

Structural Variation Detection with Read Pair Information—An Improved Null-Hypothesis Reduces Bias

Kristoffer Sahlin¹(✉), Mattias Frånberg^{2,3}, and Lars Arvestad^{3,4}

¹ KTH Royal Institute of Technology, Science for Life Laboratory, School of Computer Science and Communication, Stockholm, Sweden
`ksahlin@kth.se`

² Atherosclerosis Research Unit, Department of Medicine, Solna, Karolinska Institutet, Stockholm, Sweden
`mattias.franberg@ki.se`

³ Department of Numerical Analysis and Computer Science, Stockholm University, Stockholm, Sweden
`arve@nada.su.se`

⁴ Swedish e-Science Research Centre (SeRC), Stockholm, Sweden

Abstract. Reads from paired-end and mate-pair libraries are often utilized to find structural variation in genomes, and one common approach is to use their fragment length for detection. After aligning read-pairs to the reference, read-pair distances are analyzed for statistically significant deviations. However, previously proposed methods are based on a simplified model of observed fragment lengths that does not agree with data. We show how this model limits statistical analysis of identifying variants and propose a new model, by adapting a model we have previously introduced for contig scaffolding, which agrees with data. From this model we derive an improved null hypothesis that, when applied in the variant caller CLEVER, reduces the number of false positives and corrects a bias that contributes to more deletion calls than insertion calls. A reference implementation is freely available at <https://github.com/ksahlin/GetDistr>.

1 Introduction

Genomic structural variation, for example insertion and deletion of DNA, are common in the human population and have been linked to various diseases and conditions. The basic question scientists and clinicians want to answer is: given a DNA sample from a donor and a suitable reference genome, what structural variants does the donor have in comparison to the reference? Methods for identifying structural variants are continuously worked on, in terms of both experimental protocols and bioinformatic analysis. Short-read technologies are, despite their weaknesses, the primary data source because of the superior throughput/cost ratio. It is today important to improve accuracy of predictions and in particular to reduce the false-positive rate while retaining sensitivity. To that end, we have

worked on improving the statistical analysis of paired reads, using paired-end (PE) or mate-pair (MP) libraries, for evaluating the significance of a detected insertion or deletion.

While aligned reads are important for identifying short variants and substitutions, larger variants and variants in repetitive regions where alignment is difficult are easier detected by paired reads spanning over the region. In PE and MP protocols, reads are from the ends of DNA *fragments* from the donor. PE libraries have short-range fragment lengths (up to 100s bp), MP libraries are long range (1000s bp), and they each have their own strengths and limitations. PE libraries often has superior coverage and narrow fragment length distribution while long range MP libraries can span larger insertions and, at similar read coverage, provide higher span coverage (the number of MP pairs separated by a random position) than PE libraries, which in theory can make up for the increased variation in individual fragment lengths by increasing statistical power from more observations.

Numerous structural variation algorithms using read pairs, and their fragment length, to detect variants have been proposed. Many tools use only *discordant* read pairs for downstream calling of variants, i.e., read pairs that align at a distance smaller than $\mu - k\sigma$ or larger than $\mu + k\sigma$ base pairs from each other, where μ and σ are the mean and standard deviation of the fragment length distribution and $k \in \mathbb{R}$ [2, 5, 11, 12, 20]. This restriction may reduce the computational demand, but it sacrifices sensitivity [17] by removing observations.

There are also tools with a statistical model/approach that utilizes all read pairs. CLEVER [17] finds insertions and deletions based on statistically significant deviation of the mean fragment length of all reads¹ over a position from μ . This method finds more and smaller variants compared to methods that use only discordant reads [17]. [9] models the number of discordant and concordant read pairs (classified by a cutoff) over a region as following a binomial distribution and finds inversions and deletions based on statistically significant accumulation of discordant read pairs over regions. However, any binary classification cutoff causes loss of information [8], thus statistical power, as they do not consider how much above or below the cutoff a fragment length is². Another approach is non-parametric testing of the distribution over a region, *e.g.*, using the Kolmogorov-Smirnov test [15], but as [17] noted, this is computationally expensive. [10] presented a model to find the most likely common deletion length from several donor genomes with different fragment length distributions by maximizing the likelihood of observed fragment lengths given a deletion size and each of the distributions.

¹ With some modifications to account for heterozygous variants. Only reads that have enough overlap and similar fragment lengths are grouped together.

² Under a normal distribution, 100 continuous observations are statistically equivalent to 158 binary observations for the best possible “cut point”, which is the mean. The loss of information becomes worse the further away the cut point is from the mean, *e.g.*, $\mu \pm k\sigma$, as k increases. In practice $k \in [3, 6]$ in variant detection tools.

These methods however assumes that the probability of a fragment length being observed over a position/region follows the probability distribution of the full library fragment length distribution, which is not true [22]. Longer fragment lengths span more positions than shorter fragment lengths, so over any position in the genome there will be a bias towards read-pairs further apart than μ . This observation bias of fragment sizes has been investigated earlier in an assembly context, estimating the gap size between contigs [18, 21, 22]. The approaches given in [18, 21] are more general by using the exact (empirical) distribution over the fragment length, which also makes them computationally demanding. GapEst [22] assumes a normal fragment length distribution and derives an analytic expression for the likelihood of a gap size that scales very well, which opens up for other applications where this type of problem needs to be calculated for a large number of instances, *e.g.*, structural variation detection. There is no previous work known to the authors on incorporating this model, or a similar one, to structural variation and investigating how it affects the balance between detecting deletions and insertions.

1.1 Contribution

We use the statistical model given in [22] and present it in the context of structural variation detection. The model provides a probability distribution for the fragment sizes we observe over a position (*e.g.*, a potential *breakpoint*) or region. Given this distribution we derive a null-hypothesis distribution to detect variants. We show that the corrected null-hypothesis agrees with both simulated and biological data, while a commonly used null-hypothesis does not. We implement the null-hypothesis in the state-of-the-art fragment-length based variant caller CLEVER [17]. Although CLEVER uses constraints and assumptions that do not agree with our model, we show that the detection of insertions and deletions becomes more balanced and that the number of false positive calls decreases. This is a promising first result as we could only apply a part of our theory in CLEVER without a significant restructuring of the code. We also believe that this work is a step towards creating a statistical rigorous approach for read pair fragment lengths where we can detect indels to a much higher resolution than cutoff based ones.

In some places, we have to refer to an Appendix, which can be found in a bioRxiv (<http://dx.doi.org/10.1101/036707>) version of this paper.

2 Methods

We will review a model used to determine contig distances in scaffolding [22] and use it in the context of structural variation detection. Notation and assumptions are presented in Sect. 2.1. In Sect. 2.2 we present the probability distribution in a structural variation detection context. Section 2.4 discuss a commonly used null-hypothesis used for detecting variants with fragment length and derives an improved null-hypothesis using our model.

2.1 Notation and Assumption

We refer to our model as the *Observed Fragment Length* (OFL) model. This model carries no new concepts and makes the same assumptions as the Lander-Waterman model [14], but adds a variable and some constants. We only state it here for convenience of referencing to a model when we derive probabilities and a null-hypothesis. Read pairs are sampled independently and uniformly from the donor genome. Let G denote the length of the reference genome. Alignment of read pairs to the reference genome yields our observations: distance o between reads in a read pair, read length r , and number of allowed “inner”³ softclipped bases s [16] in an alignment, see Fig. 1a. Read-pair distances x come from a library fragment length distribution $f(x)$ (either given or estimated from alignments). We denote the mean and standard deviation of this distribution as μ and σ . Finally, a parameter δ models the number of missing or added base pairs in the reference, compared to the donor sequence. That is, if the donor sequence contains an insertion, δ is negative and we say that the donor sequence has δ added bases. Similarly, if the donor sequence contains a deletion, δ is positive and we say that the donor sequence has δ deleted bases. For a given read-pair with fragment length x , let $w_{G,p}(x)$ denote the probability that it spans over position p on a genome of size G . As we do not model that any two positions have different probabilities to be spanned over (reads are drawn uniformly), w will not depend on p and we omit it and refer to $w_G(x)$ from now on.

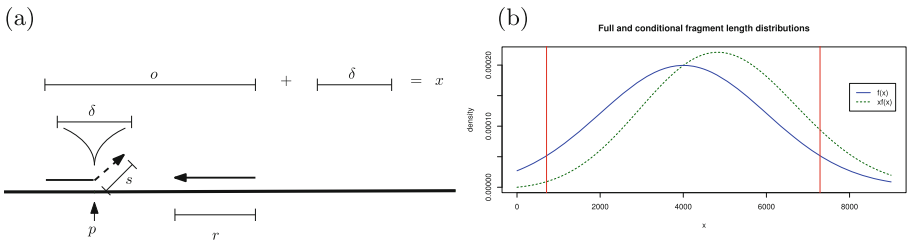


Fig. 1. (a) Constants and variables in the OFL model. The figure illustrates the scenario of an insertion in the donor genome of length δ at position p . Two reads (marked by arrows) of length r are at distance o from each other, with the left read partially aligned leaving s positions unaligned (softclipped). (b) Illustration of a full fragment size distribution $f(x)$ from $N(4000, 2000)$ (blue line), from which H_0 is derived. The green dashed line shows the observed fragment distribution over any variant free position for fragments coming from $f(x)$ for the simplified case when $r, s, \delta = 0$ (i.e., this is exactly the function $x f(x)$). Red lines indicate the $\mu \pm 2\sigma$ quantiles of $f(x)$. It is less likely to observe a smaller fragment size over any given position in the genome (see density of green distribution at red lines), as opposed to identical significance under $f(x)$.

³ We call the side of the read that is closest to its mate “inner”.

2.2 Probability Function over Observed Fragment Lengths

The distribution and probabilities derived in this section closely matches those given in [22] with the minor addition of the constant s . We restate the expressions in a structural variation context for clarity.

No Variant. First, we assume that donor and reference are identical, therefore $\delta = 0$ at any position. Given the OFL model, the probability that we observe a read pair with fragment size x over a position p on a genome of length G is

$$P(x|\delta = 0) = \frac{w_G(x)f(x)}{\sum_x w_G(x)f(x)} = \frac{\frac{(x-2(r-s))}{G} f(x)}{\sum_x \frac{(x-2(r-s))}{G} f(x)}. \quad (1)$$

Here $f(x)$ is the probability to draw a fragment of length x from the full library and $w_G(x)$ the probability that it spans over position p . The denominator is a normalization constant to make P a probability. It is assumed that $x \geq 2(r-s)$. For example, if the read length is 100 and maximum allowed softclipped bases of an aligned read is 30 a read pair with fragment length 300 will have $300 - 2(100 - 70) = 160$ possible placements where it spans position p . For simplicity, we omit the special case when p is near the end of a chromosome.

Modeling Variant at a Position. Let δ be the unknown variant size. In this case we cannot observe the true fragment length x of read pairs. What we observe is instead $o = x - \delta$ (see Fig. 1a). A modification of $w(x)$ is needed as fragment sizes is now required to span δ base pairs and have sufficiently many base pairs on each side to be mapped ($2(r-s)$). We have

$$w(x, \delta) = \frac{1}{G} \max\{x - \delta - 2(r-s) + 1, 0\}.$$

The 0 in the max function keeps the function weight to 0 in case we have no possible placing of a paired read over a variant. We can simplify this function to be expressed in o , as $o = x - \delta$, and write $w(o) = G^{-1} \max\{o - 2(r-s) + 1, 0\}$.

We see that the function w is constant for any given observation and can therefore be interpreted as a “weight” function, hence the notation w .

2.3 Probability of Variant Size δ

We can express the probability of δ given observations as $P(\delta|o)$. Lacking prior information about δ , we model it with the uniform distribution⁴. Using Bayes theorem, we get

$$P(\delta|o) = \frac{P(o|\delta)P(\delta)}{P(o)} \propto P(o|\delta)P(\delta) \propto P(o|\delta)$$

⁴ A more informative prior could improve results, *e.g.*, by fitting to the expected frequency and length of variants, studied in [4,6]. By tailoring the prior we could essentially obtain any specificity and sensitivity for a given indel size. We believe that is promising future work.

where $P(o)$ and $P(\delta)$ are constant by the assumption of a uniform distribution. We now have

$$P(o|\delta) = \frac{w(o|\delta)f(\delta + o|\delta)}{\sum_t w(t - \delta)f(t)} \tag{2}$$

where the denominator is the sum of all possible fragment sizes that can be observed given δ and f . We can now find the most likely δ using maximum likelihood estimation (MLE) over (2). The time complexity for the MLE is $O(n + \log t)$ ⁵ if $f \sim N$ (with t continuous), where n is the number of observations [22]. Note that we implicitly get $P(x|\delta)$ since $P(x|\delta) = P(o + \delta|\delta) = P(o|\delta)$.

2.4 Null-Hypothesis and Statistical Testing

Let $Y \doteq O|\delta = 0$, that is, the random variable over observed fragment lengths given $\delta = 0$. Let $\bar{y} \doteq \frac{\sum_{i=1}^n y_i}{n}$ be the sample mean of observed fragment lengths. Considering \bar{y} a random variable over experiments, it is commonly assumed that $\bar{y} \sim N(\mu, \sigma/\sqrt{n})$, *i.e.*, the distribution of the sample mean of $f(x)$ under the central limit theorem (CLT), and this distribution is used as null-hypothesis [5, 17]. We call this null-hypothesis H_0 . Furthermore, the variant size δ is estimated from observed fragment lengths o as $\hat{\delta} = \mu - \bar{o}$ [5, 9, 10, 12, 17, 20]. At first glance this formula seems reasonable since we take the expected fragment size and subtract the mean of the observations, but it has strong limitations. One is that $\hat{\delta}$ in this case has an upper bound of $\mu - 2(r - s)$ since $o \geq 2(r - s)$. This equation implies that we can never span over a sequence longer than $\mu - 2(r - s)$. We use Eq. 2 to derive the correct mean and standard deviation of Y given the OFL model, denoted μ_p and σ_p respectively. The derivation of μ_p is similar to derivation of observed fragment size linking two contigs given in [22], and the derivation of σ_p is a special case of the derivation of the variance of observed fragment size linking two contigs given in [23]. See proof in Appendix⁶.

Theorem 1. *Given the OFL model, $f \sim N(\mu, \sigma)$, and $\delta = 0$, we have $\mu_p \approx \mu + \frac{\sigma^2}{\mu - (2(r-s)+1)}$ and $\sigma_p \approx \sigma\sqrt{1 - \sigma^2(\mu - (2(r-s)+1))^{-2}}$.*

The null-hypothesis is that there is no variant, thus $\delta = 0$. Under CLT, as n increases, we therefore have $\bar{y} \sim N(\mu_p, \sigma_p/\sqrt{n})$. Notice that we can calculate μ_p and σ_p without the assumption $f \sim N(\mu, \sigma)$ by using an empirical estimate of $f(x)$ from aligned read pairs. Nevertheless, the closed expression formulas in Theorem 1 illustrates a basic feature of the model — larger variance increases the discrepancy between μ and μ_p . It is also robust to non-normality, as we will see in Sect. 3.2. In case we have enough observations to motivate the Z -test, we perform a simple Z -test and obtain a p -value based on a two sided test (both deletions and insertions are tested for) using the z -statistic

$$z = \frac{\bar{y} - \mu_p}{\sigma_p/\sqrt{n}}. \tag{3}$$

⁵ n to obtain sample mean \bar{o} , and $\log t$ to search the convex ML curve.

⁶ <http://dx.doi.org/10.1101/023929>, Sect. 5.5.

We refer to the null-hypothesis test using (3) as H'_0 . Thus, we have derived a different distribution under the null-hypothesis which we advocate should be used instead of H_0 . In case we have few observations (more often over insertions), approximation with the Z -test is poor. To get an exact test we would need to derive the distribution of $\sum_{i=1}^n Y_i$, for n observations y_i $i \in [1, n]$. This could improve power to detect insertions, but we refrain from studying this in the present paper.

3 Results

We discuss why modeling bias contributes to making deletion calls more frequent than insertions calls in Sect. 3.1. In Sect. 3.2 we show that our corrected hypothesis agrees with biological data, and in Sect. 3.3 that how indel detection is affected in CLEVER when our null-hypothesis is inserted.

3.1 Bias Between Detection of Deletions and Insertions

As donor fragments need to span over insertions ($\delta > 0$), and this probability is $w(x, \delta) = \frac{1}{G} \max\{x - \delta - 2(r - s) + 1, 0\}$ according to the OFL model, it is less likely that such fragments will be observed, as δ grows. We will therefore have a lower sample size over insertions in general. This naturally gives less power to detect an insertion compared to a deletion of similar size. However, methods using H_0 have less power than necessary. Firstly, as $\mu_p > \mu$, this gives too many significant upper quantile p -values (deletions) and too few significant lower quantile p -values (insertions). The difference in significance of observing a fragment of size $\mu + 2\sigma$ compared to observing a fragment of size $\mu - 2\sigma$ under H_0 , compared to the when observed under H'_0 is seen in Fig. 1b. Secondly, the positive skew of the OFL distribution (Fig. 1b) makes a Z -test approximation less powerful compared to an exact test, especially for small sample sizes — as is more likely for insertions.

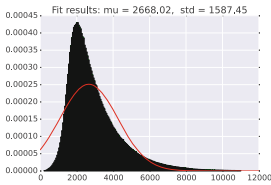
3.2 Evaluating the Accuracy of H'_0

We evaluated the accuracy of our null-hypothesis on three mate pair libraries. We used a mate pair library from *Rhodobacter sphaeroides* from [21] denoted *rhodo*, a mate pair library from *Plasmodium falciparum* used in [13] denoted *plasm*, and mate-pair data from a human individual in the CEPH 1463 family-trio⁷. For the human dataset we aligned the reads to the complete human genome, but limited analysis to chromosome 13. We call this dataset *hs13*. Table 1 shows information about the datasets. Recall from Sect. 2.4 that μ_p and σ_p are the true mean and standard deviation of fragment lengths over a position that does not contain a variant. Let $\hat{\mu}_p^c$ and $\hat{\sigma}_p^c$ refer to the estimated quantity of μ_p and σ_p from the closed formulas in Theorem 1. Similarly, let $\hat{\mu}_p^e$ and $\hat{\sigma}_p^e$ be the estimates of

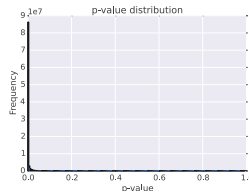
⁷ <http://www.ebi.ac.uk/ena/data/view/ERR262996>.

Table 1. Library information. Reads were aligned with BWA-MEM [16] version 0.7.12 with default parameters. Physical coverage is c , for all reads, and c (pp), for restricted proper pairs, i.e., read pairs that have both mates mapped in correct orientation and within a distance that depends on a statistical filtering of outliers based on the library distribution. The filtering bounds were roughly 10000, 6000, and 14000 bp for rhodo, plasm, and hs13 respectively. μ and σ are the mean and standard deviation of the full fragment length distribution. True mean insert-size and standard deviation over a position on the genome, μ_p and σ_p (calculated as the average over all positions in the genome) and predictions with closed formula, $\hat{\mu}_p^e$ and $\hat{\sigma}_p^e$, and exact calculation, $\hat{\mu}_p^c$ and $\hat{\sigma}_p^c$.

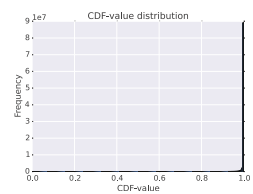
Organism	r	c	c (pp)	μ	σ	μ_p	σ_p	$\hat{\mu}_p^e$	$\hat{\sigma}_p^e$	$\hat{\mu}_p^c$	$\hat{\sigma}_p^c$
rhodo	101	43	34.5	2640	1390	3480	1534	3446	1526	3434	1143
plasm	75	4.9	4.2	2955	524	3056	511	3056	517	3053	515
hs13	80	11.1	9.0	2947	1454	3688	1780	3719	1806	3705	1241



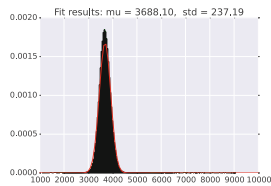
(a) $f(x)$



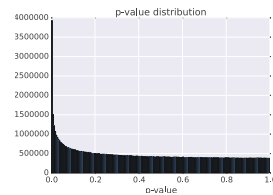
(b) p -values of H_0 based on μ, σ (naive) in equation 3.



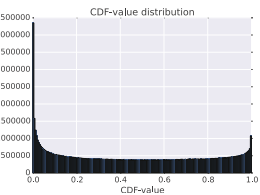
(c) CDF-values $P(X < x)$ of $f(x)$



(d) Distribution of $\mu_p, \forall p \in [1, G]$.



(e) p -values of H_0' based on $\hat{\mu}_p^e, \hat{\sigma}_p^e$ in equation 3



(f) CDF-values $P(X < x)$ of $w(x)f(x)$

Fig. 2. (a) The fragment length distribution $f(x)$ for the hs13 dataset and the red line is a best fit of a truncated normal distribution. $f(x)$ deviates significantly from a normal distribution. Although the mean of $f(x)$ is 2947, the average observed fragment length over position p (μ_p) over all positions on hs13 shows that most values occur between 3000–4500 bp with the average around 3688 bp (d) — as approximately predicted from $\hat{\mu}_p^e$ and $\hat{\mu}_p^c$. Figures (b) and (c) shows the p -value distribution and CDF values from using H_0 (i.e., using μ and σ). Figures (e) and (f) shows the p -value distribution and CDF values from using H_0' (i.e., using $\hat{\mu}_p^e$ and $\hat{\sigma}_p^e$).

μ_p and σ_p by using an empirical distribution of $f(x)$ (estimated from a sample) and summing up the probabilities in Eq. 2 with $\delta = 0$. Estimates and observed values are shown in Table 1. It is our assumption that an overwhelming majority of positions are variant free⁸. Thus, we expect a model that fits data should give a uniformly distributed p -value distribution. Our observations are summarized below.

Predicting μ_p . μ_p is estimated very well by both $\hat{\mu}_p^e$ and $\hat{\mu}_p^c$; compare Fig. 2d for hs13, and Appendix Figs. 3b and d for rhodo and plasm respectively with the estimated values in Table 1. Hence, testing $\bar{o} = \hat{\mu}_p^e$ (or $\hat{\mu}_p^c$) as in H'_0 introduce symmetrical cumulative distribution function (CDF) values, Fig. 2f, compared to CDF based on testing $\bar{o} = \hat{\mu}$ where all values are distributed around 1.0 — suggesting significant deletions, see Fig. 2c.

Predicting σ_p . The closed formula predictions of σ_p works best if $f(x)$ is normal (plasm). For rhodo and hs13, $\hat{\sigma}_p^e$ and $\hat{\sigma}_p^c$ differs significantly and $\hat{\sigma}_p^e$ should be used, compare σ_p with $\hat{\sigma}_p^e$ and $\hat{\sigma}_p^c$ in Table 1.

p -values. The p -value distribution (ideally uniform) greatly improves with H'_0 (Fig. 2e) compared to p -values obtained with H_0 (Fig. 2b). Abnormalities in the p -values are most likely explained by: alignment artifacts (some regions are more difficult aligning to), fragment length bias [1, 19], coverage bias from GC-content, and in some cases, real variants, see evaluation of hs13 dataset in Sect. 5.7 of the Appendix. Similar p -value distributions are obtained on rhodo and plasm genome (data not shown) — that should not contain any variants — indicating that most of the enrichment of low p -values on hs13 is explained by any of the former three causes.

Table 2. Insertions and deletions called with CLEVER using H_0 and H'_0 . Column δ contains the size of 50 insertions and deletions, simulated on the reference genomes by either deleting or inserting a δ bp sequence on the reference. A “0” indicates that the original biological dataset was used.

Dataset	δ	H_0				H'_0			
		TP (del/ins)		FP (del/ins)		TP (del/ins)		FP (del/ins)	
plasm	0	0	(0/0)	20	(6/14)	0	(0/0)	27	(6/21)
	2000	89	(50/39)	22	(8/22)	89	(50/39)	38	(8/30)
rhodo	0	0	(0/0)	78	(78/0)	0	(0/0)	18	(14/4)
	2000	49	(49/0)	54	(54/0)	57	(45/12)	13	(9/4)
sim-N(500,75)	75	94	(94/0)	0	(0/0)	78	(62/16)	0	(0/0)
	100	110	(100/10)	0	(0/0)	188	(100/88)	0	(0/0)
		ETP (hits)		TC (del/ins)		ETP (hits)		TC (del/ins)	
hs13	0	9 (31)		1740	(1740/0)	3 (4)		8	(8/0)

⁸ Even small variants $\delta \ll \sigma$ will not affect the model much.

3.3 Implementing the Corrected Null-Hypothesis in CLEVER

In this section we illustrate as a proof-of-concept how the corrected hypothesis H'_0 (with $\hat{\mu}_p^e$ and $\hat{\sigma}_p^e$) balances the ratio between detected insertions and deletions. We applied H'_0 in CLEVER (v 1.1). *However, we want to emphasize that we did not tailor the statistical tests as needed to fit the assumptions made by their particular method. This limits the performance improvement.* To further improve results with CLEVER, we would need to (1) implement exact tests for few observations — giving more power to detect insertions, (2) use the OFL-model for CLEVER’s discovery of positions to study, (3) based on our model adjust CLEVER’s methods to handle, *e.g.*, heterozygous variants and controlling the false discovery rate. This would require additional modeling and significant restructuring of the code and we do not consider it here. Our aim here is only to illustrate how the simple adjustment of inserting H'_0 instead of H_0 in CLEVER has a significant impact on the output. We investigated how the replacement of H'_0 instead of H_0 changed variant calls from CLEVER on hs13, rhodo and plasm as well with ideal condition simulated data denoted $sim-N(\cdot, \cdot)$ (full simulated results in Appendix Sect. 5.6). For simulated variants, similarly to [17], a prediction is classified as a true positive (TP) if the breakpoint prediction is not further than one mean insert size (*i.e.*, at most $\mu - 2r$) away from the true breakpoint. Otherwise it is classified as a false positive (FP). All variant calls on rhodo and plasm that are not from simulated variants are assumed to be false positives.

Because hs13 likely harbors true variants, we used annotated variants from dbVar [7], together with manual inspection in BamView [3], to assess if *hits* are true or false positives. For a deletion call in CLEVER with start and end coordinates p_s, p_e and a deletion in dbVar with coordinates q_s, q_e , we let $max_del = \max(p_e - p_s, q_e - q_s)$ and $overlap = \min\{0, \min(p_e, q_e) - \max(p_s, q_s)\}$. We let $hit_value = overlap/max_del$ and a call is a *hit* if $hit_value > T$, where $0 < T < 1$ is a threshold. Because dbVar contains a large amount of annotated variants from several individuals and CLEVER produces many calls under H_0 , roughly 173, 106 and 40 hits are expected by chance with $T = 0.25, 0.5, 0.75$ (estimation in Appendix Sect. 5.3), which is similar numbers to the observed hits from CLEVER: 226, 109 and 31 respectively under $T = 0.25, 0.5, 0.75$. We therefore further manually evaluated the hits produced with $T = 0.75$ by looking for coverage drop and accumulation of softclips near each breakpoint. This gave us Estimated True Positives (ETP) as a rough measure of the TP rate for hs13. Therefore, we report ETP and Total Calls (TC) for hs13 in Table 2, contrary to simple TP and FP for the other data sets where we have the ground truth.

Improvements. From Table 2 and Fig. 5 (Appendix) we see that CLEVER with H_0 detects significantly more deletions than insertions of the same sizes. Using H'_0 , reduces this bias to some extent by increasing the detection of insertions across all data sets. CLEVER also returns significantly fewer false positive deletion calls with H'_0 , see rhodo Table 2 and $sim-N(300, \cdot)$. Even though H_0 have more sensitivity in calling deletions on hs13, the signal disappears in the

overwhelming amount of total calls, compare ETP and TC for H_0 and H'_0 in Table 2.

Deterioration. A consequence of using H'_0 is fewer deletion calls, which unfortunately also removes some true positive deletion calls (Table 2 and Appendix Fig. 5). It also increases the FP insertion calls on plasm (Table 2). We believe that calling variants with the plasm library carries additional difficulties due to its GC-poor genome sequence, such as positional fragment length bias [1, 19].

Additional evidence that most calls with H_0 on hs13 are FPs are found by comparing statistics on CLEVER's deletion calls (Appendix Fig. 6) and numbers reported in recent extensive studies [4, 6]. For example, [4] provide frequency distributions for both previously discovered and new deletions on single genomes. Roughly 250 deletions have lengths over 1000 bp (inspection of plot). The simplifying assumption that large-indel distribution is uniform over chromosomes gives around 8 expected deletions⁹ in size range 1000 bp. This approximate number, and the fact that almost all calls were removed when using H'_0 corroborates, that the vast majority (in the order of > 99%) of calls with H_0 are FPs — likely a consequence of using $H_0 : \mu_p = 2410$ compared to the true value $\mu_p = 3719$.

4 Conclusions

We stated a probability distribution of observed fragment length over a position or region and derived a new null-hypothesis for detecting variants with fragment length, which is sound and agrees with biological data. Applied in CLEVER, our null-hypothesis detects more insertions and reduces false positive deletion calls. Results could be further improved by deriving an exact distribution instead of a Z -test and updating CLEVER's edge-creating conditions to agree with our model. The presented model, distribution, and null-hypothesis are general and could be used together with other information sources such as split reads, softclipped alignments, and read-depth information.

References

1. Benjamini, Y., Speed, T.P.: Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**(10), e72–e72 (2012)
2. Bickhart, D., Hutchison, J., Xu, L., Schnabel, R., Taylor, J., Reecy, J., Schroeder, S., Van Tassell, C., Sonstegard, T., Liu, G.: RAPTR-SV: a hybrid method for the detection of structural variants. *Bioinformatics* **31**(13), 2084–2090 (2015)
3. Carver, T., Böhme, U., Otto, T.D., Parkhill, J., Berriman, M.: BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* **26**(5), 676–677 (2010)

⁹ Estimated as $250 \cdot \frac{(114-16)\text{Mbp}}{3 \text{ Gbp}} = 8$, including compensation for the 16M N's at the start of the reference sequence for chr 13.

4. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J.M., Stamatoyannopoulos, J.A., Hunkapiller, M.W., Korlach, J., Eichler, E.E.: Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**(7536), 608–611 (2015)
5. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S., Wendl, M., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T., Wilson, R., Ding, L., Mardis, E.R.: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Meth* **6**(9), 677–681 (2009)
6. Chen, W., Zhang, L.: The pattern of DNA cleavage intensity around indels. *Sci. Rep.* **5**, 8333 (2015)
7. Church, D.M., Lappalainen, I., Sneddon, T.P., Hinton, J., Maquire, M., Lopez, J., Garner, J., Paschall, J., DiCuccio, M., Yaschenko, E., Scherer, S.W., Feuk, L., Flicek, P.: Public data archives for genomic structural variation. *Nat. Genet.* **42**(10), 813–814 (2010)
8. Fedorov, V., Mannino, F., Zhang, R.: Consequences of dichotomization. *Pharm. Stat.* **8**(1), 50–61 (2009)
9. Gillet-Markowska, A., Richard, H., Fischer, G., Lafontaine, I.: Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics* **31**(6), 801–808 (2015)
10. Handsaker, R.E., Korn, J.M., Nemes, J., McCarroll, S.A.: Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**(3), 269–276 (2011)
11. Hayes, M., Pyon, Y.S., Li, J.: A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data. *PLoS ONE* **7**(12), e52881 (2012)
12. Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C.: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**(7), 1270–1278 (2009)
13. Hunt, M., Newbold, C., Berriman, M., Otto, T.: A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.* **15**(3), R42 (2014)
14. Lander, E.S., Waterman, M.S.: Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**(3), 231–239 (1988)
15. Lee, S., Hormozdiari, F., Alkan, C., Brudno, M.: MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* **6**(7), 473–474 (2009)
16. Li, H., Durbin, R.: Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5), 589–595 (2010)
17. Marschall, T., Costa, I., Canzar, S., Bauer, M., Klau, G., Schliep, A., Schönhuth, A.: CLEVER: Clique-enumerating variant finder. *Bioinformatics* **28**(22), 2875–2880 (2012)
18. Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A., Prjibelski, A.D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S.R., Woyke, T., Mclean, J.S., Lasken, R., Tesler, G., Alekseyev, M.A., Pevzner, P.A.: Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**(10), 714–737 (2013)
19. Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.V., Oparina, N.Y., Polozov, R.V., Nechipurenko, Y.D., Grokhovsky, S.L.: Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* **4**, 4532 (2014)

20. Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., Hall, I.M.: Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* **20**(5), 623–635 (2010)
21. Ribeiro, F.J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J., Young, S.K., Russ, C., Nusbaum, C., MacCallum, I., Jaffe, D.B.: Finished bacterial genomes from shotgun sequence data. *Genome Res.* **22**(11), 2270–2277 (2012)
22. Sahlin, K., Street, N., Lundeberg, J., Arvestad, L.: Improved gap size estimation for scaffolding algorithms. *Bioinformatics* **28**(17), 2215–2222 (2012)
23. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., Arvestad, L.: BESST – efficient scaffolding of large fragmented assemblies. *BMC Bioinf.* **15**(1), 281 (2014)