IMPROVING VARIANT ACCURACY WITH COPY NUMBER VARIANT PIPELINE FOR TARGET SEQUENCING

George Carvalho[a], Wilder Galvão[a], Marcel Caraciolo[a], Rodrigo Alexandre[a], João B Oliveira[a]

[a] Genomika Diagnósticos, Recife-PE

INTRODUCTION AND MOTIVATION

Studies comparing human genomes have been shown that more base pairs are altered as a result of structural variants (SVs), including copy number variants (CNVs), than as result of point mutations. Structural variants were first defined as insertions, deletions and inversions greater than 1 kb size. Nevertheless, with the high-throughput sequencing becoming a routine for genome analysis, the spectrum size of SVs and CNVs have been extended to events >50 bp in length [1].

Due to the cost and the complexity of analyzing the amount of data generated by whole-genome sequencing (WGS), targeted capture sequencing including custom gene panels and whole-exome sequencing (WES) has become the predominant approach for genetic diagnostic purposes [2]. These techniques have demonstrated huge power to identify rare and private single-nucleotide variants (SNVs) and small insertions and deletions (INDELs) for genetic diagnostic purposes, however large deletions and duplications are ignored in most of the cases because copy number variation (CNV) identification from target sequencing data is less confident compared to WGS, array comparative genomic hybridization (aCGH) and the multiple ligation probe assay (MLPA) [2, 3].

For next-generation sequencing (NGS) analysis replace theses other methods, a software or a combination of tools for exon CNV detection with high sensitivity, specificity and acceptable false discovery rate is required [4]. With efforts conducted to develop an internal protocol for CNVs detection in TS data, and increase possible genetic case elucidation, the aim of this study was to analyze the performance of state-of-the-art CNV detection methods on targeted next-generation sequencing (NGS) before implementation at our laboratory.

METHODS

High coverage targeted NGS data was generated in Genomika Diagnostics using Qiagen V3 kit. We applied CNV detection algorithms to validate the sensitivity for known CNV events identified by MLPA method. Five patients positive for deletion in BRCA1 gene were selected to test four softwares: ExomeDepth, panelcn.MOPS, CoNVaDING and VisCap. This allowed calculation of the sensitivity and specificity for identified deletions.

Quality control were checked using FastQC on the raw sequence data from Illumina Miseq platform. Using python scripts from Qiagen, the FASTQ files were aligned on hg19 genome reference using BWA-MEM and PCR duplicate reads were removed using Picard MarkDuplicates, after that, Unique Molecular Identifier (UMI) sequences are extracted from BAM file for the variant calling. CoNVaDING uses BAM files as input, but is necessary a minimum set of thirty possible control samples from which the samples with the most similar overall pattern are selected as control samples.

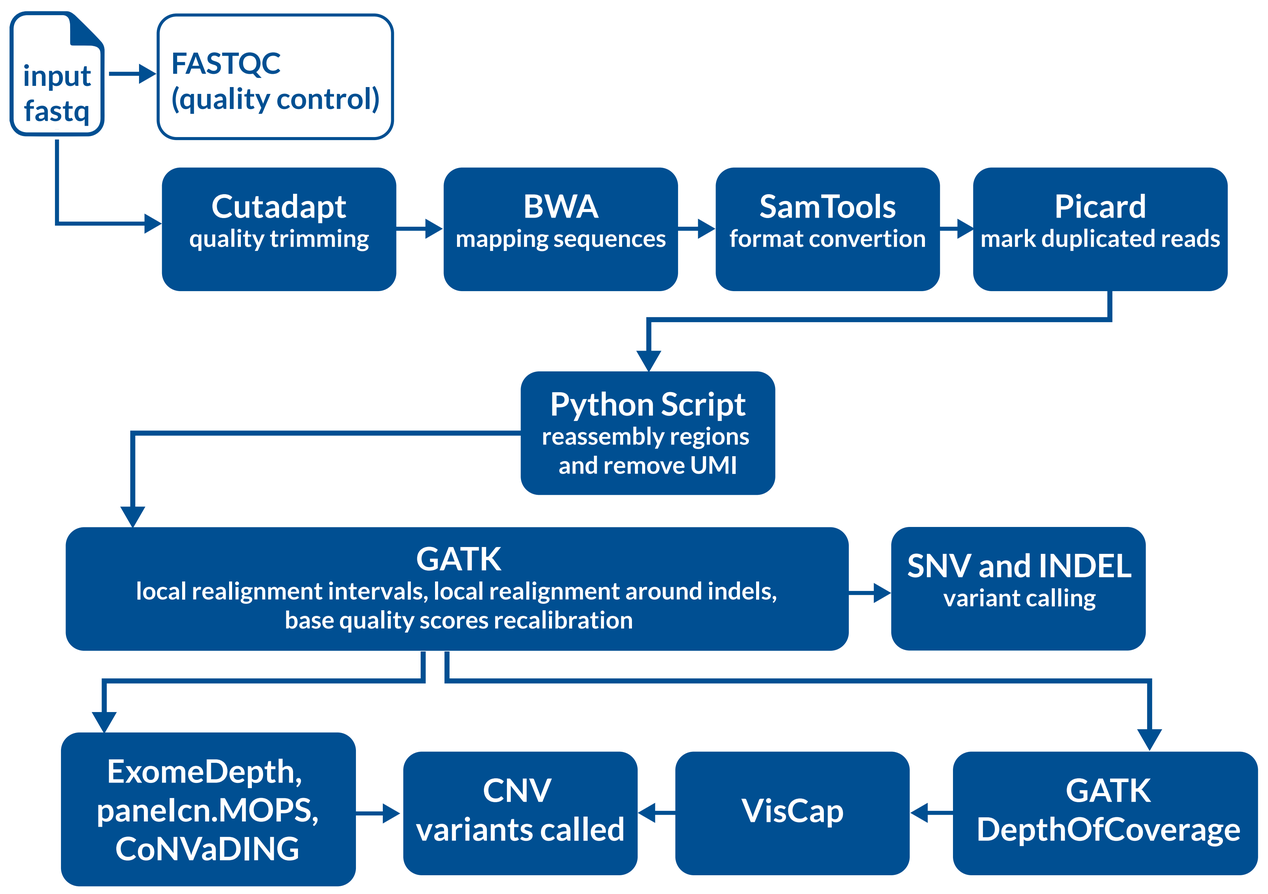

Figure 2: Pipeline workflow for target sequencing variant detection

BENCHMARK

Tool	TP	TP	TP	TP	Accuracy	Precision	Sensitivity	Specificity
CoNVaDING	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ%	XYZ
VisCap	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ%	XYZ
ExomeDepth	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ%	XYZ
panelcn.MOPS	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ	XYZ%	XYZ

Table 1: Performance evaluation of the tools in this study, we used a collection of statistics derived from a confusion matrix. We counted true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). With this data we could calculate accuracy, precision, sensitivity and specificity.

Figure 1: Classes of CNV that can be identified using read depth based softwares available for target sequencing method.

REFERENCES

1. Alkan, Can, Bradley P. Coe, and Evan E. Eichler. "Genome structural variation discovery and genotyping". Nature reviews. Genetics 12.5 (2011): 363.

2. Chen, Yong, et al. "SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data". BMC bioinformatics 18.1 (2017): 147.

3. Ellingford, Jamie M., et al. "Validation of copy number variation analysis for next-generation sequencing diagnostics". European Journal of Human Genetics 25.6 (2017): 719-724.

4. Fowler, Anna, et al. "Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN." Wellcome open research 1 (2016).

CONCLUSION

With the unavailability of well-known negative samples wasn't possible calculate statistics informations as accuracy, sensitivity and specificity. Thereby, the plan is enhance the number of samples used to the benchmark and added confirmed examples of large duplications. However, given the results of precision provided by our results, softwares as CoNVaDING can be added to the bioinformatics analysis routine as first step of CNV detection, directing for a confirmation method.