Detection of Structural Variants using third generation sequencing

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

Contact: fsedlaze@cshl.edu

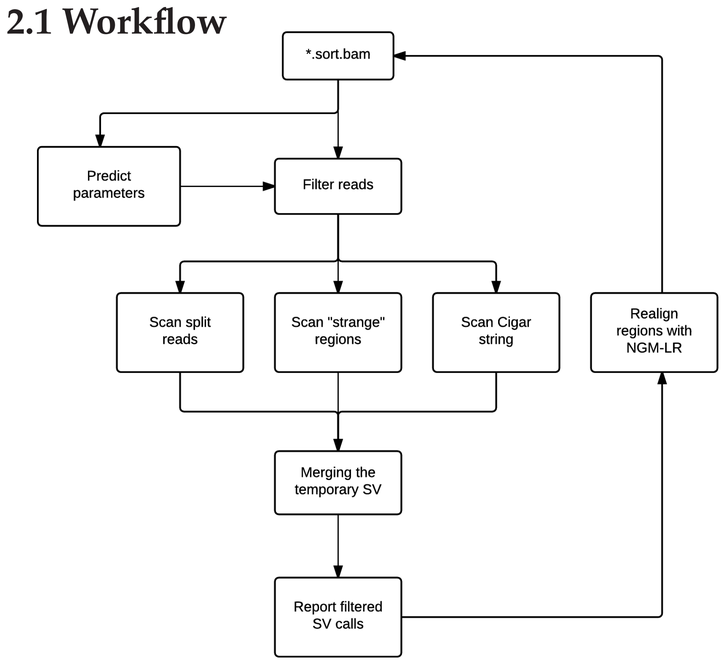
1: Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA;
2: Center for Integrative Bioinformatics Vienna, Vienna, Vienna, Austria;

1. Abstract

Structural Variants (SVs), which include deletions, insertions, duplications, inversions and chromosomal rearrangements, have been shown to effect organism phenotypes, including changing gene expression, increasing disease risk, and playing an important role in cancer development. Still **it remains challenging to detect all types** of SVs from high throughput sequencing data and it is even harder to **detect more complex SVs** such as a duplication nested within an inversion. To overcome these challenges **we developed algorithms for SV analysis** using longer third generation sequencing reads. The increased read lengths allow us to span more complex SVs and accurately assess SVs in repetitive regions, two of the major limitations when using short Illumina data. Our enhanced open-source analysis method **Sniffles accurately detects structural variants** based on split read mapping and assessment of the alignments. Sniffles uses a self-balancing interval tree in combination with a plane sweep algorithm to manage and assess the identified SVs. Central to its high accuracy is its **advanced scoring model** that can distinguish erroneous alignments from true breakpoints flanking SVs. In experiments with simulated and real genomes (e.g human breast cancer), we find that Sniffles outperforms all other SV analysis approaches in both the sensitivity of finding events as well as the specificity of those events. Sniffles is available at: **https://github.com/fritzsedlazeck/Sniffles**

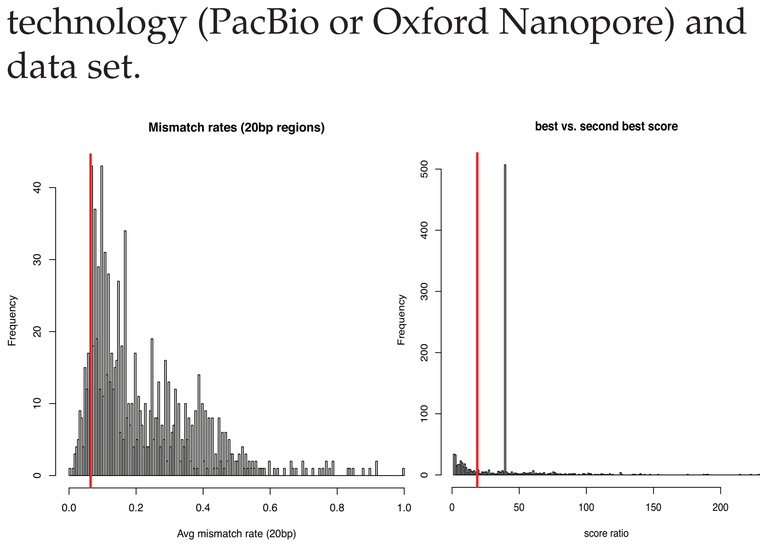
2. Sniffles

2.1 Workflow

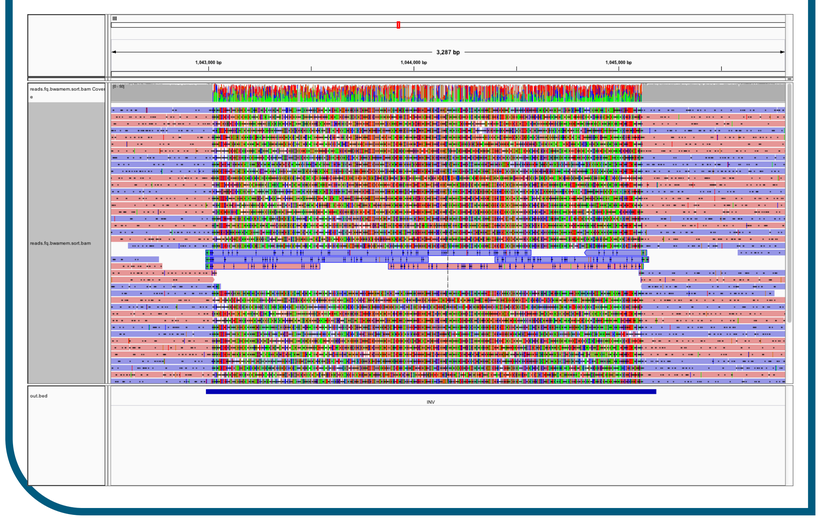


2.2 Automatic parameter estimation

Sniffles determines the parameters for each run automatically to adjust itself to the technology (PacBio or Oxford Nanopore) and data set.



This is done to distinguish between SV and alignment artifacts. Here BWA-MEM has aligned the reads through a 200bp inversion.



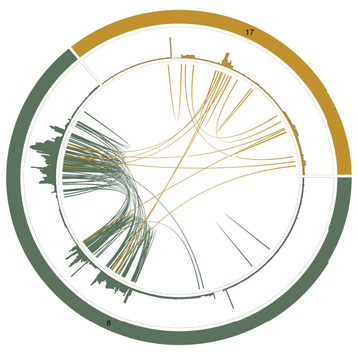
3. Showcase in SKBR3

We have sequenced this important Her2+ breast cancer cell line to 75x PacBio long read coverage with reads averaging 10kbp.

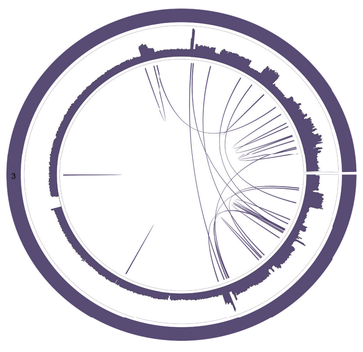
Lumpy:

Genome wide

Her2 region



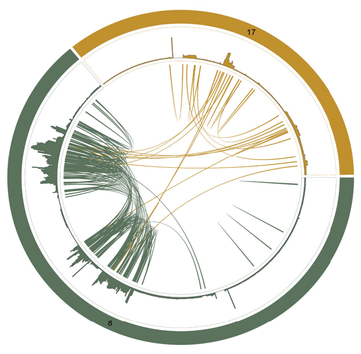
zoom in on Chr3



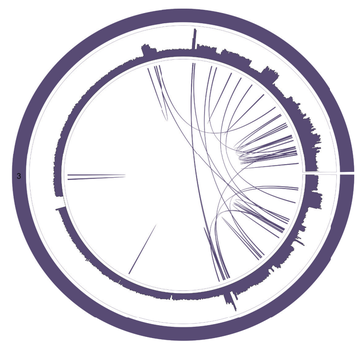
Sniffles:

Genome wide

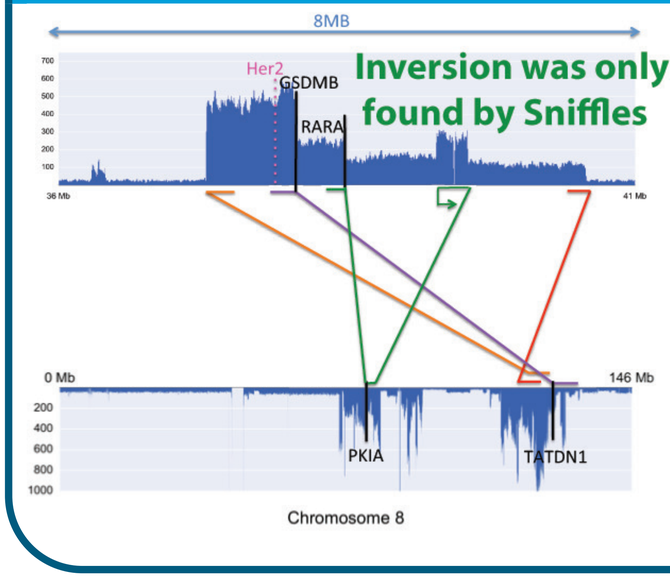
Her2 region



zoom in on Chr3



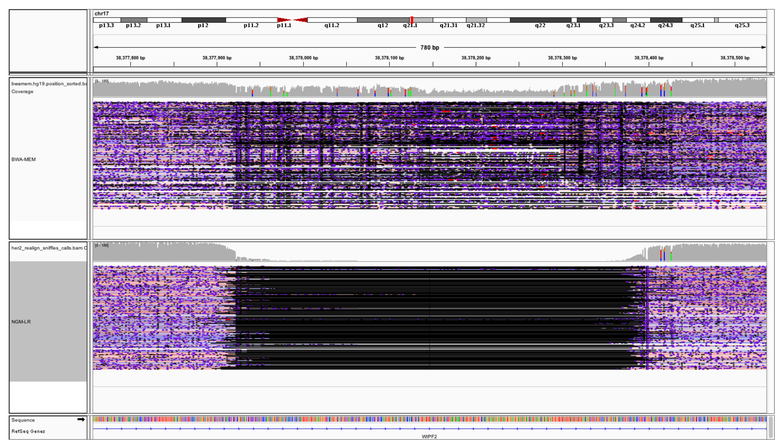
4. Her2 region



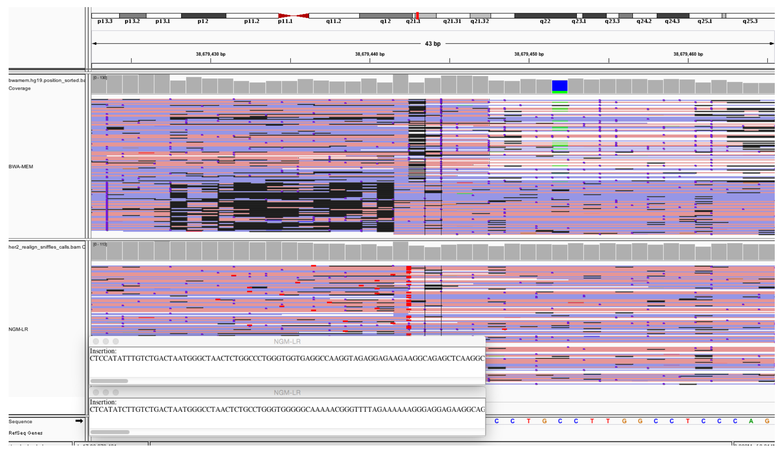
5. Realignment NextGenMap-LR

We have enhanced NextGenMap to become a long read mapper (NGM-LR) that is self aware of SVs. This is used in Sniffles to accurately detect the breakpoints.

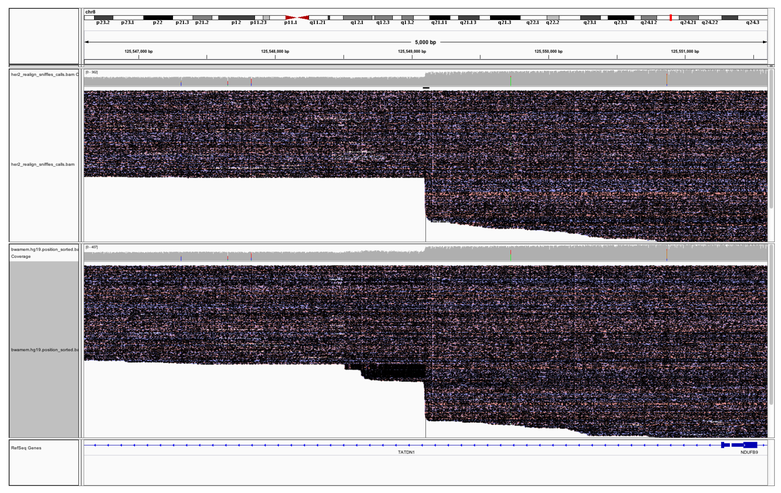
PacBio (Deletions):



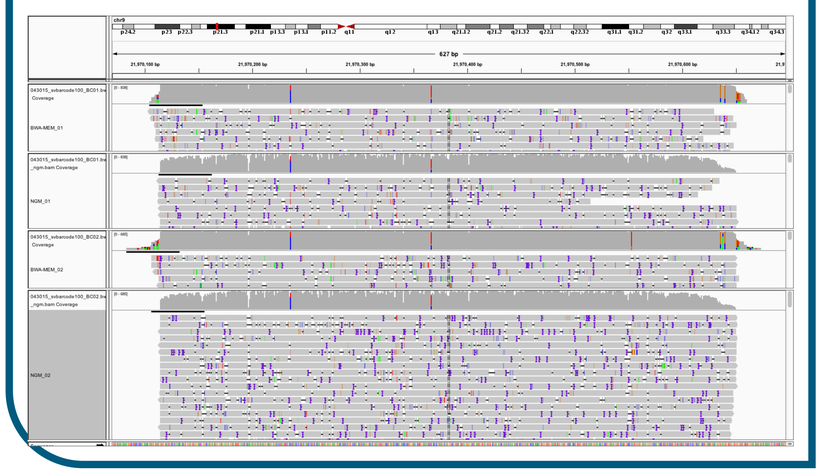
PacBio (Insertions):



PacBio (Translocations):



Oxford Nanopre (Translocation):



6. NextGenMap

NextGenMap is available at
http://cibiv.github.io/NextGenMap/

Sedlazeck et al., NextGenMap: fast and accurate read mapping in highly polymorphic genomes., Bioinformatics (2012)