Qu-SV: Quick Detection of Complex Structural Variants in Haploid, Diploid and Tumor Genomes.

Outline and Paragraphs Summary

- **Introduction**
- Structural Variation Definition
- SV findings and overview of related studies about their importance
    - Impact on phenotype specially disease and disorder cases. + non-disorders.
    - Population genetics of SVs. (+personalized)
    - +Mechanism and +evolution
    - CNVs and inversions are both important
- Newer studies like 1000 genomes – growth in numbers of SVs and methods
- **Challenges**
    - Repeat regions (especially when short reads)
    - Errors in long reads
    - Complex SVs
    - Haploid, diploid and tumor genomes
- **Review of the Methods**
    - assembly methods
    - read-count methods
    - read-pair methods
    - split-read methods
    - combined methods + …
    - a table and parameters: small SVs, Long SVs, base-pair precision, Repeat, CNV, Inversion as well,
    - our idea (Framework) – a combination of all these methods
- Long Reads for SV detection…
- **Our Method**
    - Big picture of approach
    - Main stages
    - Meta-Aligner revisited

Introduction

1) Structural[A1] Variation (SV) refers to changes in DNA sequence with length more than ~50 base pairs. In a more informative definition, Structural Variation (as its name suggests) is a broad range of genetic variations which change the structure of the genome. These variations may change the length of the chromosomes (Unbalanced SV such as novel insertion, deletion, duplication, etc.) or only change the structure without affecting their length and content (balanced SV, including inversion and reciprocal translocation). Despite single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs), structural variants requires the disruption of the sugar-phosphate backbone of the DNA and involve many base pairs [15]. In this study, SV refers to all types of variations excluding SNP/SNV.

1) Structural[A2] Variation (SV) refers to broad range of genetic variations in DNA sequence that change the structure of the genome. These variations may change the length of the chromosomes (Unbalanced SV such as novel insertion, deletion, copy-number variation) or only change the structure without affecting their length and content (balanced SV, including inversion and reciprocal translocation). Despite single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs), structural variants requires the disruption of the sugar-phosphate backbone of DNA and involve many base pairs [15]. With this in mind, in this study, SV refers to changes with length more than ~50 base pairs.

1*) Structural[A3] 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 [15]. In other words, Structural Variation (SV) refers to genetic variation in DNA sequence 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). [Fig1: SV types]

2) Recent researches in the last decade have revealed that SVs are much more frequent than previously thought [1-2] and they build up a large fraction of the human genome variation, even more than SNPs [3-3.5]. 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 phenotypic impacts of SVs (disorder and non-disorder impacts) as well as studying their population genetics [14], their formation mechanisms and genome evolution.

3) SVs, observed both as germ-line and as somatic variation, contribute to genomic disorders ranging from[A5] neurodevelopmental disorders (including schizophrenia [4] and autism [5-6]) to children's developmental disorders [7], blood diseases [8], diabetes [9], a wide range of cancers [10-11-12], etc. By influencing gene expression directly or indirectly [13], they revealed to involve not only genomic disorders but also other complex traits and phenotypes at various levels [13.5].

4) Structural variants are caused by different mutational mechanisms including DNA recombination, replication and repair-associated processes [15]. Investigating breakpoints of structural variants at base pair resolution is crucial for understanding their formation mechanisms [16]. As SVs are likely responsible for gene and genome evolution [17], the discovery of new

structural variants will help to study formation mechanisms and their population genetics to yield a better understanding of genome evolution.

Challenges

5) Detection of SVs is still a topic of concern, provided that there is an annual increase in the number of discovered alterations [?] 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[18]). Some of these shortcomings are caused by GC biases in read sampling [?], reduced read mappability 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.

CNVs are very important because … but Inversion needed too (balanced SV but important) – paper2015
Newer studies like 1000 genomes – growth in numbers of SVs and methods

Repeat…
Read mappability + Structural variants often occur in repeat-rich, recently segmentally duplicated genomic areas. Thus, mapping breakpoints can be complicated by the reduced signal-to-noise ratios of high-throughput genomics technologies.
[3.5]

…. Biases in reads (sampling and depth..)
Complex SVs:
one of major challenges in SV detection is the case of c
Furthermore, structural variants can arise from complex alterations involving numerous breakpoints (FIG. 1). Catastrophic chromosome alterations, termed chromothripsis (FIG. 1Be), in which dozens to hundreds of breakpoints are thought to form in a single dramatic structural variant formation event, constitute an extreme example of such complexity.
è [3] 2013 - Nat - c108 - Phenotypic impact of genomic structural variation insights from and for human disease

è Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144, 27–40 (2011).

Review of the methods

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.
1) **Assembly-based** methods theoretically can discover structural variants of all types by assembly of targeted genome and detecting rearrangements comparing with reference genome. These

methods are limited to read length and involved with other challenges of the assembly problem itself [22]. Expense of computation is another drawback but using such approach to discover sequence of novel insertions is Inevitable.
2) Assuming an expected distribution for read 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 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 correct[A6] ed and unbiased[A7] [-]. Another challenge in this approach is to decide if a change caused due to noise or because of a variation; so they're almost insensitive to small SVs. locating breakpoints with base pair resolution could be addressed hardly using RC. In addition, this approach is almost blind to balanced SVs especially for inversion. However, they're 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.
3) **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?)
4) **Split-Read** (SR) methods can result in breakpoints detection with base pair resolution (precision). A read capturing a breakpoint, is unlikely to be able to map, but split subsequences of that read are more tend to map to distant locations of the reference genome. … . 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 unmapped end of one-end anchored reads.

Our Method

Overview

Our approach: Filter quickly and efficiently, detect freely with graph decomposition or with local assemblies starting from informative reads.

Suppose some reads from a target genome which have some 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[Fig2]. The first group reads which cover a breakpoint become split into two parts each of which maps to a distinct region on the reference. Every read from the second category map entirely to a specific location without any division. We call the first group *informative reads* since they captured a breakpoint and they can signal a Structural Variation. The latter group reads are virtually *non-informative* reads.

Why virtually? Leave it for now but let's say they have lesser information compared to *informative* ones.

In the first stage, our Idea is to remove non-informative reads expeditiously, and to approach that Ideal read mapper for informative reads at the same time. For this purpose, we have exploited and revised an existing read aligner Meta-Aligner [Meta-Aligner] 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 to its original source on the reference genome. This way, all non-informative reads are filtered and the focus will be on mapped informative reads to detect SVs. 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.

For the second stage, (*) Unmapped reads(1) use mapped segments of the informative reads, and read depth from non-informative + reuse unmapped with multiple alignment. (2) assembly

In detail

The original version of the Meta-Aligner uses the genome statistics and suggests a minimal size for unique mapping of a subread, statistically sufficient to confidently assign the read to the location anchored by that subread. In the example of human genome the optimal size is approximately 50 bps. Excluding variants from consideration, It means that if we find a 50 bps long subread 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 50 bps unique mapping into two disjoint subreads of length 25. In a nutshell, the algorithm searches for two disjoint subreads of length 25, mapped uniquely to the genome and concordantly to each other. Finding these two subreads, the algorithm skips the remaining basepairs of the read and assigns the read to the location. Searching for 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 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 read aligned after each iteration.

In a more detailed description, Metal-Aligner divides the reads into its disjoint comprising *l*-mers, with *l* bigger than half of the minimum required length for confident unique matches. Thereafter, it start to search for unique matches for each l-mer one by one. Finding a new l-mer uniquely, Meta-Aligner checks it with previous l-mers for concordant locations and should it find a concordant pair, it skips the remaining *l*-mers.

***** More Ideas:

Find BP and start local assemblies using informative reads as primers. + use depths reported by all reads. => assembly + split + read count + ?paired-read.

In Theory - Methods:

All the data in:

Read-Count: only keeping read depth, and reads are discarded. A fraction of inversions will miss certainly (could be checked theoretically and technically). A fraction of SVs with a special length will miss certainly with respect to read length.

به طور مثال میشه نشون داد که تفاوت‌های ساختاری با توجه به اندازشون چقدر احتمال داره که miss بشن. مثلا اگر طول read فلان قدر باشه و عمق هم فلان قدر باشه و ارور هم فلان شکل باشه، پس اگر یک تفاوت با طول فلان داشته باشیم با احتمال نزدیک به یک miss میشه. همینطور درمورد inversionها. با حتی دقیق تر بگیم که اگر در ژنوم انسان باشیم و ... .

Read-pair: only mapped insert-size and orientations and reads are discarded.

مثل مورد بالا میشه عمل کرد.

References

[1] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., ... & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics*, *36*(9), 949-951..

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

[3] Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., ... & Fitzgerald, T. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704-712.

[3.5] Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*(5), 363-376.

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

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

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

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

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

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

+ nature 2014: Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes – prof. fereydoun azizi and 2 more Iranian authors (but not SV)

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

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

[12] Papaemmanuil, E., Rapado, I., Li, Y., Potter, N. E., Wedge, D. C., Tubio, J., ... & Martincorena, I. (2014). RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature genetics*, *46*(2), 116-125.

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

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

[14] Conrad, D. F., & Hurles, M. E. (2007). The population genetics of structural variation. *Nature genetics*, *39*, S30-S36.

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

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

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

[18] Puig, M., Casillas, S., Villatoro, S., & Cáceres, M. (2015). Human inversions and their functional consequences. Briefings in functional genomics, 14(5), 369-379.

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

More examples of SVs in disease:

Parkinson :

Polymeropoulos MH, Higgins JJ, Golbe LI, et al. 1996. Mapping of a gene for Parkinson’s disease to chromosome 4q21–q23. Science 274:1197–99

Singleton AB, Farrer M, Johnson J, et al. 2003. Alpha-synuclein locus triplication causes Parkinson’s disease. Science 302:841

Chartier-HarlinMC,KachergusJ,RoumierC,etal.2004.Alpha-synuclein locus duplication as a cause of familial Parkinson’s disease. Lancet 364:1167–69

Ib´ a˜nezP, BonnetAM, DebargesB, etal. 2004 .Causal relation between alpha-synuclein gene duplication and familial Parkinson’s disease. Lancet 364:1169–71

Ib´ a˜nez P, Lesage S, Janin S, et al. 2009. Alpha-synuclein gene rearrangements in dominantly inherited parkinsonism: frequency, phenotype, and mechanisms. Arch.Neurol. 66:102–8

Maraganore DM, de Andrade M, Elbaz A, et al. 2006. Collaborative analysis of α-synuclein gene pro- moter variability and Parkinson disease. JAMA 296:661–70

Alzheimer:

Rovelet-LecruxA,HannequinD,RauxG,etal.2006. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat.Genet. 38:24–26

TheunsJ,BrouwersN,EngelborghsS,etal.2006. Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. Am.J.Hum.Genet. 78:936–46

SulsA,ClaeysKG,GoossensD,etal. 2006. Microdeletions involving the SCN1A gene maybe common in SCN1A-mutation-negative SMEI patients. Hum.Mutat. 27:914–20

Epilepsy:

Not searched yet

:A1]3 types of Variation]
SV
Microindels
SNPs

:A2]3 types of Variation]
SV
Microindels
SNPs

:A3]3 types of Variation]
SV
Microindels
SNPs

A4]Could be described]

A5]Disorders need to be reorder with some helps from an expert in biology or medicine. I tried] (1
…to do it best but maybe
Parkinson + Alzheimer +
epilepsy + major depressive disorders 2014 dushlaine +

A6]Not fully]

A7]Wrong grammer]