MindTheGap: integrated detection and assembly of short and long insertions

Guillaume Rizk, Anaïs Gouin, Rayan Chikhi, Claire Lemaitre

MindTheGap : integrated detection and assembly of short and long insertions

Guillaume Rizk[1], Anaïs Gouin[1], Rayan Chikhi[2] and Claire Lemaitre[1]

[1] Inria/IRISA GenScale, Campus de Beaulieu, 35042 Rennes cedex, France

[2] Department of Computer Science and Engineering, Pennsylvania State University, USA.

Contact : guillaume.rizk@inria.fr, claire.lemaitre@inria.fr

Structural variants (SV) are large-scale structural changes in the genome, that have been shown to play an important role in evolution and disease. There are several types of SVs, in this work we focus on insertion variants: sequences that are present at one site (position) in the donor genome but are absent from the reference genome at this site.

Such variants are difficult to detect from short read sequencing data, especially when they exceed the paired-end insert size. Many approaches have been proposed to call short insertion variants based on paired-end mapping, such as SOAPindel. However, there remains a lack of practical methods to detect and assemble long variants.

We propose here an original method, called MindTheGap, for the integrated detection and assembly of insertion variants from re-sequencing data. Importantly, it is designed to call insertions of any size, whether they are novel or duplicated, homozygous or heterozygous in the donor genome. The software performs three steps: (1) construction of the de Bruijn graph of the reads, (2) detection of insertion breakpoints on the reference genome (find module) and (3) local assembly of inserted sequences (fill module). Both the detection step and the assembly step rely solely on the constructed graph.

The graph is constructed using the algorithms implemented in the Minia assembler, a memory-efficient data structure enabling high scalability.

Insertion sites are detected by scanning the reference genome and testing membership of reference k-mers in the de Bruijn graph. Homozygous and heterozygous insertions are handled using two different methods: they generate distinctive patterns, respectively a sequence of at most k-1 missing k-mers spanning the insertion site, or two consecutive k-mers with the first being right-branching, and the second left-branching.

The fill module starts from a given insert site represented by flanking kmers (L, R) and performs *de novo* assembly to attempt to reconstruct the inserted sequence. In a nutshell, a graph of contigs is constructed by performing breadth-first traversal of k-

mers, starting from L. Then all paths between L and contigs containing R are returned as putative insertion sequences.

MindTheGap was first evaluated on simulated read datasets of various genome complexities (*E. coli*, *C. elegans* and a full human chromosome). This showed that MindTheGap is able to detect with high recall and precision insertions of any size in simple genomes. On a simulated *C. elegans* dataset of homozygous insertions, MindTheGap and SoapIndel showed similar recall and precision, while needing much less computational resources. Moreover, SoapIndel appeared to be limited to insertions smaller than 200bp, whereas MindTheGap still got 79.5 % recall and 97.3 % precision on simulated 1KB insertions. When applied to real *C. elegans* and human NA12878 datasets, MindTheGap detected and correctly assembled insertions longer than 1 kb, using at most 14 GB of memory.