Methods 102 (2016) 36–49

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Structural variation detection using next-generation sequencing data A comparative technical review

Peiyong Guan [a], Wing-Kin Sung [a,b,*]

[a] School of Computing, National University of Singapore, 117543, Singapore
[b] Computational & Mathematical Biology Group, Genome Institute of Singapore, 138672, Singapore

ARTICLE INFO

Article history:
Received 18 October 2015
Received in revised form 9 January 2016
Accepted 31 January 2016
Available online 1 February 2016

Keywords:
Structural variation
Next-generation sequencing

ABSTRACT

Structural variations (SVs) are mutations in the genome of size at least fifty nucleotides. They contribute to the phenotypic differences among healthy individuals, cause severe diseases and even cancers by breaking or linking genes. Thus, it is crucial to systematically profile SVs in the genome. In the past decade, many next-generation sequencing (NGS)-based SV detection methods have been proposed due to the significant cost reduction of NGS experiments and their ability to unbiasedly detect SVs to the base-pair resolution. These SV detection methods vary in both sensitivity and specificity, since they use different SV-property-dependent and library-property-dependent features. As a result, predictions from different SV callers are often inconsistent. Besides, the noises in the data (both platform-specific sequencing error and artificial chimeric reads) impede the specificity of SV detection. Poorly characterized regions in the human genome (e.g., repeat regions) greatly impact the reads mapping and in turn affect the SV calling accuracy. Calling of complex SVs requires specialized SV callers. Apart from accuracy, processing speed of SV caller is another factor deciding its usability. Knowing the pros and cons of different SV calling techniques and the objectives of the biological study are essential for biologists and bioinformaticians to make informed decisions. This paper describes different components in the SV calling pipeline and reviews the techniques used by existing SV callers. Through simulation study, we also demonstrate that library properties, especially insert size, greatly impact the sensitivity of different SV callers. We hope the community can benefit from this work both in designing new SV calling methods and in selecting the appropriate SV caller for specific biological studies.

Contents

* Corresponding author at: School of Computing, National University of Singapore, 117543, Singapore.
E-mail address: ksung@comp.nus.edu.sg (W.-K. Sung).

http://dx.doi.org/10.1016/j.ymeth.2016.01.020
1046-2023/

1. Structural variations (SVs)

Structural variations (SVs) are large-scale changes in the genome, often more than 50 nucleotides [23]. They can be classified into different types (see Fig. 1) based on read pair information. *Deletion* is the removal of DNA sequence from the genome. *Insertion* is the addition of DNA sequence into the genome. There are two types of insertions, depending on whether the inserted sequence is from the genome of the sample. If the inserted sequence is not from the genome of the sample, the insertion is called novel insertion. One example is the insertion of the hepatitis B virus (HBV) into the human genome in hepatocellular carcinoma [96]. *Duplication* is the copying of one DNA sequence and pasting it to the genome. Depending on the pasting position, duplication can be classified as interspersed duplication and tandem duplication. *Inversion* involves the breaking of a DNA sequence at two loci and inverse it, resulting a reversed sequence. *Translocation* involves the deletion of a DNA sequence from one locus and inserting it to another locus in the genome. Depending on whether the chromosome of the source locus is the same as that of the target locus, translocations can be further classified as intra-chromosomal translocation and inter-chromosomal translocation.

Deletions, insertions and duplications alter the copy number of the genome and are thus called *unbalanced* SVs. Inversions and translocations don't change the copy number and are called *balanced* SVs.

Besides the simple SVs described thus far, combinations of these SV events can occur. For example, the duplicated sequences can be inverted before being inserted into the target loci. Duplications of this nature are called *inverted duplication* (see Fig. 1G), which can be formed via breakage-fusion-bridge cycle [11]. Chromothripsis is another type of chromosomal rearrangements, where the changes are so intense that the region involved is changed beyond recognition [107]. Various other types of SVs could happen, resulting from different formation mechanisms [79,108].

Compound SV events also occur. For example, if one parent has two normal chromosomes A and B and the other parent has a balanced translocation between A and B, the child can inherit a normal chromosome A from one parent and a rearranged chromosome B from the other parent (see Fig. 1H). If only the child's sample is available, it appears that the child has an *unbalanced translocations*, since the child has an additional copy of genes originally on chromosome A and reduced copy of genes originally on chromosome B. Such compound events can only be properly explained when all samples of the pedigree are available. Many studies are designed to sequence the pedigree (e.g., CEPH pedigree 1463, http://www.illumina.com/platinumgenomes/) and compare the variations between the parents and children to not only study the mechanisms of the SVs but also estimate the false discovery rate of various detection tools.

SVs affect the activities within our cells, including alteration of gene copy number and change of gene regulation [108]. The variation of copy numbers often leads to change of gene dosage, further disrupting and perturbing biological pathways, leading to undesired biological and physiological conditions [101]. SVs may also cause breaking and linking of genes and thus have been

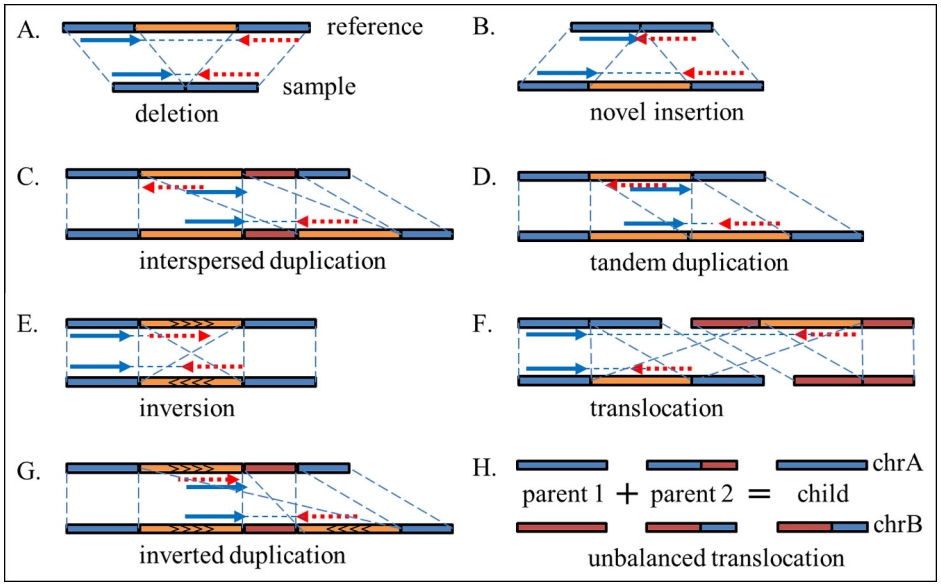

Fig. 1. Different types of SVs and discordantly mapped reads. Blue arrow representing reads from the 5′ end and red arrow representing reads from the 3′ end. The first line of each SV type in A to G represents the reference genome sequence and the last line represents the sequence in the sample. The orange-colored sequence is the sequence being deleted, inserted, duplicated or inverted. H shows a compound event leading to an unbalanced translocation.