CHAPTER 7

DISCOVERING AND GENOTYPING TWILIGHT ZONE DELETIONS

Tobias Marschall and Alexander Schönhuth

Centrum Wiskunde & Informatica, Amsterdam, Netherlands

title, edition.

7.1 Introduction

Although next-generation sequencing (NGS) experiments have become standard, the exploration of the data still poses challenges. NGS experiments usually aim at providing catalogues of genetic variants, to be used in downstream analyses of interest. In population studies, such as the "Genome of the Netherlands" or the "1000 Genomes" initiatives [5, 26], such catalogues aim to reflect the full extent of genetic diversity of populations. In cancer genome studies (see [27] for a global initiative), comprehensive lists of somatic variants are sought that help to understand cancer (sub)types and disease progression, and to select appropriate therapy protocols.

In this article, we focus on techniques for next-generation re-sequencing studies, which allow to study the differences between a *donor genome*, a genome to be investigated, and a reference genome. The workflow common to a re-sequencing study proceeds according to the following steps:

1. The DNA of the genome of interest is broken into fragments.

2. The fragments are next-generation sequenced, which yields *reads*. Thereby, a very popular and helpful technique is to generate *paired-end reads*, fragments both ends of which are sequenced with an internal part, the *internal segment*, which remains unsequenced.

3. Reads are mapped onto the reference genome, whose sequence is known in its entirety. Mapping requires *read aligners*, algorithms that allow to align reads with the reference.

4. One then tries to infer the differences between the genome under study and the reference genome, that is the genetic variants that affect the genome of interest, from the mapped reads.

See Figures 7.1, 7.2 and 7.4 for schematics on scenarios that can result from mapping paired-end reads onto a reference genome.

The computational exploration of certain classes of genetic variants, such as single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) have become standard. See, for example, the GATK [19] website for best-practice workflows. Treating other classes of variants, however, such as translocations, inversions or nested combinations of simpler variants, is still often non-standard or requires computationally advanced techniques. See for example [1, 20] for reviews on the discovery of structural variants.

Although difficult in general and still lacking best-practice workflows, the analysis of some classes of structural variants has become routine. A predominant example is the discovery of deletions of more than 200 base pairs (bp), which has been addressed by a large variety of approaches: examples are Breakdancer [2], VariationHunter [6], (MATE-)CLEVER [16, 17], DELLY [22], GASV(-Pro) [24, 25], see also the references therein, and again the above-mentioned reviews[1, 20].

Twilight Zone Deletions. In this article, we focus on deletions that are hard to discover because of their length. As mentioned above, very short deletions as well as long deletions are no longer posing fundamental difficulties, or even have become part of best-practice workflows. *Midsize deletions*, however, which we refer to as *"NGS twilight zone deletions"*, have been posing substantial computational challenges also after 2009. Only most recent advances have made their discovery possible [16, 17, 32]. Evidence of this is the fact that catalogues of deletions resulting from projects [5], where [16, 17, 32] have been in use, finally contain comprehensive amounts of such twilight zone deletions, with excellent validation rates. This is in stark contrast to earlier, related projects (in particular the 1000 Genomes project [26]).

In this chapter, we review why discovery and genotyping of mid-size deletions has been difficult and explain the techniques by which this became possible.

The organization of this chapter is as follows.

- In Section 7.2, we provide the necessary notation.
- In Section 7.3, we give the formal definition of "twilight zone" deletions. We briefly revisit the different approaches suitable for deletion discovery in resequencing studies, and we outline their pitfalls when it comes to discovering mid-sized ("twilight zone") deletions.
- In Section 7.5, we present a novel maximum likelihood approach for genotyping deletions which achieves highly favorable performance rates on twilight zone indels.
- In Section 7.6, we evaluate a comprehensive selection of state-of-the-art tools on NGS reads from a genome containing real variants (Venter's genome [12]), where NGS reads are simulated by means of the Assemblathon [3] read simulator, and current NGS technology (Illumina HiSeq and MiSeq reads).
- In Section 7.7, we discuss all results presented and point out challenges that are still open.

7.2 Notation

We predominantly focus on paired-end read data, the most widely used data in resequencing studies. Let $\Sigma = \{A, C, G, T, N\}$ be the set of nucleotides, augmented by a character (N) which represents nucleotides that could not be properly read. Throughout this chapter, reads $R = (R_1, R_2) \in (\Sigma^K)^2$ are pairs of strings of length K (where $K = 100$ in Illumina HitSeq, or $K = 250$ in Illumina MiSeq experiments) over Σ. Here, R_1 is the left and R_2 is the right end of R. We refer to single positions in the ends R_i for $i = 1, 2$ by $R_i[t]$ where $t \in [1, K]$. Let $I(R)$ be the length of the *internal segment* between the two sequenced ends R_1, R_2 of a paired-end read $R = (R_1, R_2)$. While sequence and, hence, length of the ends R_1, R_2 are known—we

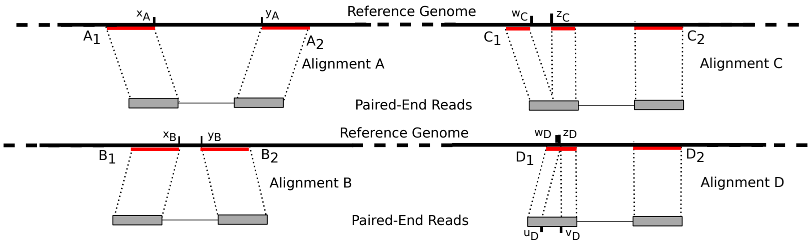

Figure 7.1 A: Alignment whose interval length indicates a deletion, B: alignment whose interval length indicates an insertion, C: alignment where a split (in the left end) indicates a deletion, D: alignment where a split (in the right end) indicates an insertion

recall that the length of R_1, R_2 is K—neither the sequence of the internal segment nor its length, $I(R)$, is known. This, of course, implies that the length of the entire fragment $2K + I(R)$ is not known either.

We write $\mathcal{G}$ for the reference genome which we also consider as a sequence over Σ.

Alignments. See Figure 7.1 for the following.

We write $A(R) = (A_1, A_2)$ for an alignment of read $R = (R_1, R_2)$ against the reference. We write x_A, y_A for the rightmost reference position of the alignment of the left read end, and the leftmost reference position of the alignment of the right end. We write $I(A) := y_A - x_A - 1$ for the length of the *alignment interval* of A.

Gaps/Splits. Alignments $A = (A_1, A_2)$ can be gapped, where gaps either indicate insertions or deletions. For the sake of simplicity, we assume that each alignment is affected by at most one gap—note that alignments of NGS fragments containing two gaps are extremely rare. For notational simplicity, we will assume in the examples and explanations to follow that A_1 displays a gap. We write w_A for the reference position that precedes the gap, and z_A for the reference position that immediately follows the gap. In turn, we refer to the position in the read that precedes the gap as u_A and the position in the read that follows the gap as v_A, that is, the reference nucleotide $\mathcal{G}[w_A]$ aligns with $R_1[u_A]$ and $\mathcal{G}[z_A]$ aligns with $R_1[v_A]$. Depending on whether $z_A = w_A + 1$ (see alignment D in Figure 7.1) or $v_A = u_A + 1$ (see alignment C), the gap indicates an insertion in the donor genome (where the inserted sequence is $R[u_A + 1, v_A - 1]$) or a deletion in the donor genome (where the deleted sequence is $\mathcal{G}[w_A + 1, z_A - 1]$).

Deletions. Let D_L and D_R be the reference coordinates of the left and right breakpoint of a deletion D. That is, reference nucleotides from (and including) position D_L till (and including) position D_R, which together form the sequence $\mathcal{G}[D_L, D_R]$, are missing in the donor genome. Let $C(D) := \frac{D_L + D_R}{2}$ be the *centerpoint* of D (which need not be an integer) and $L(D) := D_R - D_L + 1$ be the length of the

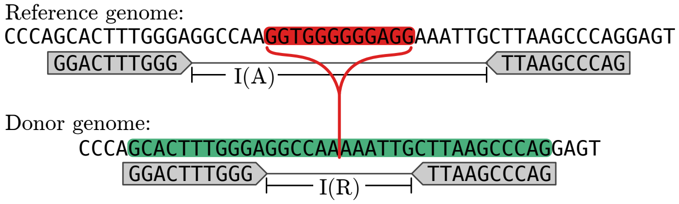
Reference genome:
CCCAGCACTTTGGGAGGCCAAGGTGGGGGGAGGAAATTGCTTAAGCCCAGGAGT
GGACTTTGGG
I(A)
TTAAGCCCAG
Donor genome:
CCCAGCACTTTGGGAGGCCAAAAATTGCTTAAGCCCAGGAGT
GGACTTTGGG
I(R)
TTAAGCCCAG

Figure 7.2 Internal-segment-size based evidence for a deletion: the piece of sequence colored in red is present in the reference but deleted in the donor genome. The length $I(R)$ of the fragment that is sequenced (in green) is determined during library preparation. When mapped back to the reference the internal segment $I(A)$ is longer than $I(R)$ due to the deletion.

deletion. We parametrize the deletion $D = (C(D), L(D))$ by its centerpoint $C(D)$ and its length $L(D)$.

7.3 Non-Twilight-Zone Deletion Discovery

Approaches available for discovering deletions from NGS read data roughly fall into four different categories: internal-segment-size based, split-read based, coverage based and assembly based approaches; see [1] for a detailed review. We solely focus on the first two types of approaches and their hybrids here. Coverage-based approaches can only discover deletions of usually at least 1000 bp in length. Assembly-based approaches face challenges that have not yet been entirely overcome. Note again that very short deletions of length up to 20 bp, can be discovered already during the initial mapping stage and therefore pose no unusual computational challenges. While common internal segment size based approaches work well for deletions of length greater than approximately 150 bp, split read aligners are able to discover also deletions longer than 20 bp, usually reaching their limits at 30-40 bp. Beyond 40 bp, deletion discovery recall of split-read aligners usually substantially drops.

7.3.1 Internal-Segment-Size Based Approaches

The basic idea that underlies internal-segment-size based approaches is that the alignment interval length $I(A)$ deviates from $I(R)$, the length of the internal segment of the read that gave rise to A, by the length of a deletion affecting the internal segment of R. That is, $I(A) = I(R) + L(D)$ for a deletion D in the internal segment of R, see Figure 7.2 for an illustration. While $I(A)$ is known, $I(R)$ is not, however. Therefore, one estimates both mean and standard deviation of the empirical distribution of $I(R)$ from uniquely mappable fragments using robust estimators [13, 14]. For modern DNA sequencing protocols, the fragment length distributions are approximately Gaussian with low standard deviations. We note already here that well-shaped (Gaussian) distribution are essential for discovering midsize deletions.

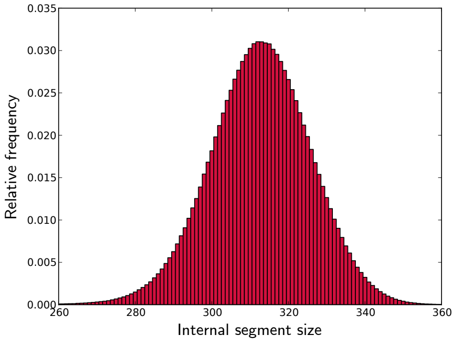

Figure 7.3 Internal Segment Size Distribution for GoNL individual

See Figure 7.3 for such a distribution, derived from the NGS reads of one of the individuals of the GoNL project [5], all of which were sequenced by BGI in 2010.

After determination of the internal segment size distribution, the following generic workflow for discovering deletions in NGS data is widely used:

1. Collect all reads whose alignments *statistically significantly deviate* in terms of alignment interval length, so-called *discordant reads*.

2. Cluster all such reads into groups that support the same deletion.

3. Make predictions from the resulting groups of discordant reads.

The majority of approaches follows this workflow (e.g. [2, 6, 8, 21, 24, 25]). They can be distinguished by their definition of discordant read, their clustering/grouping techniques, and their details in deriving predictions from groups of discordant reads. Note that handling of reads that became multiply mapped due to repetitive sequence often plays a major role [29], see for example [6] for a combinatorially principled approach.

The key factor is the definition of a discordant read, as those are supposed to represent fragments whose internal segment is affected by deleted sequence. Again, the idea is that $I(A)$, the alignment interval length of a discordant read significantly deviates from the distribution of $I(R)$. Thus, the definition of a discordant read depends on the standard deviation of the distribution of $I(R)$, which in current protocols amounts to about 15. Up to 6-7 times this standard deviation are required to obtain sufficiently low, genome-wide false discovery rates. This then translates into the fact that deletions shorter than 100 bp remain undiscoverable.

7.3.2 Split-Read Mapping Approaches

Split-read mapping approaches aim at making direct use of alignment information. As per a usual workflow, a split-read mapper processes only reads that standard read aligners fail to align correctly. This is most often due to insertions and deletions

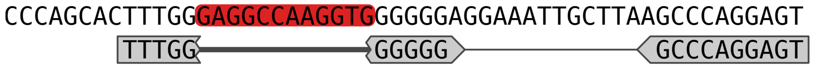

Figure 7.4 Split-Read Evidence for Deletion

that affect the read. Standard read aligners usually face difficulties in aligning such reads properly, because correctly placing longer gaps can be computationally (too) expensive.

Therefore, aligning reads affected by longer indels requires extended techniques: "split-read alignments". A generic workflow common to many split-read aligners (e.g. [4, 23, 32, 34]) looks as follows:

1. Collect all reads where one end remained unaligned and/or where one end became only partially aligned ("soft-clipped") by the standard aligner in use.

2. In case of entirely unaligned read ends, split the end in parts, or "seeds", and try to align those parts, or "seeds" (see Figure 7.4 for a resulting alignment).

3. For aligned such parts and/or for soft-clipped reads, try to align the remaining part(s) of the read somewhere "nearby".

4. For each such read, collect all possible partial alignments, and compute "split alignments", using banded alignment techniques, to connect them. Output the most likely such split alignment(s) as the alignment(s) of the read end.

To date, common split-read aligners can successfully align reads with non-negligible amounts of deletions of length up to 40-50 bp. While aligning reads exhibiting larger deletions is not impossible, the discovery rates of split-read aligners significantly decrease with increasing deletion size. Thereby, recall, that is the rate of discovered deletions, usually drops below 60-70% when reaching the 30 bp mark, which renders split-read based approaches following the workflow from above unable to discover sufficient amounts of "twilight zone deletions".

The bottleneck of split-read aligners are step 2 and 3 in the workflow from above. Split or yet unaligned parts of reads can be small, which can drastically increase the number of locations in the reference genome where these parts can be aligned. The fact that genomes are highly repetitive in general can significantly add to these difficulties—resolving the resulting ambiguity among those multiple mappings is another involved step.

Therefore, one has to limit the size of the regions in which one searches for alignments of split parts. Due to those limitations, although substantially raising the limits of standard aligners, also split-read aligners can quickly reach their limits—every implemenational detail can count. Note that the internal segment size distribution plays a decisive role also here, as it is used to appropriately quantify "nearby" in step 3 and as a guide when placing read alignments of split parts in step 2.

7.3.3 Hybrid Approaches

One decisive general advantage of split-read aligning approaches over internal segement size based approaches is the base pair resolution of deletion breakpoints: if both an internal segment size based method and a split-read aligner call a deletion, the breakpoints predicted by the split-read aligner are usually much more accurate than those of the internal segment size based approach.

However, as outlined above, split-read aligners usually cannot detect many deletions larger than 30-40 bp. The motivation of so-called hybrid approaches is to call also breakpoints of large deletions at base pair resolution. A common, generic workflow thus is:

1. Run an internal segment size-based approach and collect all deletion calls.
2. Collect all alignments nearby deletion calls collected in step 1.
3. Split-align all unaligned and partially aligned read ends in those regions. Thereby, the split-aligner is "guided" by the deletion calls of the internal segment size based approach when determining the correct placements of shorter read end parts.
4. Output all variant calls with breakpoints corrected (or removed, if no split alignments could be determined) as per the split alignments determined in step 3.

The result usually are calls for large deletions whose breakpoints come at base pair resolution. See [7, 17, 22, 33] for most prevalent approaches. In essence, the major bottleneck of hybrid approaches is step 1, that is, they inherit the computational bottlenecks of internal segment size-based approaches in terms of size range limitations.

7.3.4 The "Twilight Zone": Definition

We conclude that both internal segment size-based and hybrid approaches can discover deletions at sufficient power only in the size range of 100 bp ($\sim$6-7 times the standard deviation of the internal segment size distribution) and larger. Split-read aligners, on the other hand, are able to discover deletions only of size up to 30 bp ($\sim$2 times the standard deviation). We consequently suggest 2 to 6-7 times the standard deviation in terms of base pairs as the *"twilight zone" of NGS deletions*.

7.4 Discovering "Twilight Zone" Deletions: New Solutions

Since 2012, new solutions have been presented for discovering deletions of length 30–100 bp, at both sufficient power and precision [16, 17]. The earliest approach that immediately addressed the discovery of midsize insertions and deletions was MoDIL [11]. While successful by principle, MoDIL is too slow in practice: a single genome, sequenced at the standard coverage of 30x, needs more more than 3 days on a large computer cluster, which is no option in a population-scale genome project. In contrast, CLEVER [16] needs only 6-8 hours on a single CPU. CLEVER also signficantly outperforms MoDIL both in terms of recall and precision in discovery. We therefore focus on CLEVER and its relatives in the following. We will also mention PINDEL [32], as the possibly most favorable contemporary split read aligner, which can make considerable contributions in the twilight zone.

7.4.1 CLEVER

The key insight of CLEVER [16] is that exchanging steps 1 and 2 in the workflow outlined in subsection 7.3.1 leads to success when aiming at the discovery of insertions and deletions smaller than 100 bp. CLEVER clusters read alignments before discarding non-discordant reads. The point is that, although single concordant (= non-discordant) reads make no strong enough statistical signal for a deletion on their own, groups of concordant reads, all of which transmit a rather weak statistical signal for a deletion, can together "bundle up" to form a strong signal for a deletion. CLEVER aims at the discovery of such group signals.

While this sounds easy in principle, it is not in practice. The clear advantage of the previous approaches was that the workflow from subsection 7.3.1 allowed to discard all concordant reads, which drastically reduces the number of reads to be processed. Thus, step 2 translated into a clustering problem of decisively smaller scale: instead of clustering billions of reads, only small fractions of discordant reads, which come in amounts that are smaller by orders of magnitude, needed to be grouped and further processed. CLEVER clusters *all read alignments*. To achieve low enough runtimes in practice, CLEVER makes use of a highly-engineered, ultra-fast implementation of a max-clique enumeration technique as underlying clustering algorithm.

CLEVER solves this problem by formally collecting all read alignments into a *read alignment graph*, where each node represents a read alignment and edges indicate that two overlapping alignments are likely to reflect identical alleles. Maximal cliques represent maximal groups of read alignments all of which reflect the same allele (see the right panel of Figure 1 in [16]). If there are indel alleles in the donor genome, the max-cliques reflecting such alleles deviate from the internal segment size statistics. If sufficiently many read alignments participate in such a max-clique, they give rise to statistically significant signals even when reflecting only relatively small indels, as revealed by common multiple-sample Z-tests. The statistical model of CLEVER further allows to address that read alignments can be ambiguous due to repetitive sequence, and corrects for multiple testing, thereby keeping control of the false discovery rate.

Key to success for enumerating all such max-cliques finally is a bitvector-driven implementation of a max-clique enumeration algorithm that exploits the particular structure of read alignment graphs. See [16] and also [28] for details and corresponding runtime analyses.

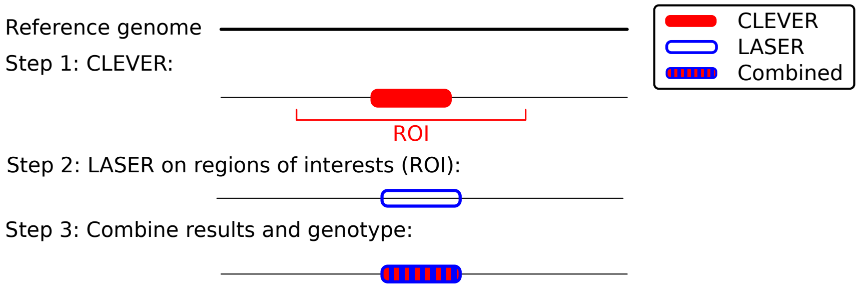
Reference genome
Step 1: CLEVER:
ROI
Step 2: LASER on regions of interests (ROI):
Step 3: Combine results and genotype:
CLEVER
LASER
Combined

Figure 7.5 MATE-CLEVER. First, the internal-segment-size based tool CLEVER discovers deletions (red). The split-read aligner LASER then finds corresponding split-read alignment (blue) in the respective regions. The resulting prediction (red-blue) is that of LASER, as split-read aligner discover deletion breakpoints at higher accuracy.

7.4.2 MATE-CLEVER

While CLEVER discovers midsize deletions at both high recall and precision (see section 7.6 below), the accuracy of the predicted breakpoints suffers from the usual deficits that are common to internal segment size based approaches. MATE-CLEVER, as a hybrid approach, aims at curing this issue, and does not only discover deletions at high recall and precision, but also at high accuracy of their breakpoints. See Figure 7.5 for an illustration of MATE-CLEVER.

The workflow of MATE-CLEVER [17] is that of a common hybrid approach, see subsection 7.3.3. Thereby, it makes use of CLEVER in the first step. Subsequently, in step 3, it makes use of a novel split-read aligner, LASER [18], which has been particularly trimmed to compute highly accurate split-read alignments reflecting also larger gaps. The output of MATE-CLEVER are mid-size (and long-size) deletions, discovered by CLEVER, where breakpoints are corrected and therefore highly accurate.

7.4.3 PINDEL

PINDEL [32] is a split-read aligner that specializes in discovering also deletions longer than 20 bp at extremely high precision, with more than 90% of all calls being true positives that are also highly accurate in terms of breakpoint annotations, across all size ranges. In doing so, it achieves clearly the highest recall rates among all (split-read) alignment based approaches. Note that GSNAP [30, 31] achieves higher recall for deletions of 30-50 bp, which, however, comes at the price of reduced precision and (much) less accurate breakpoint annotations.

Overall, PINDEL follows the workflow common to split-read aligners. Its achievements are due to an accumulation of improvements in the fine details, which in combination yield a superior method. We refer the reader to [32], the original publication, for details.

7.5 Genotyping “Twilight Zone” Deletions

7.5.1 A Maximum Likelihood Approach under Read Alignment Uncertainty

Let G_i for $i = 0, 1, 2$ represent the genotypes of an indel, where G_0 indicates absence of the indel, G_1 indicates that the indel is heterozygous, and G_2 indicates that the indel is homozygous. Let A be a read alignment. Let $\mathcal{R}$ be all reads. For $R \in \mathcal{R}$, let $A(R)$ be the alignment of R with the region we would like to genotype. We write A^+ for the event that A is the correct alignment of R, and we write A^- for the event that it is not. Note that $\mathbb{P}(A^-) = 1 - \mathbb{P}(A^+)$. We further formally consider each read $R \in \mathcal{R}$ as the disjoint union of the two events $A^+(R)$ and $A^-(R)$. Let $\mathcal{S} \subset \mathcal{R}$ be a subset of reads. In slight abuse of notation, we also consider $\mathcal{S}$ as the event that precisely the alignments of reads from $\mathcal{S}$ are correct, while all others are not. Hence,

$$\mathbb{P}(\mathcal{S}) = \prod_{R \in \mathcal{S}} \mathbb{P}(A^+(R)) \cdot \prod_{R \notin \mathcal{S}} (1 - \mathbb{P}(A^+(R)) \tag{7.1}$$

is the corresponding probability.

In the following, we consider a maximum likelihood (ML) setting, which in particular reflects that our prior belief in genotypes is the same for all types:

$$\mathbb{P}(G_0) = \mathbb{P}(G_1) = \mathbb{P}(G_2) = \frac{1}{3}. \tag{7.2}$$

Making use of an ML approach allows to attain an efficient computation scheme. We point out below that a full Bayesian approach is infeasible, due to being exponential in the number of reads that align with the region of interest—note that we wish to genotype hundreds of thousands of regions of interest, such that runtime considerations are a crucial factor.

We are interested in maximizing

$$\mathbb{P}(G_i \mid \mathcal{R}) \propto \mathbb{P}(G_i, \mathcal{R}) = \sum_{\mathcal{S} \subset \mathcal{R}} \mathbb{P}(\mathcal{S}) \cdot \mathbb{P}(G_i \mid \mathcal{S}). \tag{7.3}$$

By taking probabilities $\mathbb{P}(\mathcal{S})$ into account, we would like to appropriately address alignment uncertainty, which can be due to several factors such as multiple mappings and alignment artifacts. Let $K := |\mathcal{R}|$ be the number of reads that align to the region to be genotyped. By Bayes’ formula, Equation (7.2) further implies that

$$\begin{aligned}\mathbb{P}(G_i \mid \mathcal{S}) &\overset{(7.2)}{\propto} \mathbb{P}(\mathcal{S} \mid G_i) \\ &= \prod_{R \in \mathcal{S}} \mathbb{P}(A^+(R) \mid G_i) \cdot \prod_{R \notin \mathcal{S}} \mathbb{P}(A^-(R) \mid G_i) \\ &\overset{(7.2)}{\propto} \prod_{R \in \mathcal{S}} \mathbb{P}(G_i \mid A^+(R)) \cdot \prod_{R \notin \mathcal{S}} \mathbb{P}(G_i \mid A^-(R))\end{aligned} \tag{7.4}$$

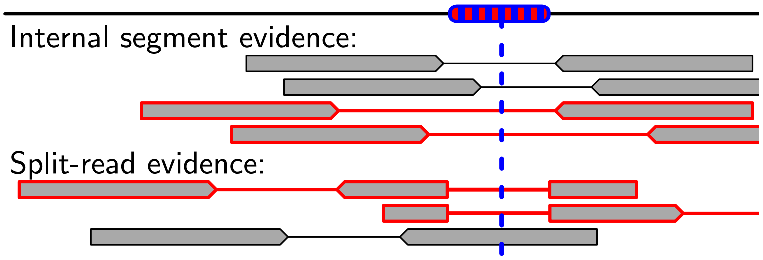

Figure 7.6 Different types of evidence for a heterozygous variant. While the gray alignment rather provide evidence against a deletion, the alignments in red rather provide evidence for it. In case of internal segment evidence, the red alignments A reflect the case $\mathcal{N}_{\mu+L,\sigma}(I(A)) > \mathcal{N}_{\mu,\sigma}(I(A))$ in (7.9), whereas the gray alignments reflect the opposite case.

where the equality is justified by assuming that reads have been generated independently of one another. Note that the computation (7.4) is not possible in the frame of a fully Bayesian approach, because only the assumption (7.2) of constant priors implies the first proportionality. This renders such an undertaking infeasible. From (7.4), we conclude that

$$\begin{aligned}\mathbb{P}(G_i \mid \mathcal{R}) &\propto \sum_{\mathcal{S}\subset\mathcal{R}} \mathbb{P}(\mathcal{S}) \cdot \prod_{R\in\mathcal{S}} \mathbb{P}(G_i \mid A^+(R)) \cdot \prod_{R\notin\mathcal{S}} \mathbb{P}(G_i \mid A^-(R)) \\ &\overset{(7.1)}{=} \sum_{\mathcal{S}\subset\mathcal{R}} \prod_{R\in\mathcal{S}} \mathbb{P}(A^+(R))\mathbb{P}(G_i \mid A^+(R)) \cdot \prod_{R\notin\mathcal{S}} (1-\mathbb{P}(A^+(R)))\mathbb{P}(G_i \mid A^-(R)) \\ &= \prod_{R\in\mathcal{R}} \Big[\mathbb{P}(A^+(R))\mathbb{P}(G_i \mid A(R)) + (1-\mathbb{P}(A^+(R)))\mathbb{P}(G_i \mid A^-(R))\Big] \qquad (7.5)\end{aligned}$$

where the second row results from expanding the third row. The last term, finally, can be computed in time linear in the number of reads $\mathcal{R}$, which had been our goal.

It remains to compute reasonable probabilities $\mathbb{P}(G_i \mid A^+)$ and $\mathbb{P}(G_i \mid A^-)$ for read alignments A. While

$$\mathbb{P}(G_i \mid A^-) = \mathbb{P}(G_i) \qquad (7.6)$$

is obviously reasonable, because the read that underlies A does not stem from the region to be genotyped, computation of terms $\mathbb{P}(G_i \mid A^+)$ require further reasoning, based on the type of evidence that A can provide about G_i.

One has to distinguish the following two cases (see Figure 7.6):

1. *Split-read evidence:* A aligns with the region of interest such that one read end stretches across the (potential) variant

2. *Internal-segment based evidence:* A aligns with the region of interest such that the internal segment of its read pair stretches across the (potential) variant

Split-read evidence. Let us first consider the case of no alignment uncertainty. That is, if read alignments are correct, then they precisely reflect the differences between the donor and the reference.

Let D be the deletion to be genotyped and let A be an alignment where, for example, A_1 stretches across the breakpoints of D. Under the assumption of no alignment uncertainty, we obtain that A stems from a chromosomal copy that is affected by D if and only if w_A and z_A precisely agree with the left and right breakpoint of D. If the split disagrees with the deletion breakpoints or there is no split, the read behind A stems from a chromosomal copy that is not affected by the deletion with probability one.

Let *A be an alignment with a split/gap that agrees with D*. By the above considerations, the read behind A stems from a chromosomal copy that is affected by D. By Bayes' formula, and (7.2), $\mathbb{P}(G_i \mid A^+) \propto \mathbb{P}(A^+ \mid G_i)$. First, $\mathbb{P}(A^+ \mid G_0) = 0$, because the read behind A cannot stem from the region, and $\mathbb{P}(A^+ \mid G_1) = \frac{1}{2}(\mathbb{P}(A^+ \mid G_0) + \mathbb{P}(A^+ \mid G_2))$, which reflects that one first randomly selects one of the two chromosomal copies, only one of which is affected by D, and then generates the read from it. The *case of B being an alignment in disagreement with D* is treated analogously, where in this case $\mathbb{P}(B^+ \mid G_2) = 0$. Transforming this into a posterior probability consequently yields

$$\mathbb{P}(G_i \mid A^+) := \begin{cases} \frac{2}{3} & i = 2 \\ \frac{1}{3} & i = 1 \\ 0 & i = 0 \end{cases} \quad \text{and} \quad \mathbb{P}(G_i \mid B^+) := \begin{cases} 0 & i = 2 \\ \frac{1}{3} & i = 1 \\ \frac{2}{3} & i = 0 \end{cases}. \tag{7.7}$$

In general, however, the assumption of no alignment uncertainty does not hold. In fact, split-read alignments can be affected by several sources of errors, the most evident of which are probably repetitive areas, such that both position and length of alignment splits disagree with the positions and the length of the true variants—nevertheless the split is indeed due to the variant. Therefore, we declare A, where $C(A) := (w_A + z_A)/2$ is the centerpoint and $L(A) := z_A - w_A$, in case of deletions, (see Figure 7.1) is the length of the split in A, to support the deletion $D = (C(D), L(D))$, with centerpoint $C(D)$ and of length $L(D)$ iff

$$|C(D) - C(A)| \leq 50 \quad \text{and} \quad |L(D) - L(A)| \leq 20. \tag{7.8}$$

While these values may seem large, they are well supported by statistics on the uncertainty of (split-)alignment.

Internal-Segment-Based Evidence Internal-segment-based evidence is provided by evaluating the empirical statistics on fragment length inherent to the library the read stems from. We develop this part here in view of fragment length statistics being approximately Gaussian. However, this can be easily generalized to arbitrary empirical statistics. Let D be the deletion to be genotyped and let $C(D)$ be its centerpoint. Let R be the read that has given rise to alignment A where $x_A < C(D) < y_A$, that is, the alignment interval of A contains the centerpoint of the breakpoints of D. Note that a centerpoint-oriented selection leads to a situation that

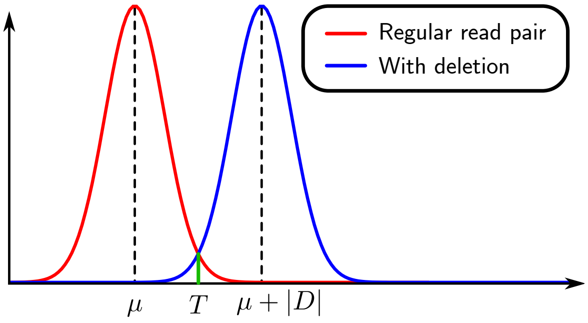

Figure 7.7 Gaussian distribution on interval size for alignments of normal reads (read) and reads indicating a deletion of length $|D|$. Alignments whose intervals are of length T provide no evidence, as both the existence and the non-existence of the deletion are equally likely.

is balanced in terms of choosing equal amounts of reads that provide evidence for and against the deletion, as outlined in [17]. Let μ and σ be mean and standard deviation of the internal segment size distribution of the library R stems from. So, internal segment length, as a random variable X, is distributed as the normal distribution $X \sim \mathcal{N}_{\mu,\sigma}$ for the library under consideration. There are two cases: first, alignments A *whose reads stem from a chromosomal copy that is not affected by* D, and second aligments B *whose reads stem from a chromosomal copy that is affected by* D. We obtain

$$I(A) \sim \mathcal{N}_{\mu,\sigma} \qquad \text{and} \qquad I(B) \sim \mathcal{N}_{\mu+L,\sigma}, \tag{7.9}$$

where the second case reflects that the alignment interval contains the deletion of length L. Refer to Figure 7.7 for an illustration. We compute that $\mathbb{P}(A^+ \mid G_0) \propto \mathcal{N}_{\mu,\sigma}(I(A))$ and $\mathbb{P}(A^+ \mid G_2) \propto \mathcal{N}_{\mu+L,\sigma}(I(A))$ as appropriate densities for the cases of no variant and a homozygous variant.

Let $Z := \frac{3}{2}(\mathcal{N}_{\mu,\sigma}(I(A)) + \mathcal{N}_{\mu+L,\sigma}(I(A)))$. In analogy to considerations for the split-read case, we arrive at

$$\mathbb{P}(G_i \mid A^+) := \begin{cases} \frac{1}{Z} \cdot \mathcal{N}_{\mu+L,\sigma}(I(A)) & i = 2 \\ \frac{1}{3} & i = 1 \\ \frac{1}{Z} \cdot \mathcal{N}_{\mu,\sigma}(I(A)) & i = 0 \end{cases} \tag{7.10}$$

as an appropriate probability distribution for reads whose alignments span the breakpoints of deletions by their internal segments.

The procedure described above is implemented as part of MATE-CLEVER [17], which can use prior information in form of the Mendelian laws, if the input consists of multiple, ancestry-related genomes. In order to genotype, MATE-CLEVER takes all (split-read) alignments resulting from step 3 of the generic hybrid approach workflow (see subsection 7.3.3), and executes the genotyping-related computations from above, by plugging Equations (7.7) and (7.10) into Equation (7.5), thereby inferring the most likeliest genotype.

Tool	Read mapper	Internal segment	Split reads	Genotyping	Version	Ref.
Bowtie2	X				2.1.0	[9]
Breakdancer		X			1.4.4	[2]
BWA	X				0.7.5a	[13]
CLEVER		X			v2.0-rc3	[16]
DELLY		X	X		0.0.11	[22]
GATK				X	2.8-1-g932cd3a	[19]
GSNAP	X				2014-01-21	[30, 31]
MATE-CLEVER		X	X	X	v2.0-rc3	[17]
PINDEL			X		0.2.4t	[32]
Socrates			X		–	[23]
Stampy	X				1.0.23	[15]
VariationHunter		X			0.3	[6]

Table 7.1 List of used software tools. *Read mapper:* Programs ticked in that column are standard read mappers. *Internal segment:* SV detection methods that use internal-segment-size information. *Split reads:* SV detection methods based on split-read aligment. *Genotyping:* Methods able to genotype SVs. *Version:* The given version was used in our experiments.

7.6 Results

7.6.1 Dataset

We downloaded all variant annotations for Craig Venter's genome from the HuRef database [12]. We generated a diploid genome using those annotations, as per the procedure described in [17] (to generate a 'father' genome), which results in a genome that is realistic in terms of both amounts of variants and also zygosity status of variants. Note that direct usage of the annotations results in a genome with an unrealistic ratio of heterozygous to homozygous deletions (it vastly overrated homozygous indels), which is likely due to the difficulties in determining the zygosity status of insertions and deletions during the original assembly stage. Here, we have resolved these issues such that the ratio of heterozygous and homozygous deletions is realistic.

Subsequently, we simulated reads for each of those copies, using Simseq, the read simulator of the Assemblathon [3], at 15x coverage, which results in 30x coverage overall. We opted to simulate reads according to two prevalent and most recent Illumina protocols: HiSeq and MiSeq, where mean and standard deviation for the size of the internal segment were 112 and 15 (HiSeq) and 250 and 15 (MiSeq) respectively.

7.6.2 Tools

Table 7.1 gives an overview of the used tools. We aimed at selecting state-of-the-art tools from different categories. From the internal-segment-size based methods, we chose Breakdancer, CLEVER, and VariationHunter. Pindel and Socrates represent split-read approaches, where MATE-CLEVER and DELLY are hybrids between internal segment size and split read. Furthermore, we included four standard read mappers in the analysis: Bowtie2, BWA (MEM), GSNAP, and Stampy. Although these tools do not exlicitly target deletion discovery, they have some capabilities of mapping reads across deletions. For these tools, we extracted all deletions that were contained in two or more read alignments to compile a set of predictions.

7.6.3 Discovery

See tables 7.2 and 7.3 for the following. We ran all tools on the two (HiSeq and MiSeq) datasets described in subsection 7.6.1. In tables 7.2 and 7.3, tools are grouped by the class of approach they belong to. The first group are internal segment size based, the second group are split-read alignment based, the third group are hybrid, and the fourth group are direct alignment based approaches.

We evaluate all tools in terms of four different categories.

1. *Strict precision*, which is the fraction of calls where the centerpoint of the breakpoints deviates by not more than 20 bp and the length by not more than 10 bp from that of a true deletion

2. *Relaxed precision* allows deviations of 100 bp for both centerpoint placement and length. Note that such calls are still statistically highly significant, because the deviations are small relative to genome length and overall numbers of calls, hence are still of great potential interest to the researcher. In essence, these calls just require further refinement.

3. We also evaluate the callsets in terms of *Recall (hom.)* and *Recall (het.)*, which are the fractions of *true homozygous* and *true heterozygous* deletions that were correctly discovered, according to the criteria for relaxed precision.

As becomes immediately clear from tables 7.2 and 7.3, split-read, hybrid and direct alignment based approaches clearly outperform the internal segment size based approaches in terms of accuracy of breakpoint annotation, as indicated by the much improved strict precision rates. It is also obvious that in the lower part of the twilight zone (see table 7.2), (split-read) alignment approaches have certain advantages, where we note that, among the alignment based approaches PINDEL excels in terms of precision, while GSNAP excels in terms of recall, on HiSeq data. On MiSeq data, BWA-MEM has clear advantages. Approaches from the other classes that achieve high recall in the lower part are CLEVER and MATE-CLEVER, certainly because they are the only such approaches tailored towards discovery of twilight zone deletions.

Tool	Prec. (strict)	Prec. (relaxed)	Recall (hom.)	Recall (het.)
	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq
Length 10–29				
BreakDancer	0.0 / 0.0	83.7 / 32.8	0.6 / 0.2	0.3 / 0.0
CLEVER	38.4 / 30.3	80.3 / 76.9	25.1 / 17.7	6.9 / 0.7
VariationHunter	3.3 / 0.4	89.1 / 53.7	0.7 / 0.4	0.5 / 0.0
PINDEL	**91.2 / 91.3**	93.0 / 93.4	**89.0 / 92.4**	**80.7 / 83.8**
SOCRATES	8.0 / 4.5	11.5 / 7.1	1.4 / 0.3	1.2 / 0.3
DELLY	– / –	– / –	0.0 / 0.0	0.0 / 0.0
MATE-CLEVER	87.3 / 89.6	**93.2 / 94.1**	23.1 / 15.3	5.8 / 2.7
Bowtie2	90.5 / 33.2	92.2 / 35.1	61.4 / 82.2	51.5 / 73.5
BWA MEM	84.8 / 80.0	88.5 / 85.0	79.6 / 86.1	72.4 / 80.8
GSNAP	69.6 / 68.4	90.0 / 90.9	83.6 / 85.5	75.2 / 73.7
Stampy	35.0 / 20.8	65.3 / 46.1	83.8 / 86.4	78.3 / 47.3
Length 30–49				
BreakDancer	7.5 / 3.0	81.4 / 37.3	18.6 / 5.7	8.1 / 0.1
CLEVER	26.2 / 19.6	71.2 / 69.2	**90.5** / 80.7	**61.8** / 15.1
VariationHunter	17.1 / 5.6	88.1 / 68.0	34.6 / 12.0	16.2 / 0.5
PINDEL	77.0 / 83.7	84.1 / 90.2	65.6 / 79.2	54.4 / 67.2
SOCRATES	44.7 / 32.7	49.1 / 35.6	9.9 / 5.7	8.2 / 3.3
DELLY	– / –	– / –	0.0 / 0.1	0.1 / 0.0
MATE-CLEVER	81.4 / **86.1**	89.8 / **93.0**	76.8 / 70.5	51.8 / 38.4
Bowtie2	87.5 / 71.7	**100.0** / 75.4	2.8 / 63.6	2.0 / 47.5
BWA MEM	**88.8** / 79.1	92.1 / 86.2	26.0 / **85.7**	21.5 / **75.3**
GSNAP	43.9 / 56.7	69.5 / 82.9	72.3 / 83.3	57.4 / 61.3
Stampy	57.4 / 33.8	85.1 / 66.0	56.3 / 84.2	46.2 / 43.5
Length 50–69				
BreakDancer	8.4 / 1.3	82.0 / 23.5	40.7 / 11.3	39.6 / 0.4
CLEVER	30.3 / 21.2	75.3 / 75.3	86.0 / **78.7**	70.2 / 16.6
VariationHunter	21.2 / 8.9	79.7 / 61.2	**86.7** / 56.0	**70.6** / 2.8
PINDEL	64.0 / 76.2	72.8 / 83.9	48.0 / 72.3	35.4 / **53.6**
SOCRATES	29.3 / 27.0	40.4 / 33.3	8.3 / 6.3	6.2 / 4.6
DELLY	– / –	– / –	1.3 / 0.0	1.0 / 0.0
MATE-CLEVER	**73.2 / 85.4**	**84.0 / 92.4**	66.0 / 64.0	51.4 / 41.8
Bowtie2	– / 74.0	– / 74.0	1.0 / 13.7	0.2 / 4.0
BWA MEM	– / 79.9	– / 86.2	5.3 / 64.0	3.2 / 46.2
GSNAP	36.7 / 46.2	65.1 / 74.4	20.0 / 25.0	17.6 / 9.2
Stampy	60.7 / 40.4	78.5 / 73.8	24.7 / 67.7	18.2 / 19.8

Table 7.2 Results for SV prediction tools on 30x data for deletions from 10 to 69 bp.

Tool	Prec. (strict)	Prec. (relaxed)	Recall (hom.)	Recall (het.)
	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq
Length 70–99				
BreakDancer	5.8 / 1.9	85.3 / 25.8	59.6 / 12.3	50.7 / 0.6
CLEVER	38.0 / 23.4	88.8 / 78.9	81.3 / **77.9**	60.2 / 16.9
VariationHunter	15.7 / 12.0	83.1 / 67.4	**83.0** / 77.0	**79.1** / 11.5
PINDEL	56.1 / 72.1	64.6 / 80.9	38.3 / 65.5	27.2 / 39.0
SOCRATES	49.1 / 55.8	54.7 / 60.5	7.2 / 8.5	10.9 / 8.9
DELLY	66.7 / –	66.7 / –	0.4 / 2.1	2.0 / 0.0
MATE-CLEVER	80.9 / **89.6**	90.1 / **94.6**	57.4 / 63.4	41.5 / **42.1**
Bowtie2	– / –	– / –	0.9 / 3.0	0.9 / 0.0
BWA MEM	– / 81.4	– / 86.4	4.3 / 40.4	2.6 / 22.3
GSNAP	– / –	– / –	13.2 / 12.3	7.7 / 0.0
Stampy	**87.5** / 34.9	**100.0** / 67.4	20.0 / 54.0	12.3 / 2.0
Length 100–149				
BreakDancer	3.1 / 0.2	78.2 / 21.8	48.7 / 19.3	48.6 / 0.4
CLEVER	31.4 / 21.6	83.6 / 70.8	69.0 / 64.0	48.6 / 14.7
VariationHunter	5.2 / 5.5	65.3 / 60.6	**76.6 / 72.6**	**66.9** / 8.0
PINDEL	61.1 / 77.6	65.5 / 85.3	16.8 / 40.1	15.9 / 24.3
SOCRATES	36.0 / 41.9	44.0 / 48.4	5.6 / 5.6	5.6 / 4.8
DELLY	40.7 / 50.0	69.5 / 80.4	6.1 / 6.1	8.0 / 5.2
MATE-CLEVER	**75.5** / 77.9	**85.4** / 87.5	44.2 / 47.7	30.3 / **32.3**
Bowtie2	– / –	– / –	0.0 / 1.0	0.8 / 0.0
BWA MEM	– / **100.0**	– / **100.0**	1.5 / 7.1	2.8 / 0.0
GSNAP	– / –	– / –	8.6 / 5.1	5.6 / 0.0
Stampy	– / 0.0	– / 0.0	6.6 / 20.8	5.6 / 0.0
Length 150–199				
BreakDancer	1.9 / 0.3	41.5 / 18.3	40.7 / 19.8	30.9 / 0.7
CLEVER	32.0 / 16.3	**86.4** / 67.0	61.5 / 63.7	36.2 / 13.8
VariationHunter	4.3 / 3.8	35.8 / 31.4	**70.3 / 67.0**	**63.8** / 8.6
PINDEL	55.4 / 69.0	58.5 / 81.0	20.9 / 34.1	11.2 / 18.4
SOCRATES	20.0 / 31.0	22.9 / 34.5	6.6 / 7.7	2.0 / 2.6
DELLY	25.0 / 56.4	45.0 / 74.4	9.9 / 12.1	7.2 / 9.9
MATE-CLEVER	**74.7 / 83.5**	86.1 / **91.3**	37.4 / 46.2	19.7 / **30.9**
Bowtie2	– / –	– / –	0.0 / 0.0	0.0 / 0.0
BWA MEM	– / –	– / –	0.0 / 0.0	0.0 / 0.0
GSNAP	– / –	– / –	0.0 / 0.0	0.0 / 0.0
Stampy	– / –	– / –	0.0 / 0.0	0.0 / 0.0

Table 7.3 Results for SV prediction tools on 30x data for deletions from 70 to 199 bp.

In the upper part of the twilight zone (see table 7.3), internal segment size based and hybrid approaches have clear advantages over (split-read) alignment based approaches, because the recall rates of alignment based approaches drops to zero or the precision considerably suffers. Among the alignment based approaches, PINDEL clearly has the best recall rates on longer deletions, at least at sufficiently high precision, and therefore makes a valuable contribution in those size ranges. Note that among the internal segment size based approaches, VariationHunter puts clear emphasis on recall, which comes at the expense of inaccurate breakpoint annotations, as indicated by very low strict precision.

The only approaches that achieve high recall rates across all size ranges of the twilight zone are CLEVER and MATE-CLEVER. The only real weakness are heterozygous deletions of length 30-50 bp, where, however, none of the other tools achieve better recall rates on HiSeq data. In this category, stepping up from HiSeq to MiSeq data, and making use of BWA-MEM is a very helpful option. On HiSeq data, heterozygous deletions of 30-50 bp can still be considered a weak spot for *all* approaches, with CLEVER achieving 61% recall, as the best performance rate.

When comparing HiSeq to MiSeq data in general, an immediate observation is that alignment based approaches tend to achieve higher recall on MiSeq data, but sometimes at the cost of lower precision. Internal segment size based approaches often incur non-negligible losses on MiSeq data, which is likely due to the longer average internal segment size for the MiSeq dataset.

7.6.4 Genotyping

See Tables 7.4 and 7.5 for the following. We evaluated five state-of-the-art protocols for genotyping deletions. We ran the *UnifiedGenotyper (UG)* and the *HaplotypeCaller (HC)* on both BWA-MEM and GSNAP alignments to produce genotyped deletion calls. We also ran MATE-CLEVER in genotyping mode, as outlined in Section 7.5.1.

It becomes evident, that, in an overall statement, MATE-CLEVER is the most favorable approach when it comes to genotyping deletions longer than 30 bp, both at sufficiently high recall (see statistic 'Number of Calls') and good precision (see column 'homozygous (correct)' in Table 7.4 and 'heterozygous (correct)' in Table 7.5). Usually, MATE-CLEVER predicts homozygosity in about 90%, and heterozygosity in about 80% of the cases correctly.

While MATE-CLEVER seemingly is the most favorable approach overall, the other approaches have certain partial strengths. Most notably, both BWA-HC and GSNAP-HC achieve better performance rates than MATE-CLEVER on homozygous deletions of 30-49 bp, both in terms of recall and precision. It remains to add, however, that the value of this remains somewhat unclear, because these tools do not achieve similarly good rates on heterozygous deletions.

When comparing UG with HC, one observes that HC leads to considerable increases in terms of recall over UG, while incurring certain losses in terms of genotyping precision. In conclusion, UG is seemingly the more conservative postprocessing method, while HC is a more aggressive variant calling postprocessor for

Tool	True annotation →	absent	heterozygous	homozygous
	↓ Number of Calls ↓	(wrong call)	(wrong type)	(correct)
	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq
Length 10–29				
BWA / UG	6109 / 7365	7.4 / 9.3	2.9 / 3.9	92.2 / 90.0
BWA / HC	6327 / 6750	9.2 / 10.0	2.7 / 2.4	90.2 / 89.8
GSNAP / UG	6392 / 6943	6.8 / 7.1	2.7 / 3.3	92.8 / 92.4
GSNAP / HC	6380 / 6721	9.3 / 9.9	2.7 / 2.4	90.1 / 89.8
MATE-CLEVER	1546 / 1222	6.7 / 6.7	7.0 / 8.6	91.8 / 90.8
Length 30–49				
BWA / UG	147 / 794	8.2 / 9.9	2.0 / 4.3	91.2 / 88.8
BWA / HC	567 / 700	10.9 / 10.4	2.6 / 2.6	88.2 / 89.1
GSNAP / UG	633 / 785	13.4 / 11.0	2.2 / 3.1	86.1 / 88.4
GSNAP / HC	579 / 686	10.7 / 10.5	3.1 / 2.6	88.1 / 89.1
MATE-CLEVER	758 / 743	9.8 / 7.9	7.4 / 10.4	86.5 / 85.3
Length 50–69				
BWA / UG	0 / 172	– / 12.8	– / 4.1	– / 86.6
BWA / HC	127 / 207	18.9 / 15.9	2.4 / 4.8	79.5 / 83.1
GSNAP / UG	31 / 43	6.5 / 9.3	3.2 / 0.0	90.3 / 90.7
GSNAP / HC	128 / 180	18.0 / 13.9	1.6 / 2.8	81.2 / 85.6
MATE-CLEVER	236 / 225	14.0 / 9.3	11.0 / 8.9	78.8 / 86.2
Length 70–99				
BWA / UG	0 / 55	– / 14.5	– / 5.5	– / 83.6
BWA / HC	81 / 117	14.8 / 13.7	2.5 / 1.7	85.2 / 86.3
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	80 / 115	17.5 / 12.2	1.2 / 0.9	82.5 / 87.8
MATE-CLEVER	135 / 154	11.9 / 5.8	3.7 / 5.2	85.2 / 90.3
Length 100–149				
BWA / UG	0 / 1	– / 0.0	– / 0.0	– / 100.0
BWA / HC	28 / 72	14.3 / 13.9	0.0 / 0.0	85.7 / 86.1
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	33 / 73	18.2 / 15.1	0.0 / 0.0	81.8 / 84.9
MATE-CLEVER	100 / 107	15.0 / 12.1	2.0 / 0.9	84.0 / 86.9
Length 150–199				
BWA / UG	0 / 0	– / –	– / –	– / –
BWA / HC	0 / 11	– / 0.0	– / 9.1	– / 90.9
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	2 / 11	100.0 / 18.2	0.0 / 9.1	0.0 / 72.7
MATE-CLEVER	39 / 51	7.7 / 7.8	7.7 / 9.8	92.3 / 84.3

Table 7.4 Genotyping performance for **homozygous** calls.

Tool	True annotation →	absent	heterozygous	homozygous
	↓ Number of Calls ↓	(wrong call)	(correct)	(wrong type)
	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq	HiSeq / MiSeq
Length 10–29				
BWA / UG	6188 / 10856	8.4 / 10.8	85.5 / 84.9	7.7 / 6.0
BWA / HC	10182 / 12464	11.6 / 10.7	80.5 / 80.6	10.1 / 11.3
GSNAP / UG	7748 / 9848	7.9 / 7.7	86.2 / 86.7	7.5 / 7.1
GSNAP / HC	10619 / 12506	11.6 / 11.3	80.3 / 80.0	10.4 / 11.2
MATE-CLEVER	1022 / 333	7.0 / 3.0	61.7 / 86.2	37.1 / 12.9
Length 30–49				
BWA / UG	119 / 971	7.6 / 9.6	87.4 / 86.2	9.2 / 6.5
BWA / HC	961 / 1456	19.6 / 10.4	72.3 / 78.4	11.1 / 15.0
GSNAP / UG	728 / 1024	22.9 / 12.3	71.2 / 83.4	8.7 / 6.2
GSNAP / HC	1122 / 1454	22.1 / 11.2	69.5 / 77.8	11.1 / 14.9
MATE-CLEVER	877 / 597	10.6 / 5.9	80.8 / 89.9	10.9 / 5.4
Length 50–69				
BWA / UG	0 / 177	– / 16.4	– / 79.7	– / 5.1
BWA / HC	262 / 438	32.8 / 14.4	53.4 / 69.4	15.6 / 18.9
GSNAP / UG	43 / 63	32.6 / 15.9	67.4 / 77.8	0.0 / 6.3
GSNAP / HC	315 / 410	32.1 / 18.3	54.6 / 66.6	15.6 / 18.5
MATE-CLEVER	313 / 233	17.6 / 6.0	75.4 / 91.4	7.3 / 3.9
Length 70–99				
BWA / UG	0 / 31	– / 16.1	– / 80.6	– / 6.5
BWA / HC	148 / 190	38.5 / 16.3	52.0 / 68.9	13.5 / 20.5
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	182 / 222	39.0 / 20.7	51.1 / 65.3	12.1 / 18.5
MATE-CLEVER	169 / 163	8.3 / 4.9	85.2 / 92.0	8.3 / 4.3
Length 100–149				
BWA / UG	0 / 0	– / –	– / –	– / –
BWA / HC	78 / 90	56.4 / 15.6	25.6 / 56.7	19.2 / 35.6
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	100 / 117	59.0 / 17.1	31.0 / 59.8	11.0 / 28.2
MATE-CLEVER	92 / 101	14.1 / 12.9	80.4 / 86.1	5.4 / 2.0
Length 150–199				
BWA / UG	0 / 0	– / –	– / –	– / –
BWA / HC	18 / 29	88.9 / 31.0	5.6 / 51.7	11.1 / 20.7
GSNAP / UG	0 / 0	– / –	– / –	– / –
GSNAP / HC	26 / 20	92.3 / 25.0	7.7 / 50.0	3.8 / 30.0
MATE-CLEVER	40 / 52	20.0 / 9.6	75.0 / 86.5	5.0 / 3.8

Table 7.5 Genotyping performance for **heterozygous** calls.

read alignments. Using HC for genotyping deletions longer than 50 bp seemingly is not an option, as precision on heterozygous deletions suffers quite substantially, achieving only 50-60%. In this size range, MATE-CLEVER seemingly is the only sound option that is available among the typing pipelines evaluated.

7.7 Discussion

In this chapter, we review the current state of the art about calling deletions of length 30-150 bp from NGS data. As deletions in this length range pose extraordinary computational and statistical challenges, they have been referred to as *"twilight zone deletions"*. Recent approaches [16, 17, MATE-/CLEVER], however, have pointed out novel and successful ways to discover twilight zone deletions at both good recall and high precision. Moreover, it was described in [17] how to reliably genotype twilight zone deletions, which we re-visit here in detail. In addition to those novel strategies, several well-maintained SV calling tools have constantly undergone improvements. Thereby, they have grown into methods by which one can at least make a good amount of calls in partial areas of the twilight zone [2, 6, 30, 31, 32]. As many of those callsets are complementary, combining MATE-/CLEVER with a reasonable selection of other tools, where we favor PINDEL, and on MiSeq data BWA-MEM in particular, should lead to successful twilight zone deletion calling pipelines. In essence, we consider the discovery of twilight zone deletions a resolved issue, at least when operating on carefully prepared sequencing libraries with reasoanably small standard deviations.

Here, we sketch the novel, successful strategies and evaluate a large range of tools, some of which have helped considerably shedding more light on the NGS twilight zone of deletions. In brief, while some advanced internal segment size based approaches "tackle" the twilight zone from above (among which [2, 6, 16]), alignment based approaches, both regular and split-read oriented (see [10, 13, 32]), tackle the twilight zone from below. A general disadavantage of internal segment size based approaches is that breakpoints predicted are rather inaccurate (see differences in *relaxed* and *strict* precision statistics). Hybrid approaches address this issue, and therefore enjoy the advantages of both internal segment size based approaches in terms of being able to call also larger deletions and alignment based approaches in terms of highly accurate breakpoint predictions.

7.7.1 HiSeq

Among all approaches evaluated, only CLEVER [16], and its hybrid version MATE-CLEVER [17] deliver comprehensive deletion callsets that span the entire size range of the twilight zone (30-150 bp). Among the internal-segment-size based callers, VariationHunter delivers good callsets for deletions of length at least 50 bp, and Breakdancer for deletions of length at least 70 bp. The high recall, however, comes at the cost of breakpoint accuracy, which can be considerably improved in both cases.

In this respect, CLEVER takes the lead among the internal-segment-size based approaches, as its strict precision rates are clearly superior.

Among the hybrid approaches tested, only MATE-CLEVER makes contributions in the twilight zone. DELLY clearly focuses on longer deletions and achieves very favorable performance rates for deletions longer than 200 bp (data not shown).

Among the (split-)alignment based approaches, PINDEL is best in terms of an overall assessment, achieving excellent performance rates for calls up to 50 bp, and also discovering non-negligible amounts beyond 50 bp. GSNAP and Stampy also deliver substantial amounts of excellent predictions, where, however, the recall of both tools becomes negligible for deletions of 40-50 bp and longer.

Heterozygous deletions 30-50 bp. A major challenge that has remained when processing HiSeq data are heterozygous deletions of length 30-50 bp. Here, CLEVER's recall is best (61%). Still, novel solutions are yet to be developed for this class of calls based on HiSeq experiments. The combination of heterozygosity and size range seemingly has (partially) remained a weak point of all classes of deletion discovery approaches, which can not (yet) entirely be overcome when making use of only HiSeq experiments.

Genotyping. In summary, one can recommend GSNAP-(UG/HC) for deletions of length up to 30 bp, and MATE-CLEVER for all deletions longer than 30 bp, with GSNAP-based genotyping pipelines also achieving competitive performance rates for homozygous deletions of 30-50 bp. Since one usually has little information on relative amounts of heterozygous and homozygous deletions, MATE-CLEVER is the superior overall choice for deletions longer than 30 bp, achieving an overall genotyping precision of greater than 90%, as the only tool on 30+ bp deletions.

7.7.2 MiSeq

MiSeq is a recent sequencing technology that allows for reads of length 250 bp and longer. As such, it holds major promises in terms of spotting also longer deletions by (split-)alignment based approaches. In fact, MiSeq experiments do not suffer from the "blind spot" of 30-50 bp heterozygous deletions, that still applies for HiSeq experiments. In particular BWA-MEM [13] and partially also GSNAP [30] profit from the advance in sequencing technology, achieving recall above 80% (BWA-MEM even 85.7%) at reliable precision rates (BWA-MEM: 86.2/79.1% relaxed/strict precision). If MiSeq technology is available and the throughput is sufficient for the application at hand, clearly these are the methods of choice.

For deletions of 50 bp and longer, CLEVER and MATE-CLEVER are again the methods of choice, achieving highest performance rates throughout, with VariationHunter as the only rival, whose calls, however, are highly inaccurate (Variation Hunter, strict precision $<10\%$).

Usage of MiSeq data leads to losses in performance for internal segment size based approaches, and hence also for hybrid approaches, which is most likely due to the increased internal segment length. For (split-)alignment based methods it often leads to a considerable increase in recall, while leading to losses in precision also here.

In summary, usage of MiSeq data in twilight zone deletion discovery has clear advantages when discovering heterozygous deletions of length 30-50 bp. In all other classes of calls, its usage leads to more "aggressive" twilight zone deletion calling, when used in combination with (split-)alignment based approaches, while HiSeq data based callsets tend to contain less false positives.

Genotyping. In the genotyping methods presented, MiSeq data usually leads to losses in recall while leading to enhanced precision for MATE-CLEVER. The explanation is that MATE-CLEVER is a hybrid approach. While MATE-CLEVER achieves best genotyping precision in general, for deletions longer than 30 bp, the HC-based tool combinations lead to more "aggressive" genotypers, with less genotyping precision, but also with a clear increase of calls that are deemed being "typable" for deletions of length 30-50 bp.

In summary, also on MiSeq data, MATE-CLEVER is the only approach that reliably types twilight zone deletion calls across all size ranges. For deletions from the lower ranges of the twilight zone, alignment based methods in combination with UG and/or HC can be helpful to generate callsets of "higher sensitivity".

7.7.3 Conclusion

HiSeq. As a general advice for HiSeq data, one can consider PINDEL the strongest approach for calling deletions of length 10-30 bp. For calls between 30-50 bp, a combination of PINDEL, CLEVER and/or MATE-CLEVER is recommended, which together should yield comprehensive callsets for homozygous deletions, but leave room for improvements for heterozygous deletions. From 50 bp on, when focusing on a single tool, MATE-/CLEVER are the methods of choice.

For genotyping, one can recommend usage of alignment based methods, such as BWA-MEM or GSNAP, in combination with UG and/or HC for deletions of length 10-30 bp (and shorter, data not shown), while MATE-CLEVER is the method of choice for deletions of length longer than 30 bp.

On a side remark, these insights lead to the selection of those tools in the most recent Genome-of-the-Netherlands project [5], which delivers callsets for deletions in these size ranges, which are decisively more comprehensive than those of other projects, such as the 1000 Genomes project [26].

MiSeq. A general advice for MiSeq data is to make use of PINDEL for calls of 10-30 bp, BWA-MEM for calls of 30-50 bp, and for MATE-/CLEVER for deletions longer than 50 bp. Still, also MiSeq data analysis technology leaves room for improvements: heterozygous deletions of length 70+ bp can still be considered to be extremely challenging, with no tool operating at a recall rate of 60% or higher, without making drastical sacrifices in terms of precision (note that VariationHunter sometimes achieves relatively high recall, without, however, achieving "operable" precision).

For genotyping, again BWA-MEM and/or GSNAP, postprocessed by UG and/or HC are the methods of choice for deletions shorter than 30 bp, while MATE-CLEVER is the method of choice for deletions of 30+ bp. For deletions of 30-50 bp, combining GSNAP with HC can make a valuable contribution, beyond using MATE-CLEVER.

[Make Snakefile available and mention where.]

7.8 Acknowledgments

A. Schönhuth acknowledges funding by the Nederlandse Wetenschappelijke Organisatie (NWO), through Vidi grant 639.072.309.

REFERENCES

1. Can Alkan, Bradley P. Coe, and Evan E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.
2. Ken Chen, John W. Wallis, Michael D. McLellan, David E. Larson, Joelle M. Kalicki, et al. Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6(9):677–681, Sep 2009.
3. D. Earl, K. Bradnam, J. St.John, A. Darling, D. Lin, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21:2224–2241, 2011.
4. Anne-Katrin Emde, Marcel H. Schulz, David Weese, Ruping Sun, Martin Vingron, Vera M. Kalscheuer, Stefan A. Haas, and Knut Reinert. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics*, 28(5):619–627, March 2012.
5. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 2014.
6. Fereydoun Hormozdiari, Can Alkan, Evan E Eichler, and S Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, July 2009.
7. Y. Jiang, Y. Wang, and M. Brudno. Prism: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, 28(20):2576–2583, 2012.

8. Jan O Korbel, Alexej Abyzov, Xinmeng Jasmine Mu, Nicholas Carriero, Philip Cayting, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009.

9. Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

10. Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

11. Seunghak Lee, Fereydoun Hormozdiari, Can Alkan, and Michael Brudno. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Meth*, 6(7):473–474, July 2009.

12. Samuel Levy, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, Sep 2007.

13. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

14. Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, Nov 2008.

15. Gerton Lunter and Martin Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Research*, 21(6):936–939, June 2011.

16. Tobias Marschall, Ivan G. Costa, Stefan Canzar, Markus Bauer, Gunnar W. Klau, Alexander Schliep, and Alexander Schönhuth. CLEVER: clique-enumerating variant finder. *Bioinformatics*, 28(22):2875–2882, November 2012.

17. Tobias Marschall, Iman Hajirasouliha, and Alexander Schönhuth. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics*, 29(24):3143–3150, 2013.

18. Tobias Marschall and Alexander Schönhuth. Sensitive long-indel-aware alignment of sequencing reads. Technical report, arXiv:1303.3520, 2013.

19. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, September 2010.

20. Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Meth*, 6(11s):S13–S20, November 2009.

21. Aaron R. Quinlan, Royden A. Clark, Svetlana Sokolova, Mitchell L. Leibowitz, Yujun Zhang, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, 20(5):623 –635, May 2010.

22. Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, September 2012.

23. J. Schroeder, A. Hsu, S.E. Boyle, G. MacIntyre, M. Cmero, R.W. Tothill, R.W. Johnstone, M. Shackleton, and A.T. Papenfuss. Socrates: Identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 2014. doi:10.1093/bioinformatics/btt767.

24. Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J. Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222–i230, June 2009.

25. Suzanne Sindi, S. Önal, L.C. Peng, H.-T. Wu, and Benjamin J. Raphael. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, 13:R22, 2012.

26. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

27. The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

28. A. Toepfer, T. Marschall, R.A. Bull, F. Luciani, A. Schönhuth, and N. Beerenwinkel. Viral quasispecies assembly via maximal clique enumeration. *PLoS Computational Biology*, 2014. to appear.

29. T.J. Treangen and S.L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13:557–567, 2012.

30. T.D. Wu and S. Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26:873–881, 2010.

31. T.D. Wu and C.K. Watanabe. Gmap: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21:1859–1875, 2005.

32. Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov 2009.

33. Jin Zhang, Jiayin Wang, and Yufeng Wu. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinformatics*, 13 Suppl 6:S6, 2012.

34. Z.D. Zhang, J. Du, H. Lam, A. Abyzov, A.E. Urban, M. Snyder, and M. Gerstein. Identification of genomic indels and structural variations using split reads. *BMC Genomics*, 12:375, 2011.