# Identification, reconstruction and validation of insertions in resequencing projects with GapFiller

F. Nadalin✉, F. Vezzi, S. Scalabrin, A. Policriti
Istituto di Genomica Applicata (IGA), Italy

## Motivations

A difficult problem in resequencing projects is the identification, reconstruction, and validation of inserted regions within an unknown genome, with respect to a reference one. In most organisms, many structural variants are found in highly repetitive regions of the genome, making their identification difficult. Recent studies demonstrate the feasibility of detecting structural variants using next-generation, paired-end sequencing reads. Reconstructing completely or at least both ends straddling the inserted sequence is a first validation step. Both reconstruction and validation are still weak in existing tools.

## Methods

GapFiller is a tool developed to fill with a single sequence the gap between paired reads produced by NGS technologies. It is based on a seed-and-extend scheme and it implements techniques to avoid errors in produced contigs. A contig spanning the whole gap is dubbed certified if the mate of the seed read is found. As a matter of fact, GapFiller can be applied to whatever pair of sequences known to lay at an estimated distance, as long as a set of uniformly distributed short reads are provided as input to fill the gap. Our pipeline to reconstruct and validate insertions in a resequenced individual is divided in two main phases: the first one consists in constructing contigs straddling the borders of putative insertions, the second one in filling the gap between them. Before running the pipeline we need to determine locations of putative insertions and to extract the reads aligning around them. More specifically, if insertions in organism A with respect to organism B are to be investigated, we first extract locations of putative insertions, using a tool designed for this purpose (i.e., BreakDancer). Then the reads of A are aligned against the reference B and those mapping next to insertions are extracted, as well as their, possibly unmapped, mates, with the proper orientation. GapFiller is then run twice: a first time to fill the gap between the extracted paired reads in order to reconstruct the borders of each insertion, and a second time to reconstruct the sequence between the certified contigs produced. In the latter phase we treat contigs on the left and on the right side of an insertion as if they were the two pairs of a paired read, respectively, and GapFiller's output will be a super-contig. The event that the super-contig obtained starting from the left contig finally matches against the right one, represents an evidence that we have reconstructed the desired sequence. Clearly, the level of confidence is a function of the number of super-contigs for each (putative) insertion. Moreover, as an important byproduct, we have the assembly of the missing sequence.

## Results

In order to check correctness we tested our pipeline on a real dataset, consisting of a 30x coverage of paired reads from the Vitis vinifera variety PN40024 (485Mbp), for which the reference genome is known. Using BreakDancer we extracted pairs of coordinates on the reference corresponding to deletions on the resequenced variety Sangiovese. This way we simulated insertions in PN40024, with the advantage of being able to check if the sequences assembled by GapFiller were correct, by simply aligning them against the reference. As a preliminar validation step, for each sequence S identifying a putative insertion we computed the maximal tails of S covered by the (certified) contigs produced in the first phase, i.e. we consider only the contigs consisting of paired reads whose gap has been successfully filled. In particular, we identified 800 putative insertions for which we were able to correctly assemble at least 200bp on both tails. The second phase is more difficult as we try to reconstruct (probably highly) repetitive regions. For a few pairs of left and right contigs we were able to entirely reconstruct the inserted sequence, with respect to the reference genome PN40024. However, this point requires a deeper analysis and we are currently working on the improvement of the final step of our pipeline.