BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 22 2012, pages 2875–2882
doi:10.1093/bioinformatics/bts566

Genome analysis

Advance Access publication October 11, 2012

CLEVER: clique-enumerating variant finder

Tobias Marschall[1,*,†], Ivan G. Costa[2,3,†], Stefan Canzar[4,‡], Markus Bauer[5], Gunnar W. Klau[1], Alexander Schliep[6] and Alexander Schönhuth[1,*]

[1]Centrum Wiskunde & Informatica, Life Sciences Group, Amsterdam, The Netherlands, [2]Interdisciplinary Centre for Clinical Research (IZKF), RWTH University Medical School, Aachen, Germany, [3]Center of Informatics, Federal University of Pernambuco, Recife, Brazil, [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA, [5]Illumina, Cambridge, UK and [6]Department of Computer Science and BioMaPS Institute for Quantitative Biology, Rutgers, The State University of New Jersey

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Next-generation sequencing techniques have facilitated a large-scale analysis of human genetic variation. Despite the advances in sequencing speed, the computational discovery of structural variants is not yet standard. It is likely that many variants have remained undiscovered in most sequenced individuals.

Results: Here, we present a novel internal segment size based approach, which organizes *all*, including concordant, reads into a *read alignment graph,* where max-cliques represent maximal contradiction-free groups of alignments. A novel algorithm then enumerates all max-cliques and statistically evaluates them for their potential to reflect insertions or deletions. For the first time in the literature, we compare a large range of state-of-the-art approaches using simulated Illumina reads from a fully annotated genome and present relevant performance statistics. We achieve superior performance, in particular, for deletions or insertions (indels) of length 20–100 nt. This has been previously identified as a remaining major challenge in structural variation discovery, in particular, for insert size based approaches. In this size range, we even outperform split-read aligners. We achieve competitive results also on biological data, where our method is the only one to make a substantial amount of correct predictions, which, additionally, are disjoint from those by split-read aligners.

Availability: CLEVER is open source (GPL) and available from http://clever-sv.googlecode.com.

Contact: as@cwi.nl or tm@cwi.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 2, 2012; revised on September 10, 2012; accepted on September 17, 2012

1 INTRODUCTION

The International HapMap Consortium (2005) and The 1000 Genomes Project Consortium (2010) have, through globally concerted efforts, provided the first systematic view on the gamut and prevalence of human genetic variation, including larger genomic rearrangements. A staggering 8% of the general human population have copy number variants (CNVs) affecting regions larger than 500 kb (Itsara *et al*., 2009). The technology enabling this advance was next-generation sequencing and the reduction in costs and increases of sequencing speeds it brought along (Bentley *et al*., 2008; Eid *et al*., 2009). The analysis of structural variation, however, has not kept up with the advances in sequencing insofar as genotyping of human structural variation has not yet become a routine procedure (Alkan *et al*., 2011). Indeed, it is likely that existing datasets contain structural variations indiscoverable by current methods. These limitations are likewise an obstacle to personalized genomics.

Here, we target *deletions or insertions (indels)* between 20 and 50 000 bp. In particular, the discovery of indels smaller than 500 bp is still challenging (Alkan *et al*., 2011; Mills *et al*., 2011), even in non-repetitive areas of the genome. That the majority of structural variants resides in repetitive areas complicates the problem further due to the resulting read-mapping ambiguities.

Categorization of our and prior work. A *(paired-end) read* is a fragment of DNA in which both ends have been sequenced. We refer to the sequenced ends of the read as *(read) ends* and to the unsequenced part of the fragment between the two ends as *internal segment* or *insert*. An *alignment A* of a paired-end read is a pair of alignments of both ends. We say that a read has been *multiply mapped* if it aligns at several locations in the reference genome and *uniquely mapped* in case of only one alignment. Existing approaches for structural variant discovery can be classified into three broad classes: first, those based on the read alignment coverage, that is, the number of read ends mapping to a location (Abyzov *et al*., 2011; Alkan *et al*., 2009; Campbell et *al*., 2008; Chiang *et al*., 2009; Sudmant *et al*., 2010; Yoon *et al*., 2009), second, those analyzing the paired-end read internal segment size (Chen *et al*., 2009; Hormozdiari *et al*., 2009; Korbel *et al*., 2009; Lee *et al*., 2009; Quinlan *et al*., 2010; Sindi *et al*., 2009) and third, split-read alignments (Mills *et al*., 2006; Ye *et al*., 2009). Refer to Medvedev *et al*. (2009) as well as to Alkan *et al*. (2011) for reviews. A major difference is that the first two classes align short reads by standard read mappers, such as BWA (Li and Durbin, 2009), Mr and MrsFast (Alkan *et al*., 2009; Hach *et al*., 2010) and Bowtie (Langmead *et al*., 2009). However, split-read aligners compute custom alignments that span breakpoints of putative insertions and deletions. They usually have advantages over insert size based approaches on smaller indels while performing worse in predicting larger indels.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Part of the work was done while the author was at Centrum Wiskunde & Informatica.

It is common to many library protocols that internal segment size follows a normal distribution with machine- and protocol-specific mean μ and standard deviation σ. On a side remark, we would like to point out that our approach does not depend on this assumption and that we also accommodate arbitrary internal segment size distributions (which may result from preparing libraries without a size selection step, as one example) to the user. One commonly defines *concordant and discordant alignments*: an alignment with interval length $I(A)$ (see Fig. 1) is concordant iff $|I(A) - \mu| \leq K\sigma$ and discordant otherwise. The constant K can vary among the different approaches. A *concordant read* is defined to concordantly align with the reference genome, that is, it should give rise to at least one concordant alignment.

With only one exception (Lee *et al*., 2009, MoDIL), all prior approaches discard concordant reads. In this article, we present clique-enumerating variant finder (CLEVER), a novel insert size based approach that takes *all*, including concordant, reads into consideration. Although a single discordant read is significantly likely to testify the existence of a structural variant, a single concordant read only conveys a weak variant signals if any. Ensembles of *consistent concordant alignments,* however, can provide significant evidence of usually smaller variants. The major motivation of this study is to systematically take advantage of such groups of alignments to not miss any significant variant signal among concordant reads.

We employ a statistical framework, which addresses deviations in insert size, alignment quality, multiply mapped reads and coverage fluctuations in a principled manner. As a result, our approach outperforms all prior insert size approaches on both simulated and biological data and also compares favorably with two state-of-the-art split-read aligners. Beyond its favorable results, our tool predicts a substantial amount of correct indels as the only tool (e.g. more than 20% of true deletions of 20–49 bp in the simulated data). Overall, CLEVER's correct calls beneficially complement those of the split-read aligner considered (Ye *et al*., 2009, PINDEL).

Moreover, we need ~8 h on a single CPU for a 30× coverage whole-genome dataset with ~1 billion reads, which compares favorably with the estimated 7000 CPU hours needed by MoDIL, the only method that also takes all reads into consideration.

1.1 Approach and related work

1.1.1 Graph-based framework Our approach is based on organizing all read alignments into a read alignment graph, whose nodes are the alignments and edges reflect that the reads behind two overlapping alignments are, in rigorous statistical terms, likely to stem from the same allele. Accordingly, maximal cliques (max-cliques) reflect maximal consistent groups of alignments that are likely to stem from the same location in a donor allele. Because we do not discard alignments, the number of nodes in our read alignment graph is large. We solve instances with more than 10^9 nodes. We determine all max-cliques in this graph by means of a specifically engineered, fast algorithmic procedure.

The idea to group alignments into location-specific, consistent ensembles, such as max-cliques here, is not new. In fact, it has been employed in the vast majority of previous insert size based approaches. We briefly discuss related concepts of the three most closely related approaches by Hormozdiari *et al.* (2009, VariationHunter [VH]), Sindi *et al*. (2009, GASV) and Quinlan *et al*. (2010, HYDRA). Although not framing it in rigorous statistical terms, HYDRA is precisely based on the same concept of max-clique as our approach. After constructing the read alignment graph from discordant reads alone, they employ a heuristic algorithm to find max-cliques. Because no theoretical guarantee is given, it remains unclear whether HYDRA enumerates them all. The definition of a 'valid cluster' in VH (Hormozdiari *et al*., 2009) relaxes our definition of a clique in a subtle, but decisive aspect. As a consequence, each of our max-cliques forms a valid cluster, but the opposite is not necessarily true. The reduction in assumptions, however, allows VH to compute valid clusters as max-cliques in interval graphs in a nested fashion, which yields a polynomial run-time algorithm. Sindi *et al*. (2009, GASV) use a geometrically motivated definition that allows application of an efficient plane-sweep style algorithm. A closer look reveals that each geometric arrangement of alignments inferred by GASV constitutes a max-clique in our sense, but not necessarily vice versa, even if a max-clique is formed by only discordant read alignments. We recall that GASV, HYDRA and VH do not consider concordant read data and hence consider read alignment graphs of much reduced sizes.

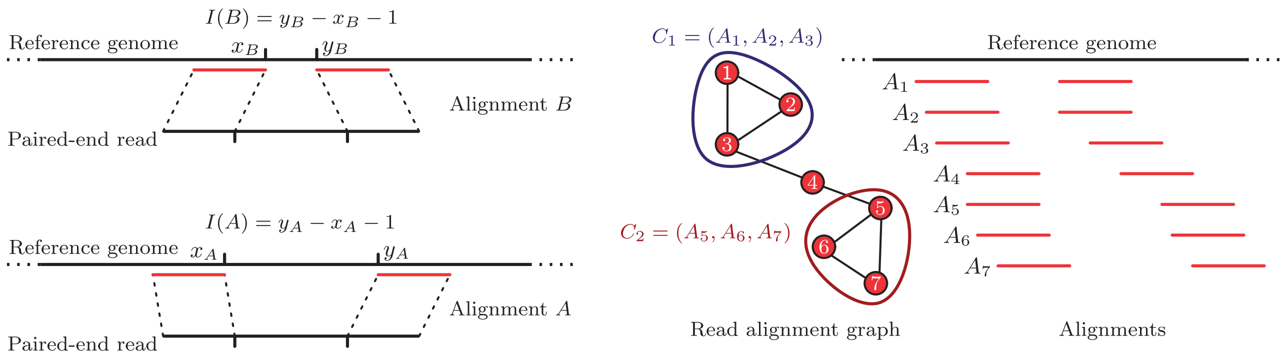

Fig. 1. Left panel: two read alignments. Assuming $I(A) > \mu > I(B)$, where μ is the mean of the true insert size distribution, alignment A is likely to indicate a deletion while alignment B may indicate an insertion. Right panel: Read alignment graph for seven closely located read alignments. Note that $1/3(I(A_5) + I(A_6) + I(A_7)) > 1/3(I(A_1) + I(A_2) + I(A_3))$. Assuming that all alignments have equal weight, C_2 is more likely to indicate a deletion than C_1 through a hypothesis test as in Equations (3) and (2). Note that we have not marked cliques (A_3, A_4) and (A_4, A_5). See Figure 2 for definition of edges

Finding max-cliques is $\mathcal{NP}$-hard in general graphs. On the basis of the idea that the read alignment graph we consider still largely resembles an interval graph, we provide a specifically engineered routine that computes and tests all max-cliques in a reasonable time—about 1 h on a current eight-core machine for a whole human genome sequenced to 30× coverage—despite that we do not discard any reads.

1.1.2 Significance evaluation

Commonly concordant and discordant reads: Testing whether $|I(A)-\mu| \le K\cdot\sigma$, to determine whether a single alignment is concordant, is equivalent to performing a Z-test at significance level $p_K := 1-\Phi(K)$, where Φ is the standard normal distribution function. However, when determining whether m consistent alignments (such as a clique of size m) with mean interval length $\bar{I}$ are *commonly concordant*, a Z-test for a sample of size m is required, which translates to

$$1-\Phi(\sqrt{m}\cdot\frac{|\bar{I}-\mu|}{\sigma}) \ge p_K \Leftrightarrow \sqrt{m}\cdot|\bar{I}-\mu| \le K\cdot\sigma. \qquad (1)$$

Due to the factor $\sqrt{m}$, already smaller deviations $|\bar{I}-\mu|$ turn out to render the alignments *commonly discordant*. In our approach, we rigorously expand on this idea. Roughly speaking, each max-clique undergoes a Inequality-(1)-like hypothesis test.

Multiply mapped reads: Although we approach the idea of not 'overusing' multiply mapped reads in an essentially different fashion, our routine serves analogous purposes as the set-cover routines of VH and HYDRA. The difference is that we statistically control read-mapping ambiguity but do not aim at resolving it.

Following Li *et al.* (2008), we compute each alignment's probability of being correctly placed. In case of a max-clique consisting of alignments $A_1,\ldots,A_n$ (all from different reads) with probabilities $p_1,\ldots,p_n$, let $A_J, J\subset\{1,\ldots,n\}$ be the event that precisely the alignments $A_j, j\in J$ are correct. We compute $\mathbf{P}(A_J)=\prod_{j\in J}p_j\prod_{j\notin J}(1-p_j)$. Let H_0 be the null hypothesis that the allele in question that—we recall that max-cliques just represent groups of alignments likely to be from the same allele—coincides with the reference genome. In correspondence to Inequality (1), we compute

$$\mathbf{P}_{H_o}(A_J) := 1-\Phi(\sqrt{|J|}\frac{|\bar{I}_J-\mu|}{\sigma}) \qquad (2)$$

with $\bar{I}_J=\frac{1}{\sum_{j\in J}p_j}\sum_{j\in J}p_jI(A_j)$, which is the probability of observing $A_j, j\in J$ when assuming the null hypothesis, given A_J. We further compute

$$\mathbf{P}_{H_0}(A_1,\ldots,A_n)=\sum_{J\subset\{1,\ldots,n\}}\mathbf{P}(A_J)\mathbf{P}_{H_0}(A_J) \qquad (3)$$

as the probability that max-clique $A_1,\ldots,A_m$ does *not* support an indel variant. We further correct $\mathbf{P}_{H_0}(A_1,\ldots,A_n)$ with a *local Bonferroni factor* to adjust for coverage-mediated fluctuations in the number of implicitly performed tests. If the corrected $\mathbf{P}_{H_0}(A_1,\ldots,A_n)$ is significantly small, it is likely that (at least) one allele in the donor is affected by an indel at that location. See Section 2 for details. In a last step, we apply the Benjamini–Hochberg procedure to correct for multiple hypothesis testing overall. Note that, among the prior approaches, only MoDIL (Lee *et al.*, 2009) addresses to correct for multiple hypothesis testing (also using Benjamini-Hochberg), although many others either explicitly (e.g. Chen *et al.,* 2009) or implicitly (e.g. Hormozdiari *et al.,* 2009; Korbel *et al.,* 2009; Quinlan *et al.,* 2010) perform multiple hypothesis tests.

Among the statistically motivated approaches, Lee *et al.* (2009), after clustering, use Kolmogorov–Smirnov tests in combination with bimodality assumptions, whereas Chen *et al.* (2009) measure both deviations from Poisson-distribution based assumptions (BreakdancerMax) and use Kolmogorov–Smirnov (BreakdancerMin) tests to discover copy number changes.

2 METHODS

2.1 Notations, definitions and background

2.1.1 Reads and read alignments Let $\mathcal{R}$ be a set of paired-end reads, stemming from a *donor (genome)* that has been aligned against a *reference (genome)*. We write A for a paired-end alignment, that is a pair of alignments of the two ends of a read (Fig. 1) and $\mathcal{A}(R)$ for the set of correctly oriented alignments that belong to read R. We neglect incorrectly oriented alignments and write $\mathcal{A}=\cup_R\mathcal{A}(R)$ for the set of all alignments we consider. We assume that $|\mathcal{A}(R)|\ge 1$; that is, each read we consider give rise to at least one well-oriented alignment. We do not discard any reads.

We write x_A for the rightmost position of the left end and y_A for the leftmost position of the right end. We write $[x_A+1, y_A-1]$ and call this the *interval* of alignment A (in slight abuse of notation: intervals here only contains integers) and $I(A) := y_A - x_A - 1$ for the *(alignment) interval length*. When referring to alignment intervals, we sometimes call x_A, y_A the left and right *endpoint*. See Figure 1 for illustrations.

2.1.2 Internal segment size statistics We write $I(R)$ for the internal segment (or insert) size of paired-end read R, i.e. the distance between the 3′ ends—the inner ends of the sequenced reads. Note that the distance between the 5′ outer ends is an equally common definition for *insert size* in the literature. In the datasets treated here, $I(R)$ can be assumed normally distributed with a given mean μ and standard deviation σ (Hormozdiari *et al.*, 2009; Lee *et al.*, 2009; Li and Durbin, 2009; Li *et al.*, 2008), i.e. $I(R)\sim\mathcal{N}_{(\mu,\sigma)}$. Estimation of mean μ and standard deviation σ from the alignments A of reads R poses the challenge that statistics on alignment insert size $I(A)$ (further denoted as $\mathbf{P}_{\text{Emp}}$) do not immediately reflect statistics on $I(R)$ because alignment insert size $I(A)$ statistics already reflect the structural variants in the dataset. As a result, statistics on $I(A)$ are fat-tailed and multimodal, even if library protocols determine statistics on $I(R)$ as normal. Here, we rely on robust estimation routines, as implemented by BWA (Li and Durbin, 2009). Note that, in general, we allow to deal with arbitrary internal segment size statistics.

2.1.3 Alignment scores and probabilities As described by Li *et al.* (2008), we determine $\log_{10}\mathbf{P}_{\text{Ph}}(A) := -\sum_j Q_j/10$, where j runs over all mismatches in both read ends and Q_j is the Phred score for position j, i.e. $10^{-(Q_j/10)}$ is the probability that the nucleotide at position j reflects a sequencing error. Hence, $\mathbf{P}_{\text{Ph}}(A)$ is the probability that the substitutions in alignment A are due to sequencing errors. The greater $\mathbf{P}_{\text{Ph}}(A)$ the more likely that A is correct, so $\mathbf{P}_{\text{Ph}}(A)$ serves as a statistical quality assessment of A. Note that to neglect single-nucleotide polymorphism (SNP) rates and indels reflects common practice (Li and Durbin, 2009; Li *et al.*, 2008), which is justified as in Illumina reads substitution error rates are higher than SNP rates, indel sequencing error rates and deletion/insertion polymorphism (DIP) rates by orders of magnitude (Bravo and Irizarry, 2010; Albers *et al.*, 2011).

Following Li *et al.* (2008) and Li and Durbin (2009), we integrate the empirical interval length distribution $\mathbf{P}_{\mathrm{Emp}}(I(A))$ into an overall score $S_0(A) := \mathbf{P}_{\mathrm{Ph}}(A) \cdot \mathbf{P}_{\mathrm{Emp}}(I(A))$ and obtain as the probability that A is the correct alignment for its read, by application of Bayes' formula

$$\mathbf{P}_0(A) = \frac{S_0(A)}{\sum_{\tilde{A} \in \mathcal{A}(R)} S_0(\tilde{A})}. \tag{4}$$

2.1.4 The read alignment graph We arrange all scored read alignments $\mathcal{A}$ in the form of an undirected, weighted graph $G = (\mathcal{A}, E, w)$. Because we identify nodes with read alignments from $\mathcal{A}$, we use these terms interchangeably. We draw an edge between alignments $A, B \in \mathcal{A}$ if we cannot reject the hypothesis that, in case they are both correct, their reads can stem from the same allele. See the subsequent paragraph for details. The weight function $w : \mathcal{A} \to [0, 1]$ is defined by $w(A) := \mathbf{P}_0(A)$. We further label nodes by $r : \mathcal{A} \to \{1, \ldots, N\}$, where $r(A) = n$ iff $A \in \mathcal{A}(R_n)$ that is alignment A is due to read R_n.

As usual, we write $\delta(A) := |\{B \in \mathcal{A} | (A, B) \in E\}|$ for the *degree* of node A. A *clique* $\mathcal{C} \subset \mathcal{A}$ is defined as a subset of mutually connected nodes, i.e., $(A, B) \in E$ for all $A, B \in \mathcal{C}$. A *max-clique* $\mathcal{C}$ is a clique, such that for every node $A \in \mathcal{A} \setminus \mathcal{C}$ there is $B \in \mathcal{C} : (A, B) \notin E$. Note that by our definition of edges, a clique is a group of alignments that can be jointly assumed to be associated with the same allele, or, in other words, to jointly support the same local phenomenon in the donor genome. Max-cliques are obviously particularly interesting: although all alignments in the clique are likely to support the same local phenomenon, joining any other *overlapping* alignment may lead to conflicts.

2.1.5 Edge computation See Figure 2 for illustrations of the following. Let A, B be two alignments. We define:

- $\Delta(A, B) := |I(A) - I(B)|$ is the absolute difference of interval length.
- $O(A, B) := \min(y_A, y_B) - \max(x_A, x_B) - 1$, where in case of $O(A, B) \geq 0$ we refer to all positions between $\max(x_A, x_B)$ and $\min(y_A, y_B)$ as their *common interval.*
- $\bar{I}(A, B) := (I(A) + I(B))/2$ is the *mean interval lengths.*
- $U(A, B) := \bar{I}(A, B) - O(A, B)$ is the difference of mean interval length and overlap. To motivate this quantity, note that, in case A and B overlap [hence, the length of common interval $O(A, B) > 0$] and are from the same allele, a deletion at that location can only happen to take place in their common interval. If $U(A, B)$ is large, then $\bar{I}(A, B)$ significantly deviates from μ and the common interval is not large enough to explain this by a large-enough deletion. Hence, it is unlikely that A,B are from the same allele.

Let X be $\mathcal{N}_{(0,1)}$-distributed and, as above, μ, σ be the mean and variance of the insert size distribution. We draw an edge between alignments A, B in the read alignment graph iff the reads of A and B are different, $O(A, B) \geq 0$ and

$$\mathbf{P}(|X| \geq \frac{1}{\sqrt{2}} \frac{\Delta(A, B)}{\sigma}) \leq 0.05 \quad \text{and} \tag{5}$$

$$\mathbf{P}(X \geq \sqrt{2} \frac{(U(A, B) - \mu)}{\sigma}) \leq 0.05 \tag{6}$$

Inequality (5) is a two-sided two sample Z-test to measure *statistically compatible insert size*. Inequality (6) reflects a one-sided one-sample Z-test for *statistically consistent overlap* (Wasserman, 2004). If two alignments A, B with $O(A, B) \geq 0$ pass these tests, we have no reason to reject the hypothesis that the alignments are from the same allele, so we draw an edge.

2.2 CLEVER: algorithmic workflow

(1) Enumerating max-cliques: We compute all *max-cliques* in the read alignment graph.

(2) We assign two P-values, $p_D(\mathcal{C}), p_I(\mathcal{C})$ to each max-clique $\mathcal{C}$, which are the probabilities that the alignments participating in $\mathcal{C}$ do not commonly support a deletion or insertion. So the lower $p_D(\mathcal{C})$ or $p_I(\mathcal{C})$, the more likely it is that $\mathcal{C}$ supports a deletion or insertion, respectively.

(3) For the thus-computed P-value, we control the false discovery rate at 10% by applying the standard Benjamini–Hochberg procedure separately for insertions and deletions. All cliques remaining after this step are deemed *significant* and processed further.

(4) Determining parameters: We parameterize deletions D by their left breakpoint D_B and their length D_L, which denotes that reference nucleotides of positions $D_B, \ldots, D_B + D_L - 1$ are missing in the donor. We parameterize insertions I by their breakpoint I_B and their length I_L, such that before position I_B in the reference there has been a sequence of length I_L inserted in the donor. Depending on whether $\mathcal{C}$ represents a deletion or insertion, we determine, defining $w(\mathcal{C}) := \sum_{A \in \mathcal{C}} w(A)$,

$$\frac{1}{w(\mathcal{C})} \sum_{A \in \mathcal{C}} w(A)(I(A) - \mu) \quad \text{respectively} \quad \frac{1}{w(\mathcal{C})} \sum_{A \in \mathcal{C}} w(A)(\mu - I(A)) \tag{7}$$

as the length D_L of the deletion, respectively, I_L of the insertion. We determine breakpoints D_B or I_B such that the predicted deletion or insertion sits right in the middle of the intersection of all internal segments of alignments in $\mathcal{C}$.

2.2.1 Enumerating max-cliques We identify nodes of the read alignment graph with the intervals of the corresponding alignments. We first sort the $2m$ endpoints of these intervals, $m := |\mathcal{A}|$, in ascending order of their positions. We then scan this list from left to right. We maintain a set of *active* cliques that could potentially be extended by a subsequent interval, which initially is empty. If the current element ℓ of the list is a left endpoint, we extend the set of active cliques according to the following rules. For the sake of simplicity, let us assume that a unique interval starts at ℓ, corresponding to a vertex A in the read alignment graph G. Let $N(A)$ be the open neighborhood of A. If $\mathcal{C} \cap N(A) = \emptyset$ for all active cliques $\mathcal{C}$, add a singleton clique $\{A\}$ to the set of active cliques. Otherwise, for each active clique $\mathcal{C}$,

(i) if $\mathcal{C} \cap N(A) = \mathcal{C}$, then $\mathcal{C} := \mathcal{C} \cup \{A\}$, otherwise
(ii) if $\mathcal{C} \cap N(A) \neq \emptyset$, add $(\mathcal{C} \cap N(A)) \cup \{A\}$ to the set of active cliques.

Finally, duplicates and cliques that are subsets of others are removed.

If the current element ℓ of the list is a right endpoint, we output all cliques that contain at least one interval ending at ℓ. These cliques go out of scope and are thus maximal. We remove intervals ending at ℓ from active cliques. Cliques that become empty are removed from the set of active cliques.

2.2.2 Run-time analysis Let k be an upper bound on local alignment coverage, c be the maximum number of active cliques and s be the size of the output. The detailed run-time analysis of Section A in the Supplementary Material gives a total running time of $\mathcal{O}(m(\log m + kc^2) + s)$. Despite these rather moderate worst-case guarantees, our algorithm is very fast in practice. See the Supplementary Material, Section A, for an analysis of the corresponding reasons.

2.2.3 P-values for cliques We proceed as outlined in the Section 1.1.2. Let $\mathcal{C}$ be a max-clique in the read alignment graph and let $w(\mathcal{C}) := \sum_{A \in \mathcal{C}} w(A) = \sum_{A \in \mathcal{C}} \mathbf{P}_0(A)$ be the *weight of the clique*. Let $\bar{I}(\mathcal{C}) := \frac{1}{w(\mathcal{C})} \cdot \sum_{A \in \mathcal{C}} w(A) \cdot I(A)$ be the *weighted mean of alignment interval*

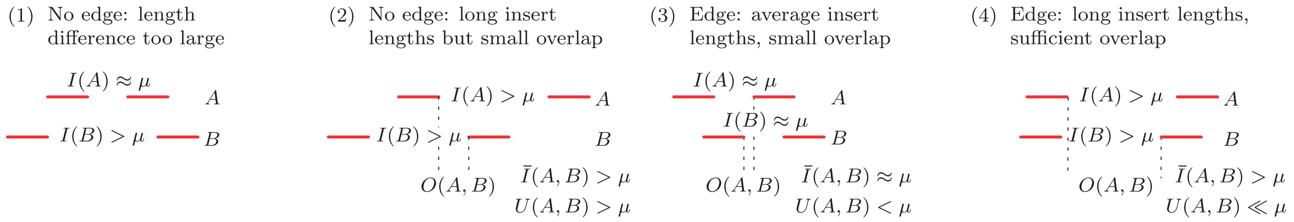

Fig. 2. Four scenarios of two overlapping alignment pairs A and B. In the *read alignment graph*, two alignments are connected by an *edge* if they are compatible, i.e. they support the same phenomenon. **(1)** Alignment A has an insert length about the expected insert length μ, suggesting that there is no variation present but alignment B has an insert length much larger than μ suggesting a deletion. Hence, A and B are not compatible. **(2)** Both alignments have similar insert lengths larger than μ, both suggesting a deletion of size $I(A) - \mu \approx I(B) - \mu$, but the overlap $O(A, B)$ is too small to harbor a deletion of this size. Thus, they are incompatible. **(3)** Both alignments do not suggest any variation and are therefore compatible. **(4)** Similar to Case (2), but now the overlap is large enough to contain the putative deletion

length of the clique. Let Φ be the standard normal distribution function. Let $\rho(\mathcal{C})$ be the number of alignments that are at the genomic location of the clique. For example, in Figure 1, $\rho(C_1) = \rho(C_2) = 7$ is just the number of alignments that overlap with one another at this position of the reference. We compute

$$p(\mathcal{C})_D := 2^{\rho(\mathcal{C})} \sum_{J \subset \mathcal{C}} \mathbf{P}_{H_0}(A_J)[1 - \Phi(\sqrt{|J|}\frac{\bar{I}(\mathcal{C}) - \mu}{\sigma})] \qquad (8)$$

$$p(\mathcal{C})_I := 2^{\rho(\mathcal{C})} \sum_{J \subset \mathcal{C}} \mathbf{P}_{H_0}(A_J)[\Phi(\sqrt{|J|}\frac{\bar{I}(\mathcal{C}) - \mu}{\sigma})] \qquad (9)$$

just as in Equations (3) and (2) with the difference that we distinguish between cliques, which give rise to deletions and insertions. $2^{\rho(\mathcal{C})}$ is the number of subsets of alignments one can test at this location, that is the virtual number of tests which we perform, so multiplying by $2^{\rho(\mathcal{C})}$ is a Bonferroni-like correction. This correction accounts for coverage fluctuations.

If $p(\mathcal{C})_D$ is significantly small then $\bar{I}(\mathcal{C})$ is significantly large; hence, the alignments in $\mathcal{C}$ are deemed to commonly support a deletion. Analogously, if $p(\mathcal{C})_I$ is significantly small, then $\mathcal{C}$ is supposed to support an insertion. Refer to Supplementary Material, Section B, for details on how the exponential sums in Equations (8) and (9) can be computed efficiently.

3 RESULTS AND DISCUSSION

3.1 Simulation: Craig Venter reads

We downloaded the comprehensive set of annotations of both homozygous and heterozygous structural variants (also including inversions and all other balanced rearrangements) for Craig Venter's genome, as documented by Levy *et al.* (2007) and introduced them into the reference genome, thereby generating two different alleles. If nested effects lead to ambiguous interpretations, we opted for an order that respects the overall predicted change in copy number. We used UCSC's SimSeq (https://github.com/jstjohn/SimSeq) as a read simulator to simulate Illumina paired-end reads with read end length 100, insert size mean $\mu = 112$ (we recall: distance between the inner ends of the sequenced reads) and standard deviation $\sigma = 15$, which reflects many biological datasets (see below). See Section J in the Supplementary Material for performance rates on $\mu = 500, \sigma = 50$ that highlights the limitations of insert size based approaches. Coverage 15× for each of the two alleles yields 30× sequence coverage overall.

3.2 Biological data: NA18507

We were further provided with reads of the genome of an individual from the Yoruba in Ibadan, Nigeria, by Illumina. Reads were sequenced on a GAIIx and are now publicly available (ftp://ftp.sra.ebi.ac.uk/vol1/ERA015/ERA015743/srf/). Read ends are of length 101. Read coverage is 30 ×, furthermore $\mu \approx 112, \sigma \approx 15$ (see the following paragraph). For benchmarking purposes, we used annotations from Mills *et al.* (2011, Gen.Res.) merged with NA18507 'DIP' annotations from the HGSV Project (http://hgsv.washington.edu/general/download/SNPs_DIPs) database, lifted to hg18.

3.3 Reference genome and alignments

As a reference genome, we used version hg 18, as downloaded from the UCSC Genome Browser. All reads considered were aligned using BWA (Li and Durbin, 2009) with the option to allow 25 alignments per read end, which amounts to a maximum of 25^2 alignments per paired-end read. BWA determined mean insert size $\mu \approx 112$ and standard deviation $\sigma \approx 15$ for both simulated and NA18507 reads. Note that we are aware that realignment of discordant reads with a more precise (but time consuming!) alignment tool, such as Novoalign (http://www.novocraft.com/main/index.php) (as suggested by Quinlan *et al.*, 2010), can lead to subsequent resolution of much misaligned sequence and hence to improved results for all tools considered.

3.4 Experiments

For benchmarking, we considered five different state-of-the-art insert size based approaches, four of which are applicable for a whole-genome study: GASV (Sindi *et al.*, 2009), VH (Hormozdiari *et al.*, 2009, v3.0), Breakdancer (Chen *et al.*, 2009) and HYDRA (Quinlan *et al.*, 2010). We ran MoDIL (Lee *et al.*, 2009) only on Chromosome 1 of the simulated data which, on our machines, required several hundred CPU hours. In contrast, we process Chromosome 1 in less than 1 h. We also consider the split-read aligners PINDEL (Ye *et al.*, 2009) and SV-seq2 (Zhang *et al.*, 2012). Details on program versions and on how we ran each method are given in Supplementary Material, Section C. In case of deletions, we define a *hit* as a pair of a true deletion and a predicted deletion that overlap and whose lengths do not differ by more than 100 bp, which roughly is the mean of internal segment size. We say that a true insertion (B_0, L_0) and a predicted insertion (B_1, L_1), where B is for breakpoint, L is for length, *hit* each other if the intervals $[B_0+1, \ldots, B_0+L_0]$ and $[B_1+1, \ldots, B_1+L_1]$ overlap. This 'overlap criterion' precisely parallels the one for deletions: if one views deletions in the reference as insertions in the donor, then the deletions in the reference (relative to reference coordinates) hit *if and only if* the insertions in the donor hit (relative to donor coordinates). Again, we also require $|L_0 - L_1| \leq 100$. We also offer results on alternative hit criteria which, instead of overlap, depend on fixed thresholds on breakpoint distance and differences of indel length in Supplementary Material, Section F. As usual, *recall* = TP/(TP + FN), where TP (= true positives) is the number of true deletions being hit and FN (= false negatives) is the number of true deletions not being hit. For Precision = TP/(TP + FP), TP is the number of predicted indels being hit and FP is the number of predicted indels not being hit. We relate recall and precision to one another and also display the *F*-measure, *F* = 2*Recall*Precision/(Recall + Precision), as a common overall statistic for performance evaluation. We refer to Exc. (= exclusive) as the percentage of true annotations, which were *exclusively (and correctly)* predicted by the method in question. Because the annotations for the biological dataset are obviously still far from complete, a false positive may in fact be due to a missing annotation. We therefore call the ratio TP/(TP + FP) relative precision (RPr.). For recall on the biological data, note that a good amount of existing annotations may be of limited reliability. Therefore, the *F*-measure is meaningless for these data and we refrain from displaying it. Last but not least, we present average deviation of breakpoint placement and differences in length for all tools in the Supplementary Material, Section G. In Supplementary Material, Section H, we present CLEVER's results on simulated data when including *true alignments* in the BAM files, or even using *only true alignments* so as to analyze its behavior relative to removal of external sources of errors.

3.5 Results

See Table 1 for performance figures. See also Section E in the Supplementary Material for a further subdivision of the 100–50 000 bp part. Boldface numbers designate the best approach, and italic numbers the best insert size based approach (if not the best approach overall). Comparing absolute numbers of true indels in the biological data with the simulated data points out immediately that the vast majority of annotations is

Table 1. Benchmarking results for simulated (Venter) and biological data (NA18507)

Dataset	Venter insertions				Venter deletions				NA18507 insertions			NA18507 deletions		
	Prec.	Rec.	Exc.	F	Prec.	Rec.	Exc.	F	RPr.	Rec.	Exc.	RPr.	Rec.	Exc.
Length range 20–49 (8786 true ins., 8502 true del.)									(2295 true ins., 2192 true del.)					
CLEVER	62.5	**53.0**	**20.4**	**57.4**	60.4	**66.8**	**15.9**	**63.4**	*7.7*	*24.1*	*8.4*	*8.9*	*44.7*	*6.6*
BreakDancer	—	5.1	0.1	—	*75.5*	7.5	0.0	13.6	—	0.3	0.0	8.2	5.8	0.0
GASV	NA	NA	NA	NA	5.4	25.8	1.8	8.9	NA	NA	NA	1.0	20.1	2.0
HYDRA	0.0	0.0	0.0	—	—	0.1	0.0	—	0.0	0.0	0.0	—	0.0	0.0
VH	32.4	8.4	0.2	13.4	66.3	8.0	0.3	14.3	0.8	3.8	0.4	4.6	4.6	0.3
PINDEL[a]	**66.1**	44.9	18.7	53.5	49.5	55.8	12.1	52.5	**13.1**	**40.0**	**25.3**	9.3	**64.9**	**26.3**
SV-seq2[a]	NA	NA	NA	NA	**96.0**	1.2	0.0	2.3	NA	NA	NA	**15.2**	1.6	0.2
Length range 50–99 (2024 true ins., 1822 true del.)									(303 true ins., 294 true del.)					
CLEVER	60.4	**86.6**	**7.3**	**71.2**	72.7	**80.7**	**6.8**	**76.5**	1.6	**70.3**	**6.9**	5.5	**79.6**	**12.2**
BreakDancer	**86.5**	56.5	0.2	68.3	**87.3**	48.1	0.3	62.0	*6.4*	15.5	0.0	*9.8*	44.2	0.7
GASV	NA	NA	NA	NA	46.1	35.0	1.5	39.8	NA	NA	NA	2.3	34.7	1.0
HYDRA	0.0	0.0	0.0	—	—	5.2	0.0	—	0.0	0.0	0.0	—	2.4	0.0
VH	55.8	76.6	1.4	64.5	66.5	65.8	1.5	66.1	1.4	62.7	2.3	4.3	57.1	1.4
PINDEL[a]	77.5	20.5	0.3	32.5	72.5	37.5	0.4	49.4	**10.8**	29.7	1.3	8.3	43.9	0.3
SV-seq2[a]	NA	NA	NA	NA	83.6	19.8	0.2	32.0	NA	NA	NA	**9.9**	28.6	0.3
Length range 100–50 000 (3101 true ins., 2996 true del.)									(165 true ins., 414 true del.)					
CLEVER	**66.2**	23.8	2.0	35.1	**87.6**	**69.9**	**4.1**	**77.7**	0.5	31.5	1.8	4.8	**70.3**	**2.7**
BreakDancer	61.0	17.6	3.0	27.4	65.8	57.7	0.0	61.5	0.9	23.0	1.8	*5.2*	62.1	0.5
GASV	NA	NA	NA	NA	0.9	49.2	1.0	1.7	NA	NA	NA	0.1	57.7	2.4
HYDRA	0.0	0.0	0.0	—	72.8	56.8	0.4	63.8	0.0	0.0	0.0	2.0	65.5	0.5
VH	60.4	**25.5**	**3.5**	**35.8**	58.8	65.1	1.5	61.8	**1.8**	**44.9**	**10.9**	3.0	70.0	1.4
PINDEL[a]	—	1.9	0.0	—	84.7	39.5	0.1	53.9	—	0.6	0.0	**5.9**	51.9	0.2
SV-seq2[a]	NA	NA	NA	NA	81.6	37.5	0.3	51.3	NA	NA	NA	3.9	34.5	0.0

Performance rates as recall, precision, exclusive predictions (Exc. which are true predictions, uniquely predicted by that tool) and *F*-measure are grouped by different indel size ranges. Dash and NA stands for 'no prediction' and 'not applicable', respectively. Insertions significantly exceeding the internal segment size (≈ 112 here) cannot be detected by insert size based approaches. PINDEL does not detect such insertions either.

[a] Split-read approach.

still missing seemingly. Therefore, all results on the biological data, in particular those on precision, can only reflect certain trends. For the simulated data, all values reflect the ground truth. As expected, performance rates greatly depend on the size of the indels. For prediction of indels shorter than 20 bp, split-read based approaches and/or read alignment tools themselves are the option of choice.

20–49 bp: CLEVER outperforms all other approaches on the simulated data and is the best insert size based approach also on the biological data. PINDEL achieves best rates on the biological data. Also, CLEVER makes a substantial amount of exclusive calls in all categories. Tables in the Supplementary Material, subsection F.2, points out that 80–90% of CLEVER's indel calls come *significantly close* to a real indel. Further analyses (Supplementary Material, Section H) demonstrate that 30% of CLEVER's false positives are due to misalignments and mapping ambiguities (see *External sources of errors* below). Obviously many of those extremely close but not truly hitting calls are due to external errors. Breakdancer makes little and highly precise calls at the expense of reduced accuracy in terms of indel breakpoint placement and length (see Supplementary Material, Section G).

50–99 bp: Here, CLEVER achieves substantially better recall and more exclusive calls than PINDEL on the biological data. On the simulated data, CLEVER again achieves best overall performance. In contrast to 20–49 bp, however, Breakdancer and VH already make significant contributions. Although VH achieves good overall performance, Breakdancer mostly excels in precision. As before, when allowing a certain offset of breakpoints (Supplementary Material, subsection F.2) or when integrating correct alignments (Supplementary Material, Section H), CLEVER's precision substantially rises from 60–72% to 72–96% across the categories.

100–50 000 bp: Also, CLEVER is best while other tools (Breakdancer, HYDRA, VH) also make decisive contributions. This documents that the current challenges for indel discovery are rather in the size range of 20–100 bp. Note that none of the tools makes predictions for insertions longer than 250 bp, see Section E in the Supplementary Material.

MoDIL: We compared MoDIL with all other tools on Chromosome 1 alone because of the excessive run-time requirements of MoDIL (CLEVER is faster by a factor of ~1000). See Supplementary Material, Section I. Overall, MoDIL incurs certain losses in performance with respect to CLEVER across all categories, but outperforms the other insert size based approaches apart from larger indels (≥100 bp). It is noteworthy that MoDIL makes a substantial amount of exclusive calls for insertions of 50–99 bp.

Accuracy of breakpoint and length predictions: See Section G for related numbers. The split-read based approaches outperform the insert size based approaches. Among the latter, CLEVER and GASV are most precise for 20–49 and 100–50 000 bp. For 50–99 bp calls, Breakdancer achieves favorable values.

External sources of errors: See Supplementary Material, Section H, for related results and a detailed discussion on to what degree misalignments and multiply mapped reads/alignment hamper computational SV discovery.

Conclusion: We have presented a novel internal segment size based approach for discovering indel variation from paired-end read data. In contrast to all previous, whole-genome-applicable approaches, our tool takes all concordant read data into account. We outperform all prior insert size based approaches on indels of sizes 20–99 bp and also achieve favorable values for long indels. We outperform the split-read based approaches considered on medium-sized (50–99 bp) and larger (≥100 bp) indels. In addition, our approach detects a substantial amount of variants missed by all other approaches, in particular, in the smallest size range considered (20–49 bp). In conclusion, CLEVER makes substantial contributions to SV discovery, in particular, in the size range of 20–99 bp.

Our approach builds on two key elements: first, an algorithm that enumerates maximal, statistically contradiction-free ensembles as max-cliques in read alignment graphs in short time and, second, a sound statistical procedure that reliably calls max-cliques that indicate variants. Our approach is generic with respect to choices of variants; max cliques in the read alignment graphs can also reflect other variants such as inversions or translocations. For future work, we are planning to predict inversions and to incorporate split read information as a unifying approach.

Conflict of Interest: none declared.

REFERENCES

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Albers,C.A. *et al.* (2011) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.

Alkan,C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bravo,H.C. and Irizarry,R.A. (2010) Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, **66**, 665–674.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chen,K. *et al.* (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Hach,F. *et al.* (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods*, **7**, 576–577.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Itsara,A. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.*, **84**, 148–161.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6** (**11 Suppl.**), S13–S20.

Mills,R. *et al.* (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830–839.

Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res.*, **16**, 1182–1190.

Quinlan,A.R. *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.*, **20**, 623–635.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.

Sudmant,P.H. *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.

The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Wasserman,L. (2004) *All of Statistics*. Springer, New York.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zhang,J. *et al.* (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics*, **13**, S6.