Approaches to Structural Variant Detection with Short Next-Generation Sequencing Data

George Horrell

June 2019

1 Abstract

1.1 Purpose

Chromosomal structural variations (SVs) occur within the genome and can have a decisive bearing on phenotype. While most early efforts to characterize genetic diversity focused on single-nucleotide polymorphisms (SNPs) and shorter indels, with the advent of next-generation sequencing techniques, attention has been given to these longer SVs. While long read sequencing (reads >1000bp in length) has proven instructive in the detection of these SVs, short-read sequencing is still the dominant sequencing technology due to price and availability of sequencing machines. Furthermore, there are some sequencing applications where only short reads can be attained. Using a golden-set of structural variants attained from calling upon long reads, structural variant calling for short reads can be tuned and improved.

1.2 Methods

We evaluated three main approaches for calling structural variants on short reads. The first approach was to improve upon split read calling, the defacto approach for calling short reads. The second approach was to focus on the alignment phase, and appraise how mapping the reads could effect the efficacy of later SV calling. The final approach trialled was to first use denovo assembly to combine reads into contigs which provide more context to the aligner. We evaluated these approaches using a golden-set provided by NIST

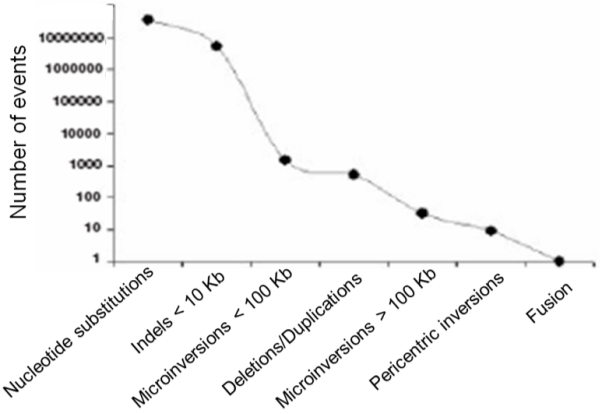

Figure 1: Frequency of different mutations

on the publicly accessible HG002 Genome in a Bottle. The tool Truvari was used to score the output of the structural variant callers.

1.3 Conclusion

Further research needs to be carried out into alignment parameters to maximize accurate detection of structural variants. Creating a split-read caller that challenges existing industrial or academic tools proved to be beyond the scope of this project. Denovo assembly appears to be a promising approach; already it outperforms existing standards for long insertions.

2 Introduction

2.1 Structural variants

Structural variations (SVs) include insertions, deletions, duplications, inversions, and translocations of DNA segments longer than 50bp in length. Since these structural variants can span single exons or larger genomic regions, they bear significant impact on phenotype. Due to their large size, they account for more of the nucleotide sequence variation than SNPs (1% for SVs and 0.1% for SNPs); indels $<$ 10kb in length are the second most common mutation (see figure 1).

We detect these structural variants at the end of a data pipeline that

starts with sequencing (converting the biological sequence into a sequence of bases known as a read), proceeds to alignment/assembly (comparing the reads to a "reference genome" and deciding where they belong in the DNA sequence) and terminates with calling SVs, based on the percieved difference between the subject genome and the reference genome.

However, SVs have received less attention since they are harder to detect with traditional sequencing technologies. Initial commercial reads were shorter than 50bp; this substantially limits the ability of algorithms to call SVs. It is impossible to call a repeat, such as a copy number variation, that is longer than the read length for the sequencing run and most SVs occur in highly repetitive regions of the genome [9].

Today paired-end reads from 100bp-300bp are employed commercially [5]. This increase in sequencing length opens the possibility of detecting shorter SVs (<300bp in length) using cheaper commercial reads. However, even at 200bp there is still the challenge of aligning reads that span variants, since there is little context to place the matching regions of the read to the reference genome.

2.2 Next generation sequencing

The advent of long-read next-generation sequencing (NGS), offered by companies like Oxford Nanopore and PacBio, substantially improved resolution of long structural variants. These long-reads are sometimes 1000s of base pairs in length and with this additional context, it is easier to align them to the correct region of the reference genome. SV callers which operate over long reads have proven highly effective at calling structural variants on synthetic data [6].

2.3 Golden-set

Unlike with synthetic data, it is impossible to know with certainty whether the calls made on real human genetic data are accurate. To combat this, publicly accessible test sets, such as the NIST Genome in a Bottle, are used as the bases for reproducible tests for structural variants using high coverage sequencing data [11]. By combining high confidence SV calls from different sequencing technologies and SV calling algorithms, a golden-set of SVs is produced. This golden-set can be used as a reference point for the output of new callers.

2.4 Relevance of short reads

Long reads provide higher resolution sequencing data, especially error corrected long reads [7]. However, these long reads are prohibitively expensive for many commercial applications, and cannot be used for all sequencing applications. In oncology, SVs are thought to contribute to oncogenesis through a range of mechanisms, and are thought to be underappreciated mutational drivers [2]. Liquid biopsies have been shown to be effective diagnostics tests for cancer and require sequencing cfDNA circulating in the blood stream [3]. cfDNA was found to have an average fragment length in humans of 144bp [10], which implies that only short reads and short read SV callers will be effective in this situation.

2.5 Split-read callers

SV callers for long and short reads have focused primarily on split reads, read depth and read pairs to provide signal about structural variation. Split reads occur when reads are aligned partially to one region and partially to another. Their occurrence can suggest the presence of any kind of structural variant. Combining information from multiple split reads can be used to make a high confidence call. Read depth refers to the number of reads aligned over the same genomic region. In repetitive regions, read depth can be used to ascertain the copy number of a repeat, which as discussed earlier, can be challenging with short reads. These two approaches underpin existing commercial SV callers, such as Manta [1].

Split-read callers for short reads are good at detecting deletions, but comparatively poor at detecting insertions. This is the case for any approach which relies purely on reference guided assembly as an upstream step to the caller, since any read which is spanned by the inserted region cannot be correctly aligned. Compared to long reads, existing short reads SV callers pick up a maximum of 49% of deletions and 11% of insertions, at 30-40x coverage (see figure 2). The majority of the missing variation appears to lie between 50-500bp, which perfectly overlaps with the average short read length used for sequencing.

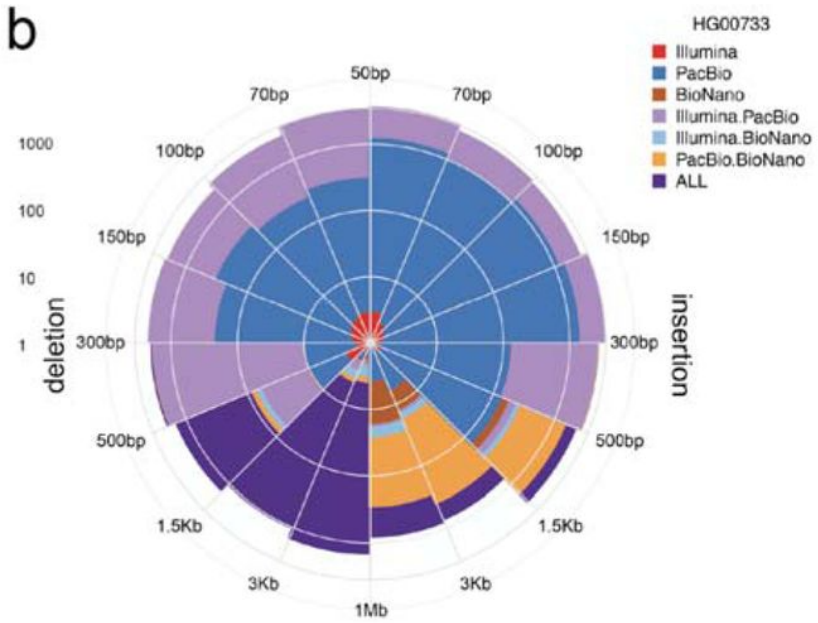

Figure 2: Recall of different sequencing technologies

2.6 Denovo assembly

While the majority of callers rely on the methods detailed above, some research has been carried out into alternative ways to call SVs. For instance, combining reference and denovo assembly has been used to phase structural variants. Denovo assembly has also been used to generate high quality contigs (long sequences formed by merging overlapping reads) which are then reference aligned. This approach does sacrifice some signal, squashing all read information which might total to over 30x coverage, down into contigs which might be 1-3x coverage. Nonetheless, these contigs with their extra length have high degrees of context which improves the quality of alignments and therefore the SV calls made [4].

3 Materials and methods

Three approaches were explored as candidate SV callers for short reads. While SVs include inversions, translocations and duplications, for the purpose of this paper only insertions and deletions were assessed.

3.1 NIST golden-set

To score the performance of the SV callers developed, we use a golden-set of calls assembled from a wide range of sequencing technologies and callers. We chose a golden-set that includes calls made on long reads and linked reads, to ensure that it is more comprehensive than the more limited calls possible with short reads. The golden-set used was the NIST SV golden-set prepared for the HG002 genome. The HG002 genome belongs to the Ashkenazim son, one of the genome donors for the Genome in a Bottle project. The callset contains 9719 structural variants in high confidence regions. Of these SVs, there are more deletions (56%) than insertions, and over 50% of the deletions and insertions are shorter than 200 bps.

Bounds	**Deletions**	**Insertions**
$50 \leq \text{SV} < 100$	1460	1501
$100 \leq \text{SV} < 500$	2034	2704
$500 \leq \text{SV} < 1000$	208	485
$1000 \leq \text{SV} < 10000$	523	729
$10000 \leq \text{SV}$	36	25
Total	**4261**	**5444**

Table 1: NIST golden-set by length

The majority of structural variants occur in repetitive regions, but this golden-set has undergone a large amount of filtering and non-random selection. We investigate whether the majority of our repeats have occurred in repetitive regions by masking our golden-set against a known set of repeats, computing an intersection of our golden-set against the simpleRepeats track from UCSC, with 100 bases of intersect padding.

The golden-set contains more deletions in repetitive regions than insertions. This is likely due to increased difficulty of detecting insertions being compounded by the added complexity of aligning reads that span repetitive regions.

3.2 Baseline: Novoalign and Manta

The NIST golden-set provides a testing interface, but we must select an existing caller to use as a baseline, to calibrate our methods against and score their success. We selected Manta, an industrial standard SV caller

Bounds	**Deletions**	**Insertions**
$50 \leq \text{SV} < 100$	688	294
$100 \leq \text{SV} < 500$	916	177
$500 \leq \text{SV} < 1000$	9	2
$1000 \leq \text{SV} < 10000$	101	0
$10000 \leq \text{SV}$	33	0
Total	**1747**	**473**

Table 2: NIST golden-set intersected with known repeats

built by Illumina, for this purpose. Manta is designed to operate optimally using the information provided by their HiSeq paired-end read sequencers. Manta takes aligned reads as input; for the baseline, we used Novoalign, a popular industrial aligner.

3.2.1 Manta

Manta has two main operating steps: candidate detection and filtering.

Candidate Detection In this first phase of operation, Manta seeks to find as many candidate SV sites as possible. It does this by looking at split-reads, paired-end reads, and alignment information over aligned reads. It then keeps a record of potential "break-end" sites; genomic coordinates where SVs start or end.

Filtering Manta's second phase ensures that only high confidence SVs are reported. It uses an assembly step, which coalesces reads from a potential break-end point into a longer contig which is then aligned with parameters sensitive to an existing structural variant. A number of probabilistic tests are used to score whether such a structural variant is likely, and if all filters are passed, then the SV is reported.

3.2.2 Performance

Manta is one of the callers used to form the NIST golden-set. Therefore, we expect the SVs called by Manta to have a high specificity against the golden-set, since the golden-set is a super-set of these Manta calls. Manta

performs well on the golden-set, yielding 60% recall on deletions, but only 23% recall on insertions. Precision is high for both.

	Deletions	Insertion	Total
Calls in golden-set	4261	5444	9705
Calls made by Manta	2689	1272	3961
Recall of Manta calls	**0.600**	**0.225**	**0.388**
Precision of Manta calls	**0.952**	**0.964**	**0.956**
F1 of Manta calls	0.736	0.365	0.522

Table 3: Manta callset

Length Manta demonstrates good recall on short and median length deletions (100-500). It has very low recall for insertions longer than 50bp.

Bounds	**Deletions**	**Insertions**
$50 \leq \text{SV} < 100$	0.61	0.499
$100 \leq \text{SV} < 500$	0.734	0.173
$500 \leq \text{SV} < 1000$	0.177	0.016
$1000 \leq \text{SV} < 10000$	0.198	0
$10000 \leq \text{SV}$	0.552	0
Total	**0.600**	**0.225**

Table 4: Manta recall on deletions and insertions by length

This is likely because of how difficult it is to align short reads around insertions. Insertions will necessarily lead to clipped reads; along with paired-end information, extracting break-ends from clipped reads is the only way to detect insertions with aligned short reads.

3.3 Approaches to improving short read SV calling

3.3.1 Darwin: Improvements to alignment

3.3.2 Split read caller

The first method used looked to improve upon existing split read callers. Darwin is configured to yield extra information about deletions. During the alignment phase it will merge split reads together, to make detection

of deletions via long alignment gaps easier. It also uses a dual-affine gap function, which should be more sensitive to long structural variants than a single affine gap.

By combining this information with the techniques mentioned above that Manta employs (break-end detection), we can call deletions based on short read alignment information. Our split read caller uses paired-end reads, split reads and an assembly phased inspired by Manta to refine the break-end points.

3.3.3 Denovo assembly for contigs

The primary issue with alignment of short reads for the later detection of structural variants is that the sequence that matches between read and reference is the issue of "vanishing score". The recurrence relation that underpins the Smith-Waterman algorithm relies on accruing large scores in matching regions which are then slowly depleted in regions containing edits. These matching regions are referred to as "context" for the regions containing the structural variant.

Short reads lack the context required to correctly align across structural variants. They either align partially with clipping or not at all. It is then left to the caller to attempt to combine the fragmented pieces of information that emerge from the alignment phase into evidence for structural variants. This process under-utilizes the power of the alignment phase, failing to capture all the information that can be gleaned from the relationship between the reads themselves.

One potential avenue for improvement here is to assemble contigs via unbiased denovo assembly using the short reads, and then align those longer assembled contigs to the reference genome. The advantage of this approach is that the generated contigs carry more context, and are easier to align to the reference genome (see figure 3).

Furthermore, the relationships that emerge from the assembly process, such as the DeBruijn graph used to connect overlapping reads can be used to guide the calling process. Furthermore, denovo assembly can utilize all the extraneous information that accompanies NGS reads; paired end information (distance between paired reads) including long range pairings, that can substantially improve the quality of the eventual assembly.

Any structure in the DeBruijn graph that resembles a "bubble", where two paths through the graph eventually reconnect, is suggestive of a het-

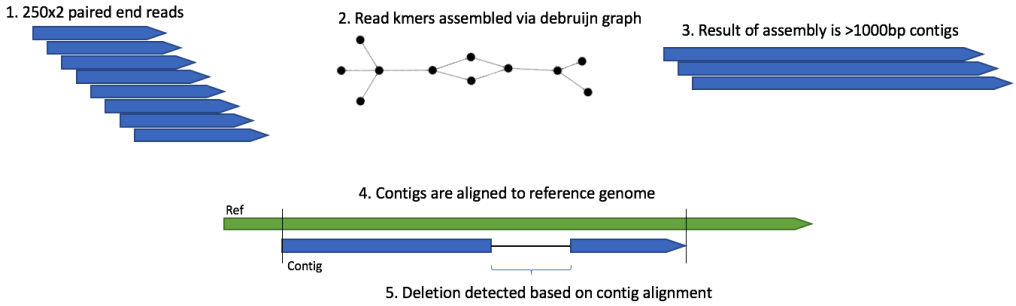

Figure 3: Alignment of contigs via denovo assembly

erozygous variant, where one allele carries the variation. This information can be exploited for both haplotyping and calling variants. If the Debruijn graph can be further extended via coloration, then handling noisier bubbles (where one allele might connect to another region of the graph) becomes possible. However, graph coloration adds some overhead to the assembly memory requirements and can make certain efficiencies impossible, such as the use of a bloom filter to reduce memory constraints [4] [8].

An extension implemented to the pipeline discussed above was to search the Debruijn graph for simple bubbles; diamond structures in the graph, where the difference in length between the two paths was over 50bp in length. Once these structures were found, "super-contigs" were generated that traced both paths through the diamond. These super-contigs were aligned, and the same SV calling algorithm was used to look for insertions and deletions.

4 Results

4.1 Darwin: Improvements to alignment

One vector for improving SV calling, is to improve the stages upstream of SV calling. Improving the quality of the alignments produced by the aligner used by an SV caller, and ensuring that the maximum amount of useful signal reaches the SV caller should improve the ultimate recall/precision of the caller.

A number of improvements were made to Darwin to attempt to improve alignment quality, including a dual-affine gap and merging of split-read pairs that suggested deletions. Some meta-data needs to be added to Darwin to

test it directly with Manta, but there is some evidence that given current parameters, that Darwin might under-perform Manta as a aligner prior to calling.

Clipped reads provide a useful metric for evaluating useful signal passed from aligner to SV caller. Correct alignment around insertions and deletions will produce clipped reads at a higher rate than alignments spanning indels/SNPs. Therefore, we expect that an aligner that is sensitive to structural variants would align a high fraction of clipped reads around known structural variants. Below we identify the percentage of clipped reads that overlap with structural variants.

	Novoalign	Darwin
Alignments	807167278	364342839
Clipped alignments	1612687	143730467
% alignments clipped	**0.2%**	**40%**
Clipped aligns. in HC regions	1514008	122464719
Clipped aligns. in HC overlapping SVs	13603	296913
% clip. aligns. overlapping SVs in HC	**0.8%**	**0.2%**

Table 5: Clipped reads overlapping SVs

Darwin produced 20x more clipped reads than Novoalign, but fewer of those clipped reads end up aligning around known SVs. This percentage only considers reads that have aligned in the high confidence region laid out by the golden-set. These results suggest that further tuning of the Darwin parameters might be required to align with sensitivity to SVs.

4.2 Split-read caller

Our improved split read caller performed poorly compared to Manta. The recall attained on deletions was 2x worse than Manta, and the precision was also roughly 2x worse than Manta. This was roughly the case at all SV lengths.

Given the attention that this approach has been given already, both in industry and academia, we decided that this approach was unlikely to yield novel results. Furthermore, given that Manta, the de-facto standard for split read callers, performed so poorly on insertions, we believed that an alternative approach might yield more impressive results.

	Deletions
Calls in golden-set	4261
Calls made by split-read caller	2490
Recall of split-read caller calls	0.205
Precision of split-read caller calls	0.426
F1 of split-read caller calls	**0.277**

Table 6: Split-read callset results

4.3 Aligned contigs via denovo assembly

4.3.1 Assembly quality

Before considering alignment and calling, a high quality denovo assembly had be generated in a reasonable time on lab hardware. As discussed above, Abyss was chosen for this; it uses a bloom filter (probabilistic hash set) to restrict the memory usage. The output of the assembly of the HG002 short reads was good, with an N50 of 29274 and an L50 of 27500. The total number of bases contained by the assembled contigs was 2.78e9. Given that the contigs are high confidence, this loss of coverage could be worth the trade-off for the high confidence context gained.

4.3.2 Calling

Denovo assembly is the most promising approach employed for structural variant detection. It performs poorly for short length structural variants but performs better than Manta for longer variants, both insertions and deletions.

	Deletions	Insertion	Total
Calls in golden-set	4261	5444	9705
Calls made by Manta	3718	2633	6351
Recall of Manta calls	**0.446**	**0.227**	**0.322**
Precision of Manta calls	**0.505**	**0.469**	**0.490**
F1 of Manta calls	0.474	0.306	0.389

Table 7: Denovo assembly callset results

Investigating these results for recall and precision at different lengths. It performs worse than Manta for short length structural variants but performs

better than Manta for longer variants, both insertions and deletions. The recall shown below trumps Manta for all insertions sizes over 100bp. The precision for these longer insertions was also high, at over 0.94 for all lengths over 100bp.

Bounds	**Deletions**	**Insertions**
$50 \leq \text{SV} < 100$	0.547	0.476
$100 \leq \text{SV} < 500$	0.561	0.187
$500 \leq \text{SV} < 1000$	0.398	0.206
$1000 \leq \text{SV} < 10000$	0.347	0.157
$10000 \leq \text{SV}$	0.034	0
Total	**0.519**	**0.264**

Table 8: Denovo assembly callset recall

As discussed above, diamond "super-contigs" were generated and then aligned for SV calling. Combining the results of these two steps together yields a final F1 score of 0.433, compared to 0.522 for Manta. While only 2000 of these super-contigs were found, the precision of the calls made on them was especially high for long insertions, over 0.9 for all lengths longer than 100bp.

Bounds	**Insertions**
$50 \leq \text{SV} < 100$	0.308
$100 \leq \text{SV} < 500$	0.944
$500 \leq \text{SV} < 1000$	1.000
$1000 \leq \text{SV} < 10000$	0.991
$10000 \leq \text{SV}$	0
Total	0.469

Table 9: Diamond supercontigs precision

Isolating the performance of our contig caller (CC) against Manta for insertions only, and removing all calls shorter than 1000bp, we see that this approach performs substantially better for this use case.

	Num calls		Recall		Precision		F1	
Bounds	CC	Manta	CC	Manta	CC	Manta	CC	Manta
$100 \leq SV < 500$	531	487	**0.187**	0.173	0.951	**0.963**	**0.312**	0.293
$500 \leq SV < 1000$	100	8	**0.206**	0.016	**1.000**	1.000	**0.341**	0.032
$1000 \leq SV < 10000$	117	0	**0.157**	0.000	**0.991**	0.000	**0.272**	0.000
$10000 \leq SV$	0	0	0.000	0.000	0.000	0.000	0.000	0.000

Table 10: Performance of contig caller against Manta for long insertions

5 Discussion

The most promising approach explored is contig alignment after denovo assembly. The high precision that accompanies the novel longer insertions generated via denovo assembly would demonstrates that the algorithm is more effective than the split read calling approach employed by Manta (for long insertions).

5.1 Super-contigs

Furthermore, the high specificity of the contigs generated via diamond alignment is encouraging, and there is evidence that with some tuning of alignment parameters, that more of the super-contigs generated could contribute towards calls. The majority of the longest super-contigs were not correctly aligned; for instance the 10kb deletion shown below:

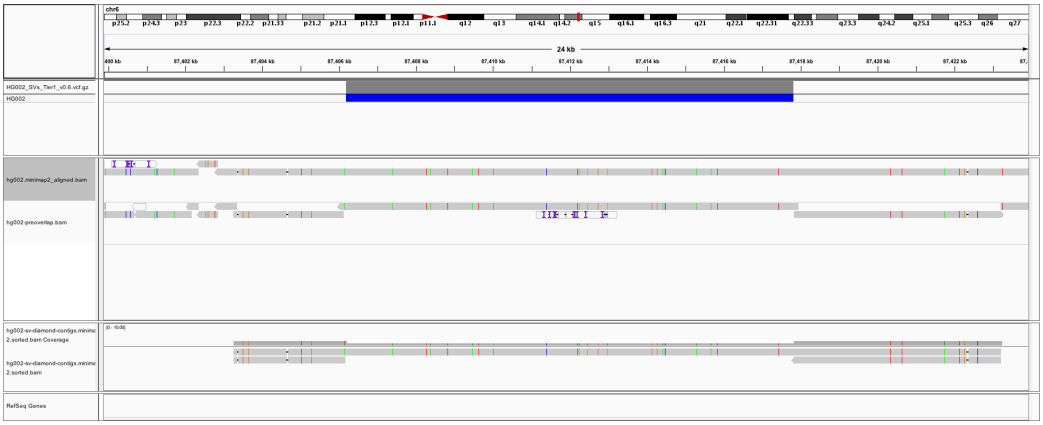

Figure 4: Misaligned 10kb deletion

In this instance, there was not enough context to align across the deletion, and the alignment was split. This would be an easy enough thing to fix up, gaining perhaps 40 or so very long deletions. There were over 140 super-contigs generated over 1000kb in length; from a random sample of twenty of these super-contigs, over 90% of them seem to correctly align with known structural variants. Combining this with the homozygous calls made during the original pipeline, and this denovo approach outperforms Manta for longer insertions and could possibly be used to augment the Manta insertions detected.

5.1.1 More control of assembly

As of now, the DeBruijn graph used to generate the super-contigs and the contigs themselves are plucked from various stages of the Abyss denovo assembly process. The Abyss assembler is nicely decomposed with good documentation; it might be possible to take more fine grained control of the assembly process and optimize for detecting structural variations.

5.1.2 Exploiting the DeBruijn graph

There is more information encoded in the DeBruijn graph than is currently used by this approach. While the detection of diamonds is quick and high confidence, finding and resolving more complex bubbles should yield more calls. As discussed by Iqbal in the colored Debruijn graph paper [4], it is possible to use the reference genome to resolve irregularities in the bubbles and make high quality calls.

5.2 Augmentation of Manta

This contig alignment approach appears to perform disproportionately well for longer structural variants. The high precision of these long calls is likely due to the filtering step of alignment; only contigs with context that aligns extremely cleanly will have a high enough score to align across the structural variant. Denovo assembly will also be capable of sequence resolving long insertions in a way that is entirely impossible with break-end callers. As a result, one definite pathway to novel results is to augment Manta results for deletions and short insertions with the long insertions discovered by the contig caller.

References

[1] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 2015.

[2] Jesse R Dixon, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nature genetics*, 50(10):1388, 2018.

[3] Ivana Bratić Hench, Jürgen Hench, and Markus Tolnay. Liquid biopsy in clinical management of breast, lung, and colorectal cancer. *Frontiers in medicine*, 5:9, 2018.

[4] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226, 2012.

[5] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

[6] Wenbo Mu, Bing Li, Sitao Wu, Jefferey Chen, Divya Sain, Dong Xu, Mary Helen Black, Rachid Karam, Katrina Gillespie, Kelly D Farwell Hagman, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. *Genetics in Medicine*, page 1, 2018.

[7] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[8] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009.

[9] Todd J Treangen and Steven L Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.

[10] Hunter R Underhill, Jacob O Kitzman, Sabine Hellwig, Noah C Welker, Riza Daza, Daniel N Baker, Keith M Gligorich, Robert C Rostomily, Mary P Bronner, and Jay Shendure. Fragment length of circulating tumor dna. *PLoS genetics*, 12(7):e1006162, 2016.

[11] Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3:160025, 2016.