Finding Deletions with Exact Break Points from Noisy Low Coverage Paired-end Short Sequence Reads

Jin Zhang and Yufeng Wu

Department of Computer Science and Engineering,
University of Connecticut,
Storrs, CT 06269, U.S.A.
{jinzhang,ywu}@engr.uconn.edu

Abstract. While there is great interest in structural variations, it is very challenging to identify them using short next generation sequencing reads, especially with noisy low coverage data. Statistical methods have been used in discovering possible structural variations by examining the changes in insertion sizes, while other methods focus on finding the exact break points by mapping the short sequence reads. Due to the low quality of data, none of the existing methods are completely satisfying. In this paper, we first find candidate deletions by mapping from both ends of a read towards the central positions. The two split parts of a read that mapped at two positions on the reference may indicate the deletion break points. After candidate deletions are found, we use paired-end reads spanning the deletions to validate them. Our method allows single nucleotide polymorphisms (SNPs) and small indels, and thus more split-reads crossing a deletion can be discovered. This increases the power in finding the deletion break points. Our mapping approach is based on Burrows Wheeler transform (BWT), which is highly efficient in mapping short reads. When applying our method on 1000 Genome pilot low coverage data, we experimented on chromosome 1 of the CEU population by using low coverage Illumina data. It shows that our method finds more deletions that have been reported by 1000 Genome low coverage structural variation release than an alternative approach. The running time of our method is faster by a factor of 3 while is searching long deletions upto 1Mb.

Keywords: Structural variation, Next-generation sequencing, Burrows Wheeler transform