Minimum Error Calibration and Normalization for Genomic Copy Number Analysis
Bo Gao, Michael Baudis
Institute of Molecular Life Sciences, University of Zürich, Switzerland

Introduction
In this study, we present a novel method named Minimum Error Calibration and Normalization of Copy Numbers Analysis (Mecan4CNA). In general, Mecan4CNA provides an advanced method for CNA data normalization especially in research involving data of high volume and heterogeneous quality. With its informative output and visualization, it can also facilitate analysis of individual CNA profiles. Mecan4CNA is freely available as a Python package and through Github.

Background
Copy number variations (CNV) are regional deviations from the normal autosomal bi-allelic DNA content. While germline CNVs are a major contributor to genomic syndromes and inherited diseases, the majority of cancers accumulate extensive "somatic" CNV (sCNV or CNA) during the process of oncogenic transformation and progression. While specific sCNV have closely been associated with tumorigenesis, intriguingly many neoplasias exhibit recurrent sCNV patterns beyond the involvement of a few cancer driver genes. Currently, CNV profiles of tumor samples are generated using genomic micro-arrays or high-throughput DNA sequencing. Regardless of the underlying technology, genomic copy number data is derived from the relative assessment and integration of multiple signals, with the data generation process being prone to contamination from several sources. Estimated copy number values have no absolute and linear correlation to their corresponding DNA levels, and the extent of deviation differs between sample profiles which poses a great challenge for data integration and comparison in large scale genome analysis.

Performance evaluation
NCI-60 cancer cell line data:
Comparison of baseline and level distance estimation with existing methods and karyotyping data on the NCI-60 tumor cell line produced coherent results.

Applications
Baseline deviation of TCGA data:
The generally good quality of these datasets allowed us to show that baseline variation is a common and recurring problem among copy number data.

Normalized data as GISTIC input:
To estimate the method’s impact on downstream analyses we performed GISTIC analyses on original and normalized (using Mecan4CNA) data from the Cancer Genome Atlas (TCGA). The normalized data showed prominent improvements of both sensitivity and specificity in detecting focal regions.

Availability
1. Python installation: pip install mecan4cna
2. github: https://github.com/baudisgroup/mecan4cna
3. Software license: MIT
4. Contact: b.gao@imls.uzh.ch