Genome analysis

TCGA.CNV.BaselineCorrection (TCBC): Interactive Web-Application for Baseline Correction, Plotting, and Quality Control of Copy Number Data from Cancer Samples.

Nour-al-dain Marzouka[1], Jonas Hagberg[2], Ann-Christine Syvänen[1]

[1] Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.

[2] National Bioinformatics Infrastructure Sweden (NBIS), Sweden.

ABSTRACT

Copy number (CN) variants play important roles in the genetics of many diseases, especially in cancer. Detection of CN variants can be performed using data from different technologies and platforms, and large number of tools and algorithms are available for this purpose.

One of the important challenges in CN calling, regardless of the source of the data, is to determine the correct baseline (i.e. normal chromosomes level) in the samples. As a general rule, the more accurate the baseline is, the more accurate the calling of duplications/deletions will be. Numerous large-scale chromosomal aberrations in cancer samples cause baseline shifting in CN data, which is a known problem that may result in erroneous CN calling. Consequently, a baseline correction process is needed before any further analysis is performed.

The Cancer Genome Atlas (TCGA) is one of the largest databases for CN data from cancer cells, and it is widely used in cancer research. However, TCGA provides the CN data without baseline correction.

Previously, we developed an R package CopyNumber450kCancer (Marzouka et. al., 2015) to correct the baseline and to re-center cancer CN data. Here, we provide a web-application with a user-friendly interface for baseline correction, quality control (QC), and interactive visualization of CN data from TCGA or other sources.

AVAILABILITY AND IMPLEMENTATION:

TCGA.CNV.BaselineCorrection (TCBC) is freely available and includes example data and CN data from TCGA at: https://copynumber.shinyapps.io/RTCGA-CNV-BaselineCorrection/ and on GitHub: https://github.com/NourMarzouka/RTCGA-CNV-Baseline-Correction

Contact: nour.marzouka@medsci.uu.se

1 INTRODUCTION

The Cancer Genome Atlas (TCGA) database is an important source for copy number (CN) data, and is widely used by researchers in the cancer field. TCGA provides 3 different levels of CN data categorized based on the level of data processing. Researchers in cancer are usually interested in the processed data (i.e. level 3) rather than the raw data. Level 3 data represents segmented data where the signals from probes are grouped based on their locations and their signal levels to form large contiguous regions. Using segmented data for cancer samples requires sample re-centering (i.e. baseline shifting), plotting, and quality control before downstream analysis.

Conventionally, segmentation data is centered on the median or the mean of the log values for the segments (Mermel et al., 2011) which causes a significant shift in the center (i.e. baseline) due to the large chromosomal alterations in cancer samples (Lipson *et al*., 2007). Previously, we showed that re-centering cancer samples on the supposed baseline (i.e. normal chromosomes level) reduces the false positives significantly and avoids the cancer-specific problem of erroneous CN calling (Marzouka et al., 2015). After the baseline correction the segments with different signal levels can be interpreted as normal, losses or gains.

To date, there is no open source, freely available tool to systemically resolve this particular problem in TCGA CN data that might also be present in other similar datasets.

2 DESCRIPTION

To resolve the baseline shifting problem in TCGA data and to provide additional QC measurements, we provide a web-application with a user-friendly interface denoted **T**CGA.**C**NV.**B**aseline**C**orrection (briefly TCBC) which can be freely used on-line or locally on all operating systems with installed R (version > 3.0). TCBC provides a novel functionality for QC assessment and interactive plots for baseline correction in CN data.

3 BASELINE ESTIMATION

For the baseline correction step, TCBC uses an algorithm from the R package CopyNumber450kCancer (Marzouka *et. al*. 2015). Briefly, the correct baseline is estimated at the maximum density peak for the segments. This method was tested on CN data from 764 acute lymphoblastic leukemia samples and showed >95% accuracy rate (Marzouka *et. al*. 2015). TCBC allows the user to interactively adjust the proper baseline in case the auto-correction function selects a suboptimal baseline.

4 TCGA INPUT DATA

Instead of the direct access to the TCGA database, TCBC uses TCGA data from the Bioconductor.org R package RTCGA.CNV (Biecek P (2015)) which contains CN data (level 3) from TCGA servers. RTCGA.CNV package is frequently updated and contains the different cancer types from the TCGA database.

5 USER INPUT DATA

TCBC uses a tabular input data structure. It requires one file that contains segmented genomic regions for all samples with intensity log values and number of probes in each segment.

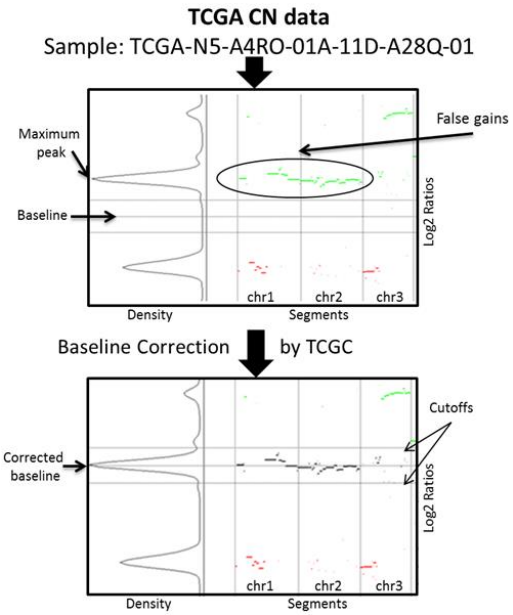

Figure 1: The upper plot shows part of the segmentation data for a cancer sample from TCGA. Red color represents deletions and amplifications are shown in green. The horizontal lines represent the cut-offs and the zero level (i.e. baseline). The lower plot shows the sample after the baseline correction.

As an additional optional input file, a sample list with comments can be uploaded. The comments (e.g. karyotypes) will be displayed on the generated plots, and this will help the user to review the baseline correction.

6 QUALITY CONTROL

CN data quality varies between samples. QC measurements help the researchers to assess the quality of the data and to exclude samples of poor quality. TCBC provides 8 different QC measurements which can be used regardless of the source of the data or the technology that was used to produce the data, in addition to the visual revision of the sample plots.

The QC measurements includes the number of the segments, Standard Deviation (SD), InterQuartile Range (IQR), Median Absolute Pairwise Difference (MAPD), the highest segments density peak sharpness, the area under the curve between the cut-offs, and finally the heights (i.e. values) of the density function at the upper and lower cut-offs. Only the last three measurements are sensitive to the baseline position (i.e. sample center) and the cut-offs. Lower values of the density function at the cut-offs indicate fewer segments located near the cut-off and lower possibility to have segments with log values higher than the cut-off by chance or due to the noise in the data.

7 OUTPUT

TCBC generates 4 output files. 1) Segmentation file, which is similar to the input segmentation file but with shifted log intensity values based on the new baselines in the samples. 2) Corrected segmentation plots. 3) QC measurements. 4) Record for all the shifts that were applied on the samples by the user or by the tool.

The corrected CN data can be downloaded as tables and plots. To facilitate the tracking of the modifications in the data, TCBC provides a file that contains the shift value for each sample that was performed by TCBC and the user.

8 INTERACTIVE REVISION

TCBC produces plots where the user can interactively review and change the baseline estimation. User comments (*e.g.* karyotyping) can also be displayed on the plots to facilitate the reviewing step and help the user to decide the correct baseline.

9 CONCLUSION

Cancer CN data suffers from incorrect sample centering, which is mainly due to the presence of large chromosomal aberrations and the use of the median/mean of the segments as the center for the sample. This problem affects the accuracy of the downstream analysis significantly. We developed a user-friendly web-application to correct the baseline in CN data, and to perform QC for the samples.

ACKNOWLEDGEMENTS

Support by NBIS (National Bioinformatics Infrastructure Sweden) is gratefully acknowledged.

FUNDING

This work was supported by The Swedish Cancer Society [15 0353] and the Swedish Research Council for Science and Technology [C0524801]

Conflict of Interest: none declared.

REFERENCES

Biecek P (2015). RTCGA.CNV: CNV (Copy-number variation) datasets from The Cancer Genome Atlas Project. R package version 1.0.2.

Lipson D. (2007) Determining the center of array-CGH data. In: Computational Aspects of DNA Copy Number Measurement. Technion – Israel Institute of Technology, Computer Science Department, pp. 105–110.

Marzouka, N., Nordlund, J., Bäcklin, C.L., Lönnerholm, G., Syvänen, A.-C., and Almlöf, J.C. (2015). CopyNumber450kCancer: Baseline Correction for Accurate Copy Number Calling from the 450k Methylation Array. Bioinformatics.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.