CGARS: Cancer Genome Analysis by Rank Sums

--- User's Manual ---

Abstract:

Cancer genomes are characterized by the accumulation of point mutations and structural alterations such as copy number alterations and genomic rearrangements. Among structural changes, systematic analyses of copy number alterations have provided deeper insight into the architecture of cancer genomes and had led to new potential treatment opportunities. During the course of cancer genome evolution, selection mechanisms are leading to a non-random pattern of mutational events contributing to fitness benefits of the cancer cells. CGARS is designed to dissect random from non-random patterns in copy number data and thereby to assess significantly enriched somatic copy number aberrations across a set of tumor specimens or cell lines. In contrast to existing approaches, the method is invariant to any strictly monotonous transformation of the input data, which results to an insensitivity of differences in tumor purity, array saturation effects, and copy number baseline levels.

Installation:

CGARS runs under Linux and Mac OS X; Windows is currently not supported.

1. Please download the tar-archive: *"cgars.tgz"* from http://www.uni-koeln.de/med-fak/cgars/ and decompress the archive by executing *"tar xvzf cgars.tgz"*.
2. Enter the directory: *"cgars"* and run: *"sh install.sh"*.
3. Run: *"make"* to build the executable "*./cgars"*.

Running CGARS:

CGARS can be run under two different modes: 1) to perform the core analysis and 2) to visualize detected peaks. Please note that the visualization module requires a fully operable installation of R (http://cran.r-project.org).

SYNOPSIS:

cgars [command] <options>

COMMANDS:

analyze	analyze copy number data
plot	plot copy number aberrations (requires R)

The "analyze" module:

SYNOPSIS:

cgars analyze [options] <name for output files>

DESCRIPTION:

-h -? -help	help
-cn	filename of raw copy number data
-m	rank matrix file
-s	name of segment file
-b	genome build (hg18, hg17, mm9) [hg18]
-nocnv	represses removal of CNV prone probes
-l	file to exclude regions form analysis
-mwd	minimal number of probes for a segment [20]
-alpha	smoothing parameter for segmentation [0.00001]
-uq	upper quantile for amplifications (list, e.g, 0.15,0.1,...) [0.15]
-lq	lower quantile for deletions (list, e.g, 0.15,0.1,...) [0.15]
-alfq	q-value cutoff SCNA filter [0.25]
-keepXY	keep X and Y chromosomes

The name of the output files (prefix) is usually after the options. For sake of simplicity we refer to this name in the following as: *$out_name*.

Options:

-cn This option requires the name of the file holding the raw copy number data. Details about the format of the input data: please see *Input Data* section.

-m During the first analysis of each data set CGARS generates the rank matrix and stores it under *$out_name.mat*. In a subsequent run using the same data set, the generated rank matrix can be used directly (instead of recalculating it) to save time and memory. In this case the *"-cn"* option is omitted and instead *"-m $out_name.mat"* is used.

-s CGARS generates a file: *$out_name.seg,* which holds information about the smoothing (segments). To enforce the method to use the smoothing of a previous run: *"-s $out_name.seg"*. This option might be useful to compare results directly, e.g., cell lines vs. primary tumors.

-b This option specifies the genome build underlying the raw copy number file. Currently implemented is for human: hg18 and hg19 and for mouse: mm9. By default this option is set to hg18. Note that the most current human genome build (hg19) we soon be available.

-nocnv If this option is set, no internal list for copy number variation (CNV) will be used.

-l	With this option a file containing a list of positions can be specified that should be excluded from the analysis. The format of this list is a tab delimited file where the first column is the chromosome (1,2,...,X,Y), the second is the start position, and the third is the end position. Sorting of this list is arbitrary.
-mwd	With this option the minimal segment size can be set. By default, the minimal segment size is 20. Low-resolution arrays may require smaller values.
-alpha	This option specifies the degree of smoothing. Larger values lead to less smoothing and smaller values to a stronger smoothing.
-uq	Selection of the upper quantiles. If multiple quantiles are considered, the different quantile selections have to be separated by commas (without banks in between).
-lq	Selection of lower quantiles.
-alfq	Significance threshold of the peel-off filter. Probes with the strongest contribution to an identified peak are successively removed until this significance threshold is exceeded.
-keepXY	Usually, X and Y chromosome are excluded from the analysis. If this option is set these chromosomes will also be analyzed.

Input Data

The program only accepts raw copy number data as input. The format of this data is a tab-delimited file as shown in the following example:

SNP	*Chromosome*	*PhysicalPosition*	*S00501*	*S00050*	*S00832*	*S00022*	*S01578...*
CN_473963	1	51599	1.66731	2.48207	3.23007	1.29837	2.08127...
CN_473964	1	51672	2.14241	1.85938	2.20043	1.45043	1.7367...
CN_473965	1	51687	2.39072	1.80866	2.18474	1.31353	1.82445...
CN_477984	1	52016	1.85889	2.27857	1.63643	1.91214	2.34428...
CN_473981	1	52784	1.48395	1.77011	2.02233	1.45814	1.73668...
...							

The header row (in italics) specifies besides fixed fields like *SNP, Chromosome, and PhysicalPosition* the sample name of each specimen included in the analysis. The following rows contain the actual data, where the first column is a unique probe identifier, the second column specifies the chromosome of the probe (as a single number, or X, and Y), the third column holds the position on the chromosome, and the following columns are the raw copy number signals of each probe.

Examples:

Three example data sets are available that can be downloaded from http://www.uni-koeln.de/med-fak/cgars/. Please find below examples how to analyze these data sets with CGARS:

1) 63 samples of small cell lung cancer:

To analyze:
cgars analyze -cn SCLC.cn -uq 0.25,0.05 -lq 0.25,0.05 SCLC_cgars
To plot the results:
cgars plot -a 0.01 SCLC_cgars

2) 146 samples of squamous cell lung carcinoma:

To analyze:
cgars analyze -cn SQ.cn -uq 0.25,0.05 -lq 0.35,0.25 SQ_cgars
To plot the results:
cgars plot -a 0.01 SQ_cgars

3) 70 samples of pulmonary carcinoids:

To analyze:
cgars analyze -cn CA.cn -uq 0.05,0.15 -lq 0.05,0.15 CA_cgars
To plot the results:
cgars plot -a 0.01 CA_cgars

License:

CGARS: Cancer Genome Analysis by Rank Sums
Copyright (C) 2013, Martin Peifer

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public Licensealong with this program. If not, see <http://www.gnu.org/licenses/>.