MGvizCNV: a QC Machine Learning approach for CNV evidence scoring

Pedro Pons-Suñer[1,2], Pablo Marin-Garcia[1,2], David Gómez-Peregrina[2,3], Anastasiya Onofriychuk[1,2], Pablo Cano[2], Rodiel Martinez-Jimenez[1,2], Ana Barbara Garcia-Garcia[4,5] and Jose M. Juanes[1,2]

1. Kanteron Systems, Valencia (Spain); 2. MGviz.org, Cambridge (UK); 3. Sarcoma Translational Research Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona (Spain); 4. CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Madrid (Spain); 5 Genómica y Diabetes, INCLIVA, valencia (Spain)

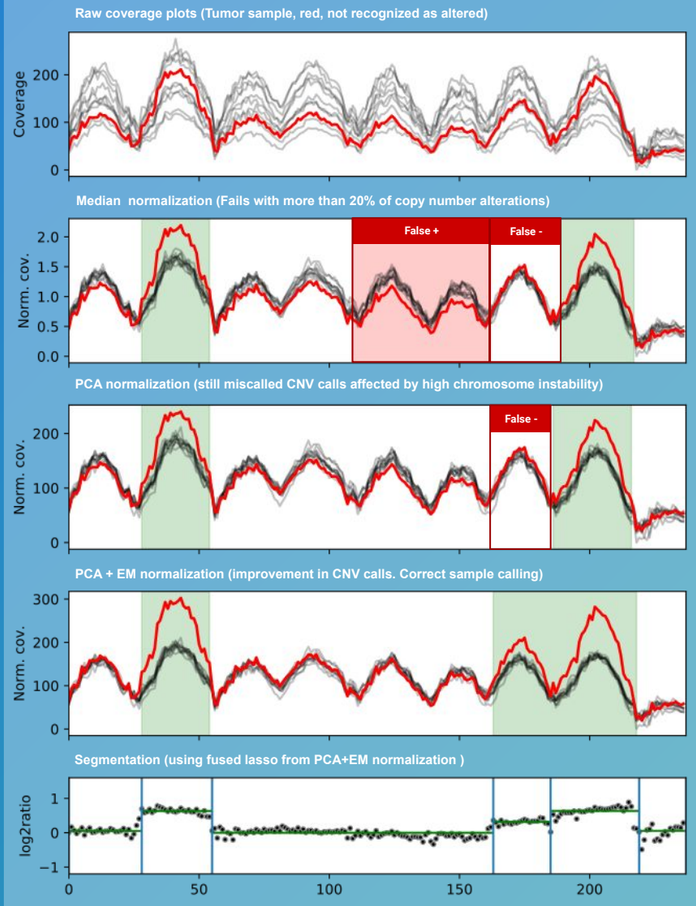
kanteron systems

MGviz
www.mgviz.org

Introduction

Copy Number Variants (**CNV**) are differentiating events between individual genomes and play an important role in cancer and mendelian diseases like neurodegenerative diseases. The current CNV detection methodologies using Next Generation Sequencing (NGS) are based mainly on measuring the relative losses or gains of the reads coverage for a genome region. However, read counts usually present various types of random noise and biases, which need to be mitigated in order to accurately analyze copy numbers. Currently there is no consensus yet on methodologies or thresholds to use for accurately calling CNVs, resulting in many different tools calling too many CNV candidates with little concordance between their results.

CNV detection

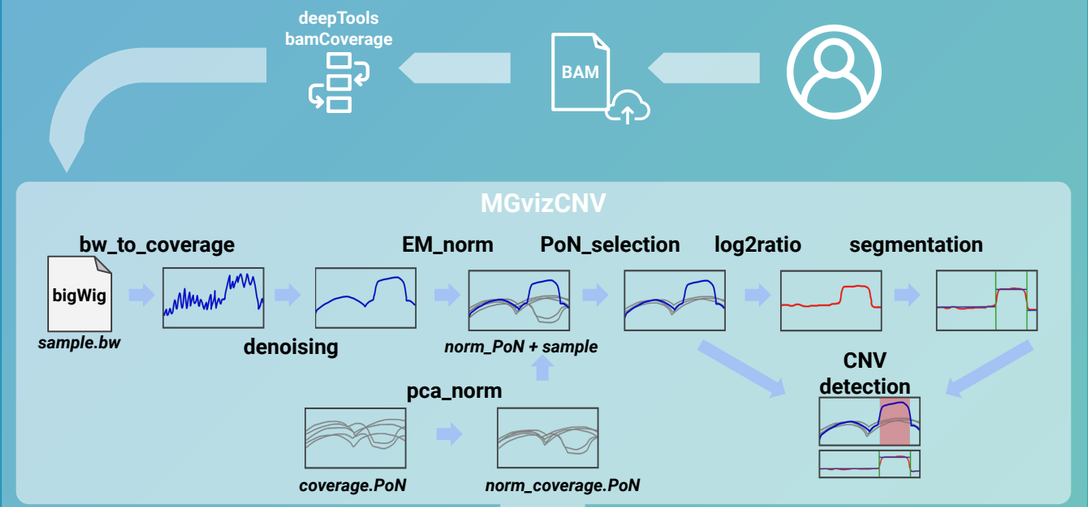


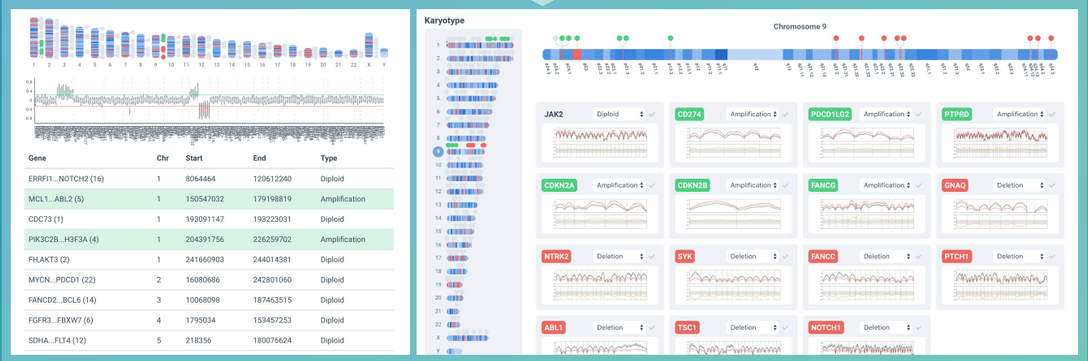
Methods

Naive normalization methods such us median normalization are not appropriated for cancer samples with high chromosomal instability.

MGvizCNV project has evaluate different signal processing techniques (including novel Machine Learning methods) for data normalization and denoising. This work employs several techniques such as wavelet shrinkage denoising and total variation denoising to smoothen readcount data while preserving high frequency information such as breakpoints. Also, a principal component analysis and expectation-maximization based technique is used to reduce undesired biases between samples. The log2ratio is then segmented via CBS and fused lasso techniques.

The images on the left represent a tumor sample (red) and a population of controls (black) samples for reference. Different normalization methods has been tested, being the PCA+EM Normalization and fused lasso segmentation the more accurate. See docs at **github** repo.





MGvizCNV can be integrated in explorer/Advisor views on interactive visualization tools

Conclusion

We have created a workflow for CNV assessment and implemented a web application for exploring and annotating the complexity of CNV detection. This work features a machine learning layer for denoising, normalization and segmentation that allow accurate CNV detection.

Also can clusterize a given sample with similar individuals in a reference population and detect CNVs as anomalies that deviate from the assigned population, following a statistical approach that allows the user to select a desired sensitivity level for detection.

Code an docs at **https://github.com/MGvizPro/MGvizCNV**