CNVfinder: a Python tool for copy number variation detection on exome sequencing data from amplicon-based enrichment technologies

Valengo AE, Rosati R
Instituto de Pesquisa Pelé Pequeno Príncipe - Faculdades Pequeno Príncipe
valengoandressa@gmail.com

Fork me on GitHub

CNV: meaning and detection

Copy Number Variations (CNV) have been identified in a wide range of species and are source of genetic variation in humans. Even though most common CNVs are not associated to medically relevant traits, some structural variations can also play a role in common diseases. In addition, copy number alterations (CNA) appearing as somatic aberrations in neoplasms can be associated with crucial tumor characteristics. Although DNA microarray technology is currently the most refined approach to identify CNVs, whole exome sequencing (WES) has increasingly emerged as a front line diagnostic platform for genotype-phenotype associations. This has led to an interest in leveraging WES data to also provide CNV information.

Motivation and goals

Most software available for CNV detection is focused on data obtained by hybridization capture exome enrichment methods (most notably, from Illumina platforms). However, the same **tools may not be as successful in processing data from amplicon-based enrichment methods**, such as the AmpliSeq technology used in Ion Torrent^{TM} sequencers. In addition, there is a **lack of** CNV/CNA detection tools with **user-friendly interfaces** and **advanced visualization features**.

The purpose of this project was to develop a CNV/CNA detection tool specifically tailored on WES data obtained through amplicon-based enrichment protocols.

Our solution

We developed **CNVfinder**, which was implemented using **Python 3.x** applying **Object Oriented Programming**. CNVfinder employs the read depth of amplicons in order to detect gains or losses of number of copies in regions. CNVfinder reads Binary Alignment Map (BAM) files and analyzes read depth data for each amplicon in order to detect genomic CNVs/CNAs. Amplicon data are loaded from the relevant Browser Extensible Format (BED) file and processed into a series of unique, non-overlapping genome targets (Figure 1). CNVfinder can also apply information of the allele frequencies of variants called by the sequencing workflow, allowing for more accurate results.

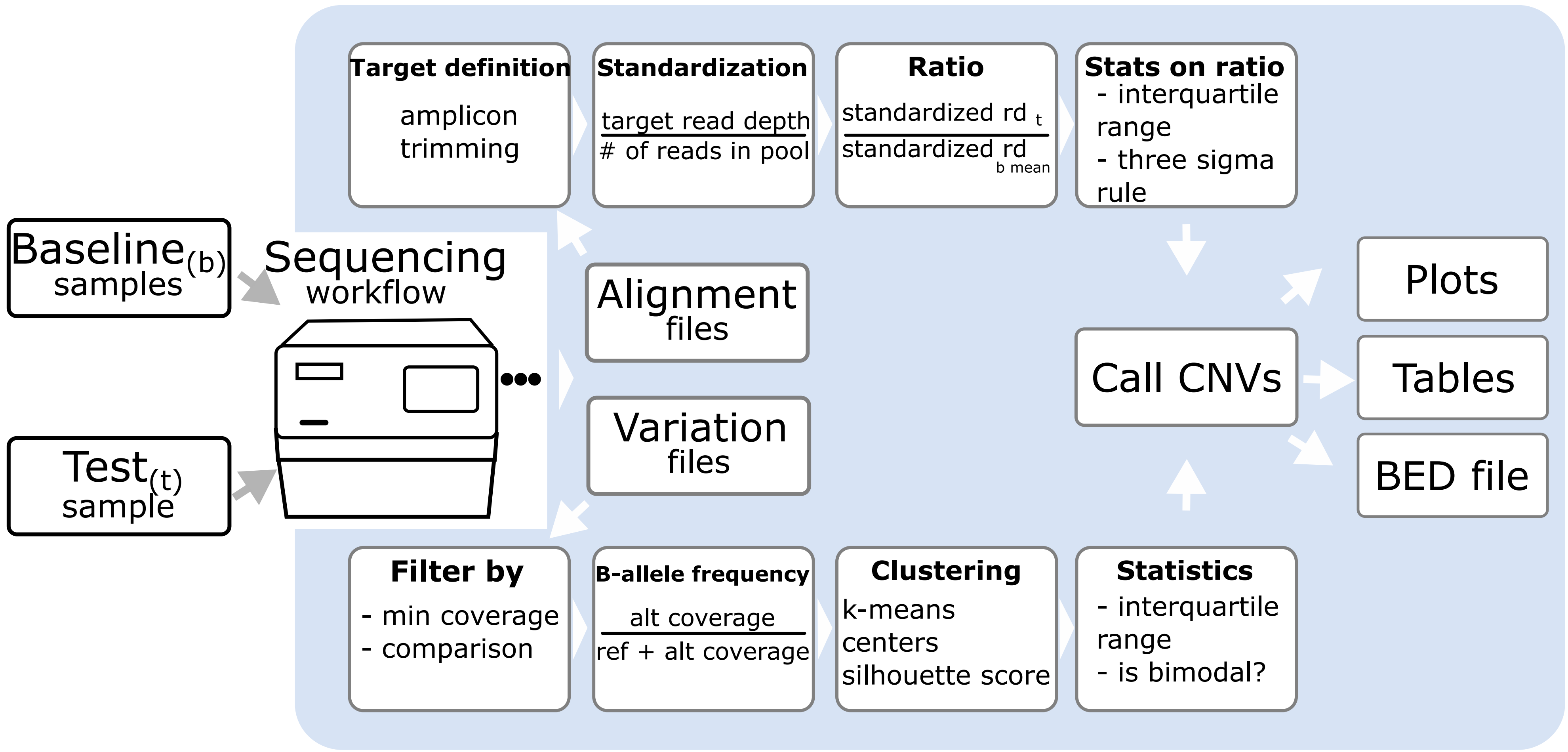

Figure 1: An overview of CNVfinder approach for CNV detection

CNVfinder: features and availability

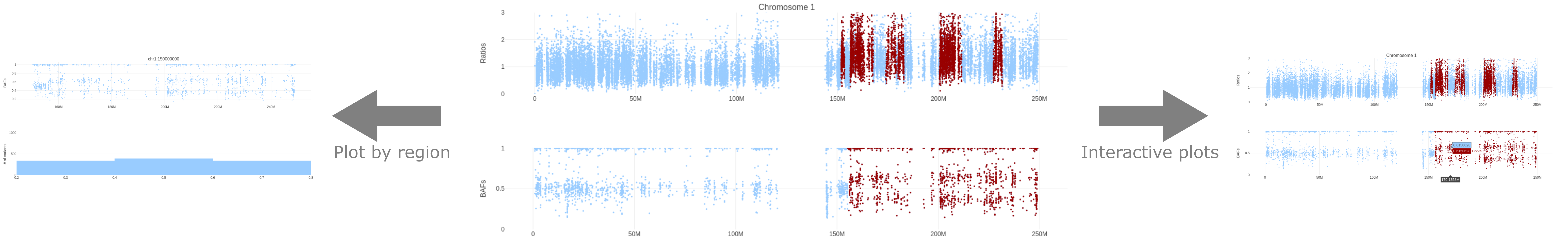

Figure 2: Example of plots created by CNVfinder

Figure 2 shows examples of plots created by CNVfinder as part of the output for a given set of baseline and test samples. The middle plot represents potential CNVs (red colored). Plots generated by CNVfinder are interactive (zoom, show values, save as PNG), allowing for a **better visualization** of the results (right plot). In addition, it is possible to **limit the plot to certain genomic regions** (left plot), in order to best understand, for example, how the data (ratios and/or B-allele frequencies) are distributed.

CNVfinder is **structured as a Python package, allowing developers to incorporate our code in their own CNV detection solutions**. Furthermore, it can be **easily installed** with PIP (Python's package management system). CNVfinder is available under the **MIT License** and its **source code is hosted at GitHub** (https://github.com/ip4-team/cnvfinder).

Installation command: pip3 install -U cnvfinder

```
pip3 install -U cnvfinder
```