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

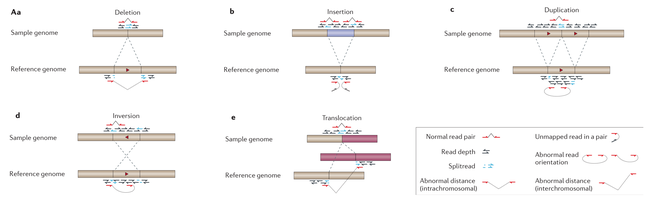
Philipp Rescheneder[1], Fritz J. Sedlazeck[2], Maria Nattestad[2], Arndt von Haeseler[1], Michael C. Schatz[2]

[1] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria, [2] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Introduction

Characterizing genomic **structural variations (SV)** is vital for understanding how genomes evolve. Furthermore, SVs are known for playing a role in a wide range of diseases including cancer, autism, and schizophrenia. Nevertheless, due to their complexity they remain harder to detect and less understood than single nucleotide variations. Recently, **third-generation sequencing** has proven to be an invaluable tool for detecting SVs. The markedly higher read length not only allows single reads to span a SV, it also enables **reliable mapping** to repetitive regions of the genome. However, current sequencing technologies like PacBio show a raw read error rate of 10% or more consisting mostly of indels. Especially in repetitive regions the high error rate causes current mapping methods to fail finding exact borders for SVs, to split up large deletions and insertions into several small ones, or in some cases, like inversions, to fail reporting them at all. Here we present *NextGenMap-LR* for long single molecule PacBio reads which addresses these issues.

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.

Example Alignments

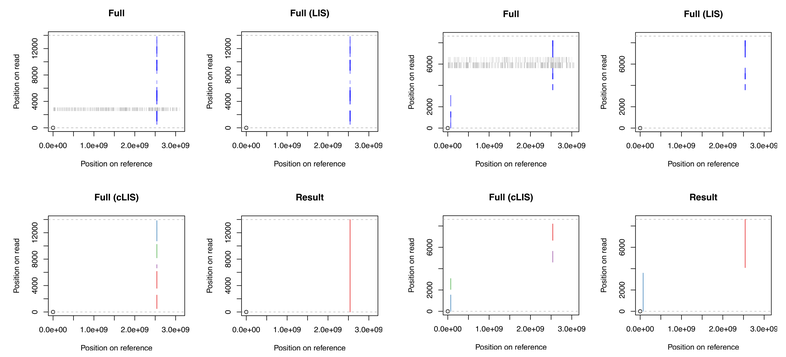
Reads from a SKBR3 breast cancer sample were aligned using BWA-mem 0.7.10 with "-x pacbio". Although, BWA-mem in general produces very accurate alignments, larger **SVs often cause misalignments**. The figure shows example regions containing a 300bp deletion (left) and a 200bp insertion (right).

NextGenMap-LR

NextGenMap-LR comprises four main steps:

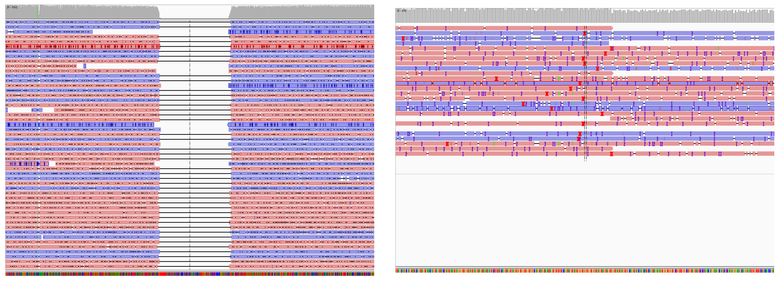
1. Identify initial anchors
2. Verify anchors with vectorized Smith-Waterman algorithm (scores only)
3. Filter anchors and find candidate regions for the alignments
4. Compute the full alignment between the read and the respective candidate regions using a modified version of the Smith-Waterman algorithm

Candidate Search



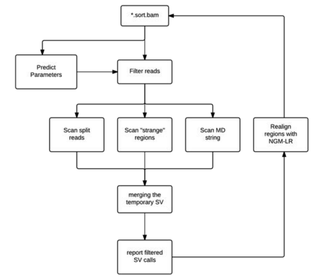
The high-quality anchors retrieved from the initial k-mer search are used to determine whether a read spans a large or none linear SV and has to be **split** (right) or can be **aligned contiguously** (left).

Alignment Step



To compute the final alignment(s) we use a banded 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 alignment computation.

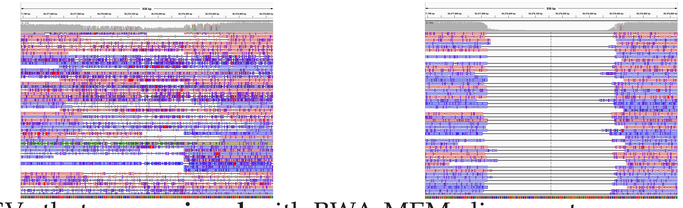
Realignment



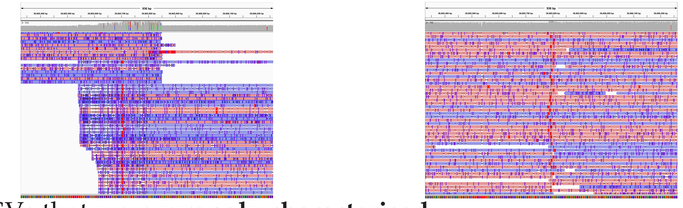
Currently we use *NextGenMap-LR* in combination with **Sniffles**. Sniffles is an extremely efficient structural variation caller developed for long third-generation sequencing reads. It is able to call structural variations as well as identify regions that most likely contain misaligned reads.

Results

To evaluate Sinffles and *NextGenMap-LR* we realigned all reads around regions identified by Sniffles (based on BWA-MEM alignments) on chromosome 8 and 17 and found:



(1) SVs that were **missed** with BWA-MEM alignments



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

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

Outlook

Currently we are working on applying *NextGenMap-LR* to the **full human genome** and to **Oxford Nanopore** data. The program will soon be available at www.cibiv.at/software/ngmlr