which is still disturbed by the reference bias.

In fact, a complete alignment is usually not necessary for genotyping of target SVs. Some researchers suggested that a traversal list of variants supported by each read is sufficient for genotyping [50,89]. BayesTyper, which is also a k-mer based method, adopts a kind of pseudo-alignment model [89]. This method compares the

This implementation extends the recurrence relation and the corresponding scoring matrices of dynamic programming across junctions in the local graph [39,94]. Reads aligned to a single graph location with the best mapping quality score were retained to genotype breakpoints [39]. A read is considered to support a node if its alignment overlapped the node by at least 10% of the read length, and a similar criterion is applied to the definition of supporting paths [39]. Finally, Paragraph uses an expectation–maximization algorithm to estimate genotype likelihood-based allele frequencies based on the realignment coverage of each allele [39].

### 4.2. Sequence-to-graph alignment and genotyping models

Sequence-to-graph alignment is a fundamental operation for graph-based genotyping [91]. In general, classical algorithms for sequence-to-sequence alignment, such as the Smith-Waterman (SW) algorithm [92], cannot be directly applied to genome graphs. Nonetheless, Paragraph applies an extended generalization of Far- rarr's striped SW algorithm [93] to local graph alignment [39,94].

| Category         | Tools             | Graph | Output format | Description                                                                 |
|------------------|-------------------|-------|---------------|----------------------------------------------------------------------------|
| Graph Construction | seqwish           | VG    | GFA           | A VG building from a set of sequences and alignments between them [96]       |
|                   | Cuttlefish        | DBG   | GFA           | A colored compacted DBG building from a collection of genome references [97,98] |
|                   | ODGI              | VG    | ODGI          | A suite of tools that implements scalable algorithms [99]                  |
|                   | Pandora            | DAG   | FASTA         | A pan-genome graph structure and algorithms for identifying variants [100] |
|                   | Simpltrigs        | DBG   | FASTA         | A compact representation of DBG [101]                                      |
|                   | Bifrost           | DBG   | GFA           | A parallel algorithm enabling the direct construction of the compacted DBG [102] |
|                   | libbdsg           | VG    | GFA           | Tools allow for construction and manipulation of genome graphs with dense variation [103] |
|                   | minigraph         | VG    | ODGI          | A graph-based data model to represent multiple genomes [104]               |
|                   | SevenBridges      | DAG   | –             | A computationally graph genome implementation [105]                        |
|                   | vg                | VG    | VG            | A toolkit of computational methods for creating and manipulating VG [90]    |
|                   | Wheeler graphs    | DBG   | DOT           | A framework for BWT-based data structures [106]                           |
| Graph alignment  | GraphChainer      | VG    | GAM           | A algorithm to co-linearly chain a set of seeds in an acyclic VG [107]     |
|                   | BlastFrost        | DBG   | GFA           | Query Bifrost data structure for sequences of interest [108]               |
|                   | A*                | –     | ALN SAM       | A seed heuristic enabling fast and optimal sequence-to-graph alignment [109,110] |
|                   | Giraffe           | VG    | GAF GAM       | A pan-genome short-read mapper that can map to a collection of haplotypes [64] |
|                   | GraphAligner      | VG    | GAF GAM       | A tool for aligning long reads to genomes [91]                             |
|                   | SPAAligner        | DBG   | GFA           | A tool for aligning long diverged nucleotide and amino acid sequences to assembly graphs [111] |
|                   | Vargas            | DAG   | SAM FASTA     | A heuristic-free algorithm to find the highest-scoring alignment [112]     |
|                   | PaSGAL            | DAG   | TSV           | A parallel algorithm for computing sequence to graph alignments [113]       |
|                   | HISAT2            | DBG   | SAM           | A tool can align both DNA and RNA sequences using a graph Ferragina Manzini index [114] |
|                   | V-ALIGN           | DAG   | TXT           | A tool based on dynamic programming that allows gapped alignment directly on the input graph [115] |
unbiased distribution of k-mers from sequencing reads to the k-mer profile along paths representing the most likely haplotypes [89]. The posterior distribution over all possible genotypes is estimated according to the counts of k-mers in the reads based on a generative model [89]. However, approaches based solely on the k-mer counts cannot reliably genotype variants in repetitive regions because unique k-mers may not exist for the variants [4,72]. In a recent study, Ebler et al. proposed PanGenie, which integrates information from k-mer tables and genetic variation across the input panel haplotypes [72]. They utilized information from known haplotype sequences to infer genotypes based on neighboring variants, therefore avoiding the inability to genotype in the absence of unique k-mers [72]. Since PanGenie and BayesTyper bypass the time-consuming alignment step, they are much faster than the remaining mapping-based methods.

5. Summary and outlook

The rapid development of LRS in recent years has revitalized SV studies. Taking advantage of LRS technology, researchers have successfully conducted large-scale SV studies in diverse populations worldwide [30–32], yielding a considerable number of novel SVs and complete gap-closed genome assemblies. However, genotyping SVs in a large-scale short-read sequencing cohort remains challenging. Traditional mapping-based genotypers are biased towards the reference allele [22]. Therefore, researchers have made great efforts to eliminate the reference bias by representing both the reference and the alternative allele using a linear or graph genome. Notably, most recent population-scale studies of SVs have used pan-genomic models to eliminate reference bias and successfully scaled up the population size for genotyping from a few hundred to thousands of samples [4,32,63–65], facilitating other genotyping-based downstream analyses. Recently, the Telomer-to-Telomere (T2T) Consortium and the Human Pangenome Reference Consortium have successfully announced their exciting progress in constructing complete and error-free T2T assemblies of all chromosomes as well as full-spectrum genomic variant collections [116–118], which will further promote the application of pan-genomic approaches in population genetic studies.

Genotyping SVs using pan-genome graphs is still at a nascent stage. There is still an urgent need in this field for efficient tools to construct complex graphs and perform sequence-to-graph alignments. For example, complex SVs often occur in repetitive regions and are nested with other small variants. Despite the potential for reliable genotyping of complex SVs by bidirectional variation graphs and nested DAGs, complex SVs have not been comprehensively analyzed in population-scale studies. Little is known about their contribution to genetic evolution or their interaction with other variants. The same problem is faced by mosaic and low-frequency SVs, which have been reported to be risk factors for neurological diseases [82,119]. Besides, it remains unclear whether pan-genomic approaches will become mainstream in clinical diagnostics. Some researchers argue that graph-based genotyping relies on single-base resolution breakpoints, making it more suitable for studying common variants rather than somatic or pathogenic variants [120]. In addition, graph-based genotyping approaches are not entirely mature, with competing implementations and data formats [22]. There is an urgent need for a benchmark to evaluate the genotyping performance of graph-based genotyper with uniform criteria.

CRediT authorship contribution statement

Cheng Quan: Writing - original draft, Visualization. Hao Lu: Writing - original draft, Writing - review & editing. Yiming Lu: Writing - original draft, Writing - review & editing. Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the General Program (31771397 and 81672369) of the National Natural Science Foundation of China (www.nsfc.gov.cn), the National Key R&D Program of China (No. 2017YFA0504301), and the Chinese Key Project for Infectious Diseases (No. 2018ZX10732202 and 2017ZX10203205).

References

[1] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011;12:363. https://doi.org/10.1038/nrg2858.
[2] Sedlazeck FJ, Lee H, Darby CA, Schartz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018;19:329–46. https://doi.org/10.1038/s41576-018-0003-4.
[3] Mahmoud M, Gobet N, Cruz-Dávalos D, Mounier N, Dessimon G, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol 2019;20:246. https://doi.org/10.1186/s13059-018-1828-7.
[4] Ebert P, Audano PA, Zhu Q, Rodríguez-Martín B, Porubsky D, Bondér M, et al. Haplootype-resolved diverse human genomes and integrated analysis of structural variation. Science 2021;372. https://doi.org/10.1126/science.ahj7117.
[5] Collins RL, Brand H, Karaczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature 2018;558:444–51. https://doi.org/10.1038/s41586-018-2297-8.
[6] Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature 2020;578:112–21. https://doi.org/10.1038/s41586-020-1913-9.
[7] Fudenberg G, Pollard KS. Chromatin features constrain structural variation across evolutionary timescales. Proc Natl Acad Sci 2019;116:201808631. https://doi.org/10.1073/pnas.1806311116.
[8] Chang C, Scott AJ, Davis JR, Tsang ER, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet 2017;49:692–9. https://doi.org/10.1038/ng.3838.
[9] Audano PA, Sulovari A, Graves-Lindsay T, Gedda N, Méggy K, Pernecky C, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med 2018;10:95. https://doi.org/10.1186/s13059-018-0606-6.
[10] Fudenberg G, Pollard KS. Chromatin features constrain structural variation across evolutionary timescales. Proc Natl Acad Sci 2019;116:201808631. https://doi.org/10.1073/pnas.1806311116.
[11] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011;12:363. https://doi.org/10.1038/nrg2858.
[12] Almarri MA, Bergström A, Prado-Martínez J, Yang F, Fu B, Dunham AS, et al. Population structure, stratification, and introgression of human structural variation. Cell 2020;184:654–64.e16. https://doi.org/10.1016/j.cell.2020.02.027.
[13] Alkan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol 2021;22:159. https://doi.org/10.1186/s13059-021-02382-7.
[14] Chen L, Abel HJ, Das I, Larsson DE, Ganel I, Kanchi KL, et al. Association of structural variation with cardiometabolic traits in Finns. Am J Hum Genetics 2021;108:583–96. https://doi.org/10.1016/j.ajhg.2021.03.010.
[15] Cortés-Criado I, Lee H, Ji X, Yuan D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet 2020;52:331–41. https://doi.org/10.1038/s41588-019-0576-7.
[16] Ganevskov S, Goodwin S, Sherman RM, Sedlazeck FJ, Arun G, Bhatia S, et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. Genome Res 2020;30:1258–73. https://doi.org/10.1101/gr.260497.119.
[17] Hadi K, Yao X, Behr MI, Darshei P, Xanthopoulos C, Tian H, et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell 2020;183:197–210.e32. https://doi.org/10.1016/j.cell.2020.08.006.
[19] Yoshitaka S, Yutaka Z, Masahide S, Ayako S. Application of long-read sequencing to the detection of structural variants in human cancer genomes. Comput Struct Biotechnol J 2021;19:4207–16. https://doi.org/10.1016/j.csbj.2021.07.030.

[20] Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet 2019;1:19. https://doi.org/10.1038/s41576-018-0190-3.

[21] Chaisson MJ, Zhao X, Malhotra A, Porubsky D, Basile A, T et al. Multi-platform discovery of haploïte-resolved structural variation in human genomes. Nat Commun 2020;11:10.1038/s41467-019-11443-3.

[22] Coster W, Weissenbacher MH, Siedzalek FJ. Towards population-scale long-read sequencing. Nat Rev Genet 2021;22:21–16. https://doi.org/10.1038/s41576-021-0363-7.

[23] Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet 2020:1–18. https://doi.org/10.1038/s41467-020-02345-9.

[24] Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al. One reference genome is not enough. Genome Biol 2021;20:291. https://doi.org/10.1186/s13059-021-02474-x.

[25] Lu T-Y, Consortium T, Munson KM, Lewis AP, Zhu Q, Tallon LJ, et al. Profiling and de novo assembly of a human genome with ultra-long reads. BioRxiv 2021:2021.02.06.430068. https://doi.org/10.1101/gr.114876.110.
