NextGenMap-LR: Highly accurate read mapping of third generation sequencing reads for improved structural variation analysis

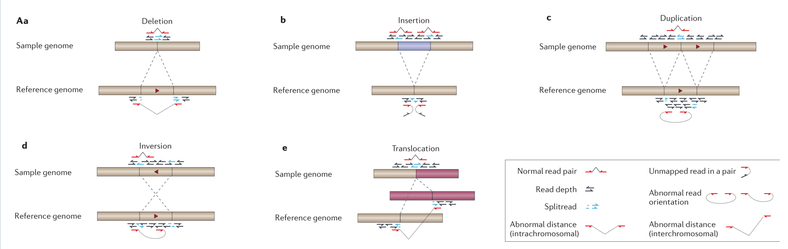
Philipp Rescheneder[1], Fritz J. Sedlazeck[2], Arndt von Haeseler[1,3], Michael C. Schatz[2]

[1] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria, [2] Department of Computer Science, Johns Hopkins University, Johns Hopkins University, Baltimore, MD, USA, [3] Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

1. Introduction

Characterizing genomic **structural variations (SV)** is vital for understanding the mechanisms behind a wide range of diseases including cancer, autism, and schizophrenia. Nevertheless, due to their complexity they remain harder to detect and less understood than SNPs. Recently, **third-generation sequencing** has proven to be an invaluable tool for detecting SVs. The higher read length allows single reads to span a SV and reliable mapping to repetitive regions of the genome. However, current technologies show a raw read error rate of 10% or more consisting mostly of insertions and deletions. Due to this high error rate **current mapping methods fail to find** exact borders for **SVs**, split up large SVs into several small ones, or completely fail to detect certain SVs. Here we present ***NextGenMap-LR*** for long single molecule PacBio and Nanopore reads which addresses these issues. Using simulated data, verified SVs from healthy human samples and human cancer samples we show how the combination of highly accurate NextGenMap-LR alignments and ***Sniffles***, a structural variation caller specifically developed for noisy long-read data, **enables the full characterization of complex SVs** even at low coverage.

2. Structural Variations



Different types of structural variations (SVs) [a]

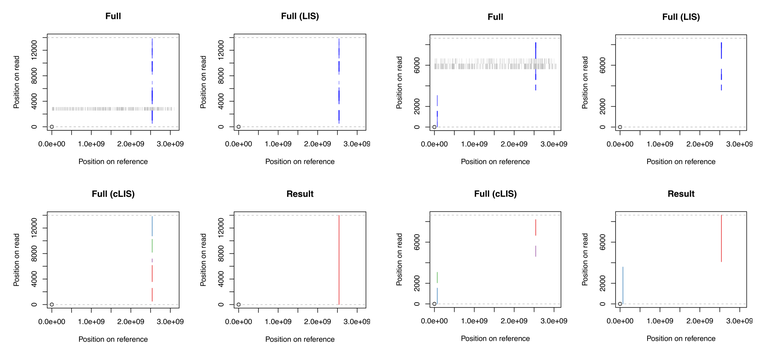
[a] Weischenfeldt, J., Symmons, O., Spitz, F., Korbel, J.O. Phenotypic impact of genomic structural variation: Insights from and for human disease (2013) Nature Reviews Genetics, 14 (2), pp. 125-138.

3. Example Alignments

Example regions containing a 300bp deletion (left) and a 200bp insertion (right).

SKBR3 breast cancer reads aligned using BWA-mem with "-x pacbio". Although, BWA-mem in general produces very accurate alignments, larger **SVs often cause misalignments**.

5. Candidate Search

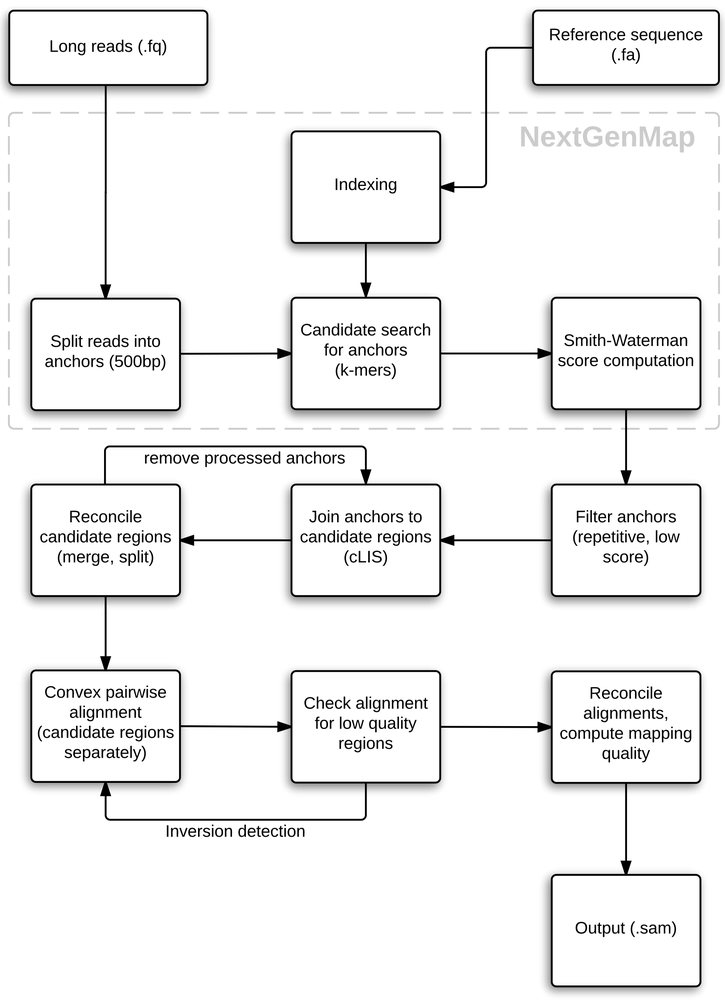


The high-quality anchors retrieved from the initial k-mer search are used to determine whether a read spans a large or none linear SV and has to be **split** (right) or can be **aligned contiguously** (left).

6. Alignment Step

To compute the final alignment(s) we use a Smith-Waterman algorithm. To account for both the **sequencing error** (short and randomly distributed indels) and real **genomic variations** (typically, longer indels), we employ a heuristic non-affine gap model (gap decay) that penalizes gap extensions for longer gaps less than for shorter ones and does not increase the time complexity of the algorithm.

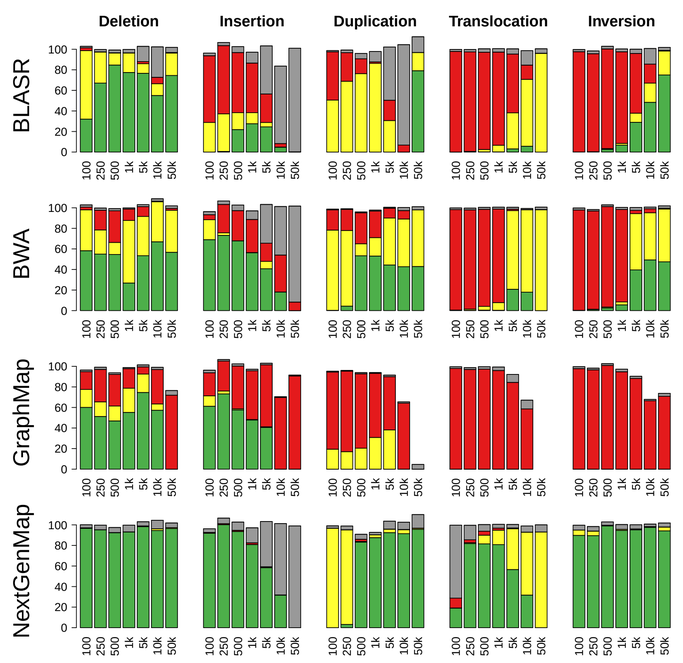
4. Workflow



NextGenMap-LR comprises four main steps:

1. Identify initial anchors
2. Verify anchors with vectorized Smith-Waterman algorithm (scores only)
3. Filter anchors and find candidate regions for the alignments
4. Compute the full alignment between the read and the respective candidate reference regions

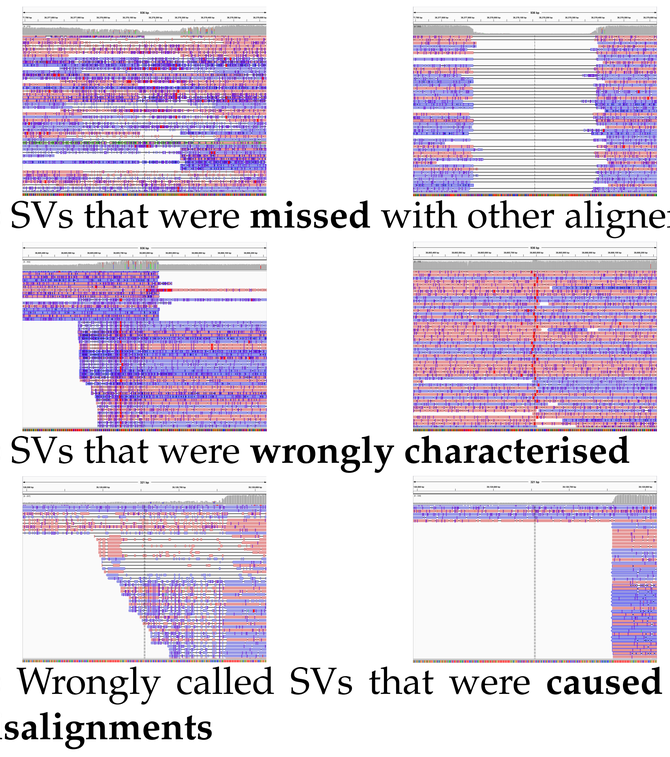
7. Simulated data



Green: correct SV indicated and correct break points (+/- 10bp) found, Yellow: correct SV indicated, Red: wrong SV indicated, Grey: no SV indicated

8. Skbr3 data

To evaluate *NextGenMap-LR* we mapped a full SKBR3 and the genome in a bottle PacBio data set and compared the results to BLASR and BWA-MEM alignments:

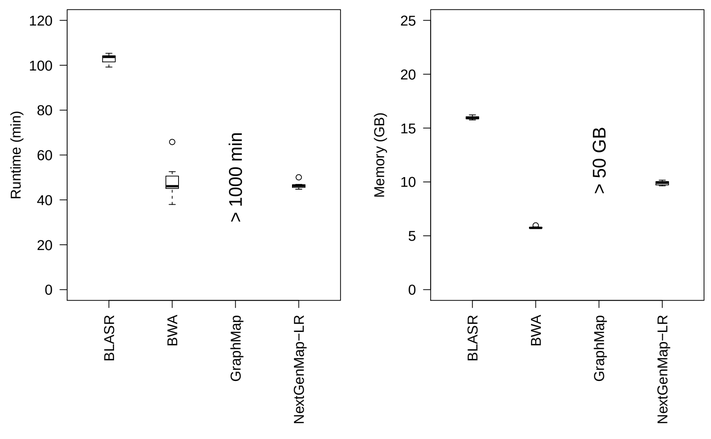


(1) SVs that were **missed** with other aligners

(2) SVs that were **wrongly characterised**

(3) Wrongly called SVs that were **caused by misalignments**

9. Runtime & Memory



Ten fastq files contining 1 GBp of read data were aligned using 10 CPU threads.

10. Sniffles

To accurately detect structural variations in PacBio and Oxford Nanopore data, NextGenMap-LR is tuned to work hand in hand with the structural variation caller Sniffles. Using NextGenMap-LR alignments, Sniffles outperforms all other SV analysis approaches in both sensitivity and specificity.