[Updating]:
Qu-SV: Quick Detection of Structural Variants in Haploid, Diploid and Tumor Genomes

Amirhossein Afshinfard, Damoun Nashta-ali, Seyed Abolfazl Motahari

February 17, 2018

Abstract

[updating - please contact a.afshinfard@gmail.com for slides and online presentation - the content of this paper is not reviewed and corrected for academic writing and grammatical mistakes yet - we are updating the results and figures]

Structural Variants are one the main contributors to different types of genomic disorders. Because of the many challenges in detection (such as repeat regions, sequencing biases and errors, heterozygous variations, etc.), proposing novel and versatile methods is still a topic of concern. In this study, we propose an ultrafast novel approach to detect these changes from sequencing reads, regardless of that the genome is a haploid, diploid or polyploid. In this approach, non-informative reads are being identified and eliminated very quickly and efficiently. Simultaneously, anchoring variation regions, informative read segments build up clusters of events along the genome with some events being connected because of sharing some identical reads. Parsing the resulted sparse graph of events, the algorithm discovers the exact location, type, and sequence of each alteration. Evaluations of this method were successful in both speed and accuracy and additionally showed its ability to detect complex variants.

1 Introduction

Structural Variants (SVs) are generally defined as genetic variations in DNA sequence other than mutations which involve one or a few base pairs. SVs, despite single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs), requires the disruption of the sugar-phosphate backbone of DNA and thus involve more base pairs [1]. In other words, Structural Variation (SV) refers to a variation that changes the structure of the genome. This variation may change the length of the chromosome (Unbalanced SV such as novel insertion, deletion, copy-number variation CNV) or only change the structure without affecting its length and content (balanced SV, including inversion and reciprocal translocation). Figure 1 shows three types of these variations.

Recent research in the last decade has revealed that SVs are much more frequent than previously thought [2, 3] and they build up a large fraction of the human genome variation, even more than SNPs [4, 5]. This opens a new field in genetic and genomic studies and motivates lots of research on methods for detection of structural variants as a primary step for further investigations. The next desired steps are revealing

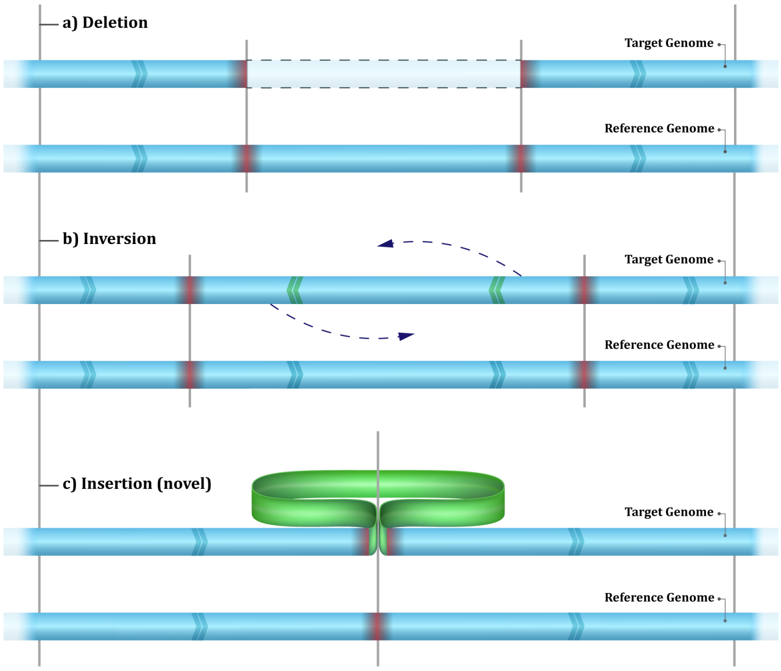

Figure 1: Structural Variation. a) a region on the target genome is deleted, b) an inverted region on the target genome (a balanced SV), and c) a novel sequence inserted.

phenotypic impacts of SVs (disorder and non-disorder impacts) as well as studying their population genetics [6], their formation mechanisms, and genome evolution.

SVs, observed both as germ-line and as somatic variation, contribute to genomic disorders ranging from neurodevelopmental disorders (including schizophrenia [7] and autism [8, 9]) to childrens developmental disorders [10], blood diseases [11], diabetes [12], a wide range of cancers [13, 14, 15], etc. By influencing gene expression directly or indirectly [16], they revealed to involve not only genomic disorders but also other complex traits and phenotypes at various levels [17].

Structural variants are caused by different mutational mechanisms including DNA recombination, replication, and repair-associated processes [1]. Investigating breakpoints of structural variants at base pair resolution is crucial for understanding their formation mechanisms [18]. As SVs are likely responsible for gene and genome evolution [19], the discovery of new structural variants will help to study formation mechanisms and their population genetics to yield a better understanding of genome evolution.

2 Challenges

Detection of SVs is still a topic of concern, provided that there is an annual increase in the number of discovered alterations [20, 21] as there are biases in each method for detection of different types and various lengths. Many methods are limited to detection of longer SVs and some of them are able to search for specific types (e.g. many misses balanced SVs which are of great importance [22]) due to their nature.

Some of these shortcomings are caused by fundamental biases in sequencing [23], reduced read mappability or poor Genome Mappability Score (GMS) [24], especially in repeat-rich regions and thus reduced signal-to-noise ratios. In Addition to these difficulties, it is more challenging should we consider complex SVs, diploid genomes, and sequencing reads from tumor samples containing somatic mutations. [Updating]

3 Current Approaches

There are four general approaches to detect structural variants each having specific advantages and limitations. We will review these methods quickly and list their Strengths and weaknesses.

Assembly-based methods theoretically can discover structural variants of all types. Having built the assembly of the targeted genome, these methods detect structural variants by aligning it to the reference genome. methods of these types are limited to read length, and are involved with other challenges of the assembly problem itself [25]. the computational cost is a case in point. A novel idea in this regard is to use this approach locally (local assembly). using such approach to discover the sequence of novel insertions is inevitable.

Assuming an expected distribution for sequencing depth, **Read-count** (RC or Read-Depth) methods search for SVs by checking the divergence of local estimated distributions from expected ones. The assumption itself is challenging due to the non-uniform behavior of read depth, and some factors found to affect read counts. In example, GC-content has been revealed to affect sequencing process [- -] and result in non-uniform read coverage which can somewhat be corrected and unbiased [- -]. Another challenge in this approach is to decide if a change caused due to noise or because of a variation; so theyre almost insensitive to small SVs. locating breakpoints

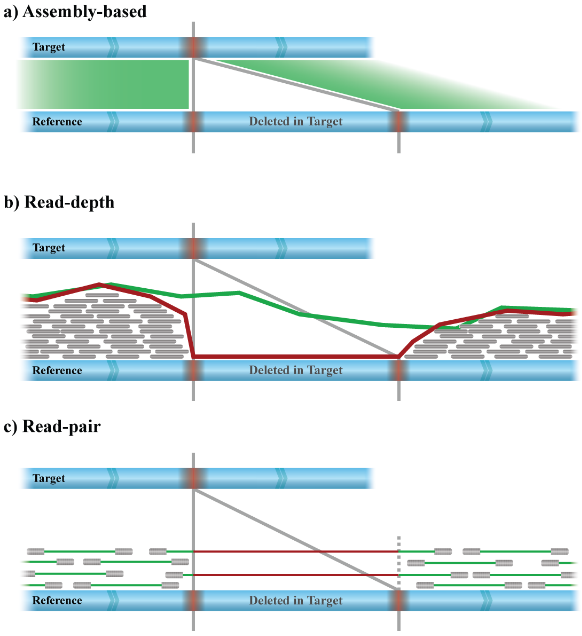

Figure 2: current approaches, [updating]

with base pair resolution could be addressed hardly using RC. In addition, this approach is almost blind to balanced SVs, especially for inversions. However, theyre really powerful to discover deletions and Copy Number Variations (CNVs). Methods based on RC can estimate exact copy number while others are weak or at least are faced with many challenges in this regard.

Read-pair (RP) methods utilize the capability of paired-end reads (or mate-pair reads) to detect SVs. Paired-end reads expected to have a specific orientation and a known insert-size distribution with respect to the reference genome, the sequencing technology, and its parameters. Changes in orientation is a sign (an indication of) of inversion occurrence, while insertion and deletion events result in higher or lesser insert-size. Like RC, the breakpoint resolution is not exact and again it is difficult to decide if a change in insert-size distribution is caused by noise or by a variation event, especially in the case of small SVs. However, they are likely to have more information for such cases comparing with RC and also able to detect inversion. As a limitation, these methods are not sensitive to insertions longer than the insert-size and the detection needs more process steps on one-end aligned reads with using other methods like RC and SR. CNVs will be a big challenge for methods of this class. (smaller SVs in comparison with RC?)

Split-Read (SR) methods can result in breakpoints detection with base pair resolution (precision). A breakpoint-spanning read (an informative read) is unlikely to map, but split subsequences of that read tend to map to distant locations of the reference genome. [Incomplete] better than others for small SVs. Not good for CNVs and poor in repeat regions. Limited to read length. Applicable to the case of paired-end reads, checking the unmapped end of one-end anchored reads.

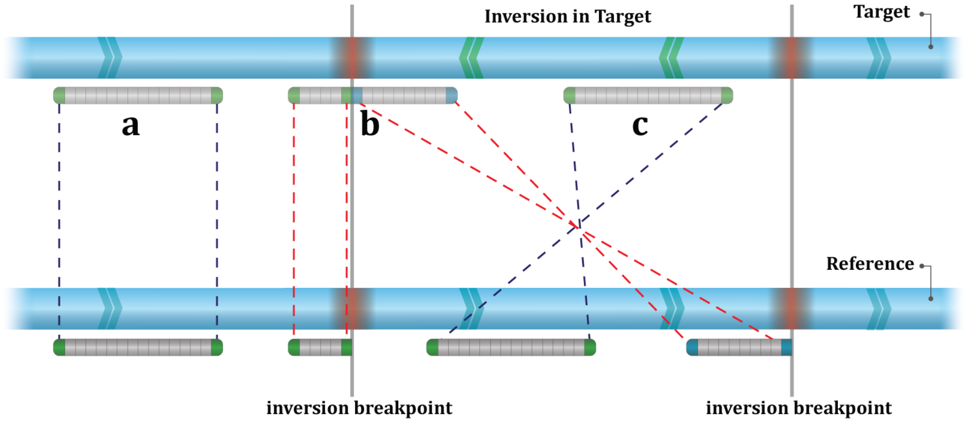

Figure 3: informative vs. non-informative reads from the target genome an their mappings via the assumed ideal aligner to their corresponding locations on the reference. a) a non-informative read in a normal region, b) an informative read which is split for mapping, and c) a non-informative read inside an inversion event

4 Proposed method

4.1 Method Overview

Suppose some reads from a target genome containing SVs. Assume that we have an Ideal read mapper meaning that it can assign each part of a single read to its corresponding location on the reference genome. With this attention, reads are divided into two categories (figure 3). The first group reads which span a breakpoint become split into (at least) two parts each of which maps to a distinct region on the reference. Every read from the second category maps entirely to a specific location without any division. We call the first group informative reads since they capture a breakpoint and thus they can signal a Structural Variation. The latter group reads are virtually non-informative reads. Why virtually? Leave it for now, but lets say they only contain useful information about the depth of their location.

In the first stage, which is the mapping step, our Idea is to remove non-informative reads expeditiously, and to approach that Ideal read mapper for informative reads

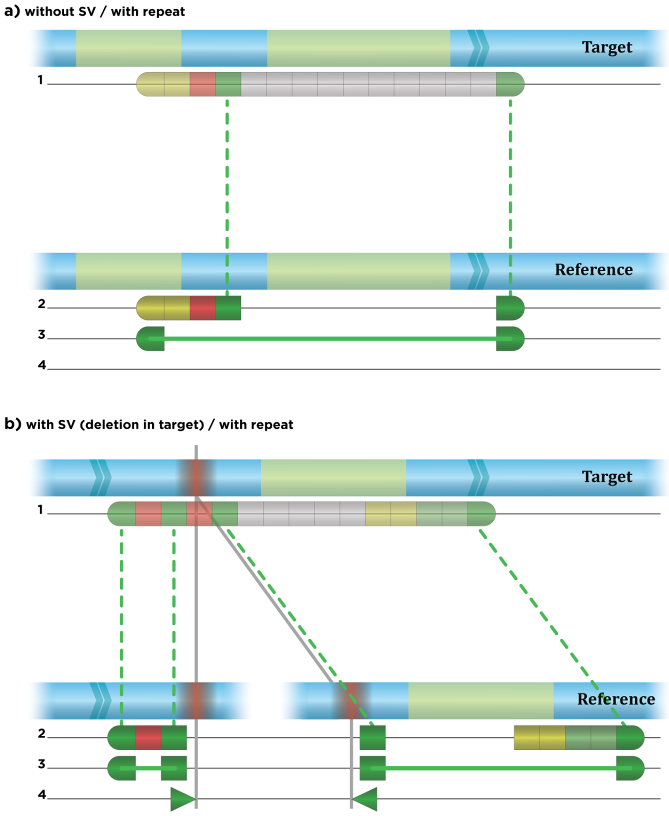
a) without SV / with repeat
Target
1
Reference
2
3
4
b) with SV (deletion in target) / with repeat
Target
1
Reference
2
3
4

Figure 4: read chunking and mapping strategy a) non-informative read b) informative read [updating] / line 1: read sampled from the target genome; line2: mapping of the read chunks to reference genome; line3: after extension; line 4: re-alignment (optional) and breakpoint detection. / Green: unique true matches; Red: unmapped or wrong matches due to sequencing errors or small mutations; Yellow: non-unique chunks

at the same time. For this purpose, we have exploited and revised an existing read aligner, Meta-aligner [26], so it can detect non-informative reads very quickly and efficiently without reporting any informative read as non-informative. In the same run, the revised version is able to split the informative reads into multiple segments and assign each segment into its original source on the reference genome. This way, almost all non-informative reads are identified and their contribution to local depths are considered. Then they are filtered, and the focus will be on split-mapped informative reads to detect SVs in the next stage. Indeed, informative reads pile up a very small fraction of the total number of reads, resulting in an elbowroom for applying algorithms of high complexity in the second stage without concerning about the computational cost.

As mentioned before, non-informative reads are not non-informative at all. Their contribution in local depths is useful, especially for detecting some types of CNVs (like RC methods). Thus, without any extra computation, they account for depth before being removed. Additionally, they are saved in buckets for further access in the second stage if needed.

In the second stage, overlapping informative reads which share the same direction (whether the breakpoint is right or left) contribute to piling up distinct events. Informative reads with more than one location (contributing to more than 1 event) connect some of these events together resulting in a sparse graph. In this graph, events are nodes and identical shared reads between events are edges. Multiple edges is equal to one weighted edge. For detection of SV types from subgraphs in the resulting sparse graph, a statistical approach measures the probability of each subgraph and maximize it given SV $type_i$ (equatuin 1). According to equations 1 and 2, it is done by likelihood maximization. There are five general types of SVs. The ith type which maximizes the likelihood is returned as the type of the SV for that subgraph.

[Updating:]

$$\begin{aligned}\underset{i}{\operatorname{argmax}}\, \mathbb{P}\left(type_i \mid subgraph\right) &= \underset{i}{\operatorname{argmax}}\, \frac{\mathbb{P}\left(subgraph \mid type_i\right)\mathbb{P}\left(type_i\right)}{\mathbb{P}\left(subgraph\right)} && (1)\\ &= \underset{i}{\operatorname{argmax}}\, \mathbb{P}\left(subgraph \mid type_i\right) && (2)\end{aligned}$$

In equation 1, $\mathbb{P}\left(type_i\right)$ is considered as uniform as the distribution of the SV types are unknown. They are not achievable from the current discovered SVs as current methods are biased. Additionally, $\mathbb{P}\left(subgraph\right)$ is not effective in maximization and can be eliminated.

The next step is to define $\mathbb{P}\left(subgraph \mid type_i\right)$ which we will discuss in 4.3 in the incoming official preprint.

Another approach, exploited in the second stage, builds up local assemblies around locations anchored with informative reads. Resulting local assemblies being re-aligned to the reference genome and identify the type of each SV by rearrangement minimization. This approach helps in the detection of Complex SV with complicated breakpoints near each other.

4.2 Stage one in detail (Qu-Break)

4.2.1 Meta-Aligner

The original version of the Meta-Aligner [26] uses the genome statistics and suggests a minimal size $2l$ for unique mapping of a subread, statistically sufficient to confidently assign the read to the location anchored by that subread. In the example of the human genome, the optimal size is approximately 50 bps. Excluding variants from consideration, It means that if we find a 50 bps long sub-read from the read uniquely in the reference, it is enough to be confident that this is the right location for that read. Another great idea of the Meta-Aligner is to divide this needed $2l$ bps unique mapping into two disjoint sub-reads of length l (2x25 for Human genome).

In a nutshell, the algorithm searches for two disjoint sub-reads of length 25, mapped uniquely to the genome and concordantly to each other. Finding these two sub-reads, the algorithm skips the remaining base pairs of the read and assigns the read to the location. Searching for a sequence of length 25 for unique hits in the reference can be done very quickly, making the algorithm extremely fast. Not only does this approach eliminate the cost of mapping of a large portion of many reads, but it also overcomes the challenge of repeat regions when mapping reads with a minimal mapping size. + an example of a run by Meta-Aligner and proportion of reads aligned after each iteration.

4.2.2 Revision: Qu-Break

When considering SVs, some reads (informative reads) are not related to a single location when aligned to the reference genome. Thus, two disjoint sub-reads cannot suggest the location of the read as a whole. However, two sub-reads of an acceptable distance of length L_{in} can decide about the original location of the proportion between the two sub-reads. L_{in} should be small enough that it is possible to distinguish between small scale-mutations and SVs. We will discuss calculating L_{in} in another research but let's assume that it is five times as big as the smallest SV which we want to detect. As a result, if the distance between two concordant sub-reads is less than L_{in}, the location for that proportion of the read is known. Relatively, if we find two pairs of concordant sub-reads, it means that this read captures a breakpoint (assume there is no errors nor noises).

With this in mind, the aim is to revise Meta-Aligner in order to detect non-informative reads quickly and with high accuracy. At the same time, it should find

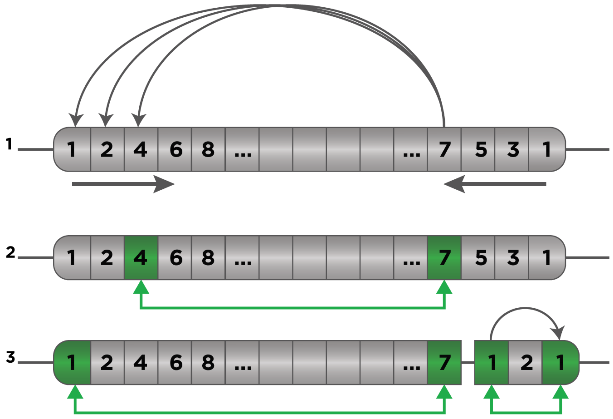

Figure 5: read chunking and mapping strategy - line1: the order of unique mapping and checking for concordance for disjoint sub-reads. line2: found concordant sub-reads. line3: first pair after extension and second pair after recursively checking the remaining chunks

sub-read pairs in informative reads. For this purpose, we propose Qu-Break. This algorithm divides a read of length L_{read} into disjoint sub-reads (chunks) of length l_{sub} like what is depicted in figure 5. The algorithm starts by searching for unique hits for the first and last chunks and continue by searching for inner chunks one by one. Finding two concordantly-mapped sub-reads, the algorithm stops the unique search assuming the proportion between two concordant sub-reads as resolved. Thereafter, an extension phase starts to widen the resolved proportion. We will discuss the extension step later in 4.2.4. After extension, assuming the remaining unresolved basepairs as a new read, the algorithm search for concordant sub-reads in a recursive manner.

With the procedure discussed above, the algorithm find non-informative reads very quickly and by checking only a small proportion of the reads. Statistical analysis of the number of searched sub-reads is available in section 4.2.3. Additionally, practical results for human chr19 is available in section 5.1. In addition, a small proportion of reads, informative reads, are split into multiple segments each of which mapped to a specific location and threaded to each other (figure).

Qu-Break has two other optional procedures which are useful for further research in the second stage. Firstly, Qu-break can save a encoded mapping depth from non-informative reads which is crucial for investigation about the frequency of each SV in input genomes (in case of diploid or polyploid genomes like tumor genomes). Secondly, Qu-Break can save non-informative reads in buckets. Each bucket is for a fixed proportion of the genome. This way, it is possible to access the reads for a specific range very quickly.

4.2.3 Mapping analysis

[Updating:] statistical analysis on the number of checked sub-reads. => speed
figure 6 is related.

4.2.4 Extension and informative l-mer

[updating] figure 6 is related.

4.3 Stage two in detail

4.3.1 Weights analysis

[Updating:] Statistical analysis - expected weights of the nodes and edges - figure 6

4.3.2 Detection method

There are two approaches for this step discussed in summery in section 4.1. The detailed version will be available on the official preprint.

5 Results

[updating - Results on Speed, Recall and Accuracy] Speed ! Recall ! Accuracy !

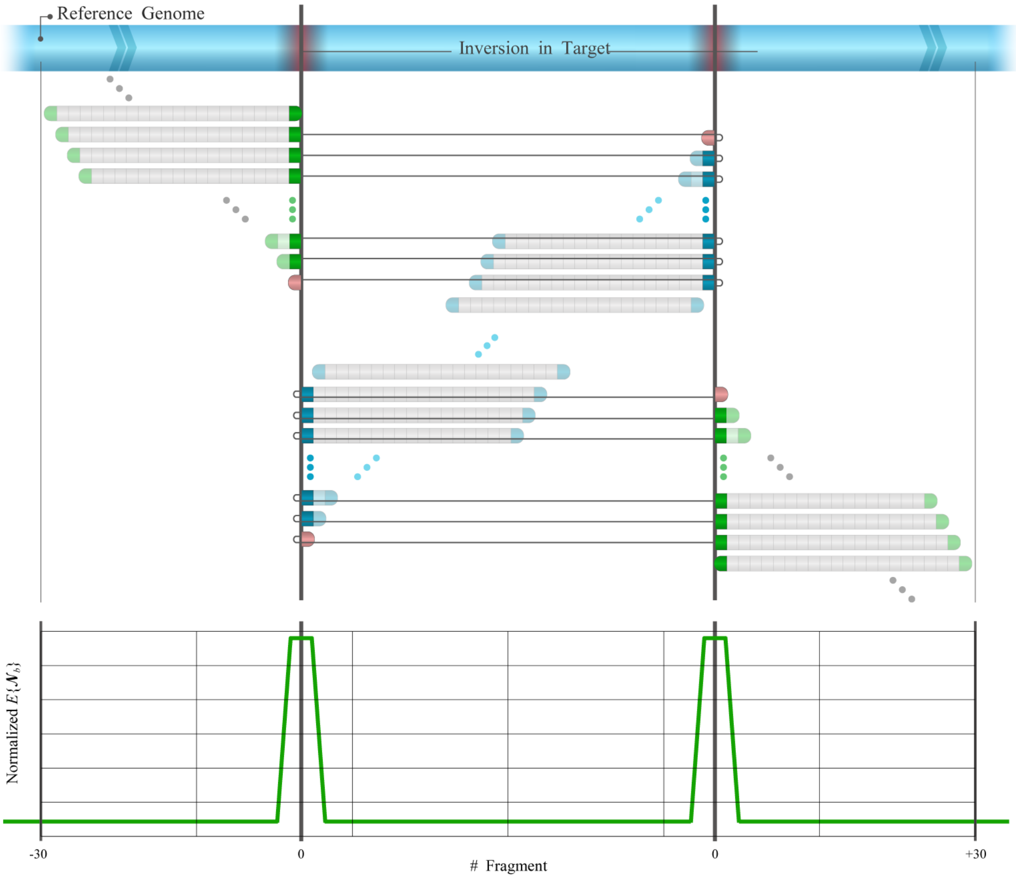

Figure 6: Informative chunks collaborate to signal for SVs - the expected weight for nodes in the sparse graph - Informative read chunks are visible around the breakpoints (green and blue with 100% opacity)

5.1 Mapping statistics for Chr19

5.2 Synthetic data

5.3 Real data

References

[1] C. M. Carvalho and J. R. Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238, 2016.

[2] A. J. Iafrate, L. Feuk, M. L. Rivera, M. N.and Listewnik, P. K. Donahoe, Y. Qi, ..., and C. Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949, 2004.

[3] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, ..., and N. Navin. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

[4] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, ..., and T. Fitzgerald. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704, 2010.

[5] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363, 2011.

[6] D. F. Conrad and M. E. Hurles. The population genetics of structural variation. *Nature genetics*, 39:s30–s36, 2007.

[7] A. Sekar, A. R. Bialas, H. de Rivera, A. Davis, T. R. Hammond, N. Kamitaki, ..., and G. Genovese. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183, 2016.

[8] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, ..., and B. Thiruvahindrapduram. Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488, 2008.

[9] R. K. Yuen, B. Thiruvahindrapuram, D. Merico, S. Walker, K. Tammimies, N. Hoang, ..., and M. J. Gazzellone. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine*, 21(2):185–191, 2015.

[10] D. A. King, W. D. Jones, Y. J. Crow, A. F. Dominiczak, N. A. Foster, T. R. Gaunt, ..., and E. A. Jones. Mosaic structural variation in children with developmental disorders. *Human molecular genetics*, 24(10):2733–2745, 2015.

[11] L. M. Boettger, R. M. Salem, R. E. Handsaker, G. M. Peloso, S. Kathiresan, J. N. Hirschhorn, and S. A. McCarroll. Recurring exon deletions in the hp (haptoglobin) gene contribute to lower blood cholesterol levels. *Nature genetics*, -(-):–, 2016.

[12] M. Zanda, S. Onengut-Gumuscu, N. Walker, C. Shtir, D. Gallo, C. Wallace, ..., and S. S. Rich. A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes. *PLoS Genet*, 10(5):e1004367, 2014.

[13] J. M. Tubio. Somatic structural variation and cancer. *Briefings in functional genomics*, -(-):elv016, 2015.

[14] E. Papaemmanuil, I. Rapado, Y. Li, N. E. Potter, D. C. Wedge, J. Tubio, ..., and I. Martincorena. Rag-mediated recombination is the predominant driver of oncogenic rearrangement in etv6-runx1 acute lymphoblastic leukemia. *Nature genetics*, 46(2):116–125, 2014.

[15] N. Waddell, M. Pajic, A. M. Patch, D. K. Chang, K. S. Kassahn, Bailey P., ..., and M. C. Quinn. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501, 2015.

[16] E. R. Gamazon and B. E. Stranger. The impact of human copy number variation on gene expression. *Briefings in functional genomics*, 14(5):352–357, 2015.

[17] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251, 2009.

[18] A. Abyzov, S. Li, D. R. Kim, M. Mohiyuddin, A. M. Sttz, N. F. Parrish, ..., and J. O. Korbel. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature communications*, 6, 2015.

[19] H. H. Kazazian. Mobile elements: drivers of genome evolution. *science*, 303(5664):1626–1632, 2004.

[20] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, ..., and M. K. Konkel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75, 2015.

[21] J. Huddleston, M. J. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, ..., and P Peluso. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685, 2017.

[22] M. Puig, S. Casillas, S. Villatoro, and M. Cceres. Human inversions and their functional consequences. *Briefings in functional genomics*, 14(5):369–379, 2015.

[23] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, ..., and D. B. . Jaffe. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51, 2013.

[24] H. Lee and M. C. Schatz. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16):2097–2105, 2012.

[25] C. Alkan, S. Sajjadian, and E. E. Eichler. Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61–65, 2011.

[26] D. Nashta-ali, A. Aliyari, A. A. Moghadam, M. A. Edrisi, S. A. Motahari, and B. H. Khalaj. Meta-aligner: long-read alignment based on genome statistics. *BMC bioinformatics*, 18(1):126, 2017.