GASOLINE: Germline And SOmatic structuraL variants detectioN and gEnotyping

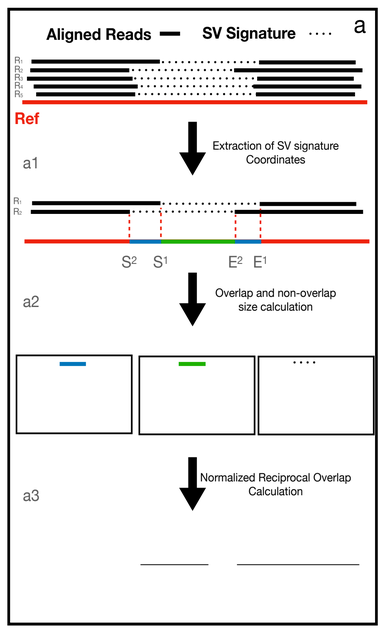
G.Mattei, R.Semeraro, A. Mingrino, C. Caprioli, E. Colombo, C. Ronchini, L. Mazzarella, P.G. Pelicci, A.Magi

ABSTRACT

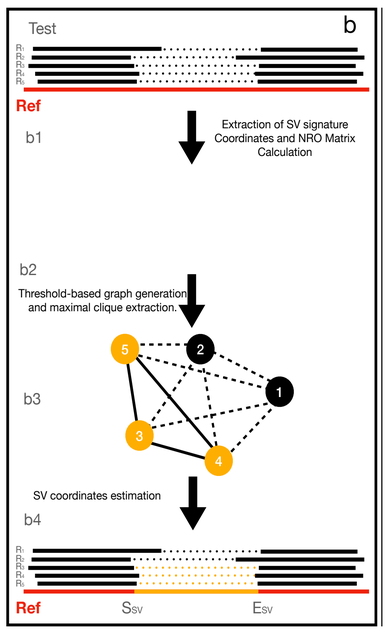
Identification of SVs from long-reads data requires complex computational methods, which are based on intra- and inter-alignment SV signatures (gapped/split- read alignments) approaches. These methods allow detection of deletions, inversions and translocations of any size, with insertions and duplications limited by read length. The SV signatures are clustered by reciprocal overlap of genomic coordinates, possibly leading to partial recovery and underestimation of allelic fraction, and genotyping errors. Furthermore, owing to high error rates, alignment of long reads can generate imprecise genomic coordinates of SV signatures with variances of tens of bp, which may prevent identification and clustering of SVs signatures generated by small events (50-500 bp). Large SVs (tens or hundreds of kb) are less affected by error rate, yet they need large reciprocal overlap to prevent inclusion of signatures from other events, possibly leading to the underestimation or loss of small SVs signatures. A further limit of most currently available computational tools is their poor flexibility for analyses of sample pairs. Since they are designed to detect germline variants from single samples, identification of somatic variants, for example, requires separate analyses of paired samples and discarding SVs from germline samples. To overcome these limits, we developed a novel tool, GASOLINE (Germline And SOmatic structuraL varIants detectioN and gEnotyping), which groups SV signatures using a sophisticated clustering procedure based on a modified reciprocal overlap criterion (Normalized Reciprocal Overlap, NRO) and allows detection of both large and small SVs with high accuracy. We extensively tested the new tool on simulated and real cancer datasets and we demonstrated that it outperforms NanomonSV in the detection of small and large somatic variants. Notably, when applied on COLO829 cell line and matched normal sample, GASOLINE identified 6 genuine somatic SVs that were missed by Valle-Inclan et al. by using five different sequencing technologies and state-of-the art SV calling approaches.

METHODS

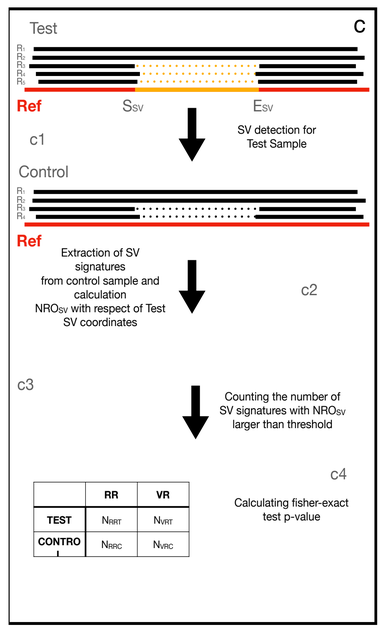
ALGORITHM



For each pair of reads with ends near the related SV signature, the coordinates from both SV ends related to the signature (S1, E1 and S2, E2) are extracted. Next, for the considered pair of reads, are scored the interval range of both ends of the signature (NO112 = S2 - S1 ; NO212 = E2 - E1), the range of the SV (O12 = S2 - E1) and the potential size of signatures (L1 = S1 - E1 ; L2 = S2 - E2). These values are necessary in order to calculate the normalized reciprocal overlap (NRO12) that allows to group both small and large SV signatures with high accuracy reducing the effect of imprecise alignment by taking into account both overlapping and non-overlapping regions (Fig a). This process is reiterated for each pair of reads.



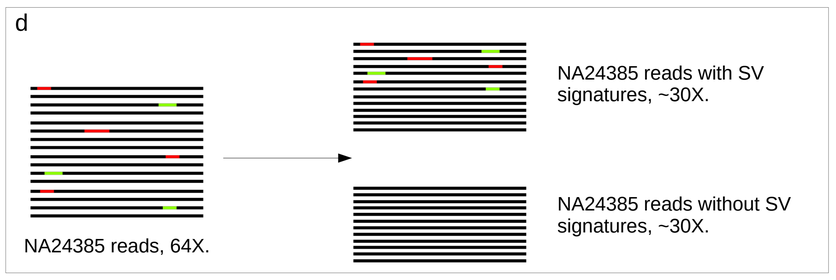
Once NROij has been calculated for all the signature pairs, values are used to create the NRO adjacency matrix to cluster intervals by using a graph-based approach. Nodes are the SV signatures and the edges between two signatures exist if NROij > Thr, where Thr is a predefined reciprocal overlap threshold depending on the SV length. The edge expresses the confidence that two signatures are generated by the same SV event. The undirected graph is then used to extract maximal cliques (groups of fully connected nodes) by using the Eppstein-Löffler-Strash algorithm (Eppstein et al., 2010). The coordinates for the SV are then extracted from the clique by calculating the median of all start and end coordinates (Fig b). Since more SVs may fall within the same range, more cliques are extracted according to the maximum number of links. We then calculate different statistics to filter out low quality signatures that include: cohesion score (the ratio between the number of links in the extended clique and the maximum number of link), mean mode and standard deviation of start and end coordinates



Somatic signatures are then detected by comparing the NROSV of each cluster of signatures with those from the control sample (NROsv = NROsv-test - NROsv-ctrl). Signatures with a NROSV larger than a predefined threshold are considered somatic. The statistical significance of each somatic SV is then calculated by applying the fisher exact test using for the contingency table the NSRT / NRRT (number of reads from the control with/without SV signatures), the NSRC / NRRC (number of reads from the control with/without SV signatures) (Fig c).

BENCHMARK

Due to the lack of high-quality gold standard datasets that enable the benchmarking, we simulated somatic SVs of different sizes by using the 64x nanopore WGS data generated by the GIAB consortium for the NA24385 sample. We randomly splitted the NA24385's reads in two bam files among which only one containing SV signatures: 330 heterozygous SVs (169 INS and 161 DEL) between 50bp and 5kb, obtaining a 30x bam file with SV and a 30x control bam file. GASOLINE was tested for different values of NRO thresholds (0.5, 0.6, 0.7, 0.8, 0.9) and NONorm (000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000). We found that the best combination is Thr = 0.5 and NONorm = 5000 for both minimap2 and NGLMR aligned data.



RESULTS

The results reported in Fig e and Fig f clearly show that GASOLINE outperforms the performance of NanomonSV for data aligned with both minimap2 and NGMLR and for both small (50-500bp, Fig e) and large (> 500bp, Fig f) SVs, especially for INS, where NanomonSV completely failed the identification of this class of variants. These analyses also demonstrate that filtering out SVs on the basis of somatic p-values allows to drastically increase precision, by removing a large fraction of false positive calls, at the expense of a minimal decrease in recall (removal of true positive calls): the best trade-off between precision and recall for both deletions and insertions is reached with p-value=5x10−3 p-value=1x10−3.

To test the capability of our tool in detecting all SVs subtypes we exploited the nanopore data generated by Shiraishi et al. for the COLO829 cell lines and we compared its performance with those of NanomonSV by using the Valle-Inclan et al. true-set SVs as benchmark. With the exception of deletions with minimap2 alignment our tool outperformed NanomonSV in the detection of all SV subtypes. Remarkably, the few deletions missed by our method are supported by a small number of signatures and although detected by GASOLINE they do not reach a statistically significant p-value, and are filtered out from somatic validations.