rCGH : a comprehensive array-based genomic profile platform for precision medicine

Supplementary methods

Frederic Commo[1,2], Justin Guinney[2], Charles Ferté[1,2], Brian Bot[2], Celine Lefebvre[1], Jean-Charles Soria[1,3], Fabrice André[1,3]

1) INSERM U981, Gustave Roussy, University Paris XI, Villejuif, France
2) Sage Bionetworks, Seattle, WA
3) Department of Medical Oncology, Gustave Roussy, Villejuif, France

Workflow description

The rCGH package is designed to read and process aCGH data, individually.
In order to keep a high level of flexibility, as well as to allow a better control of the entire process, we have let the workflow decomposed into 5 five steps described below. A convenient way to run multiple profiles would be to wrap the different steps into one global function, and to loop over a list of files to process. In such case, each sample is processed independently, and the genomic profiles shout not be impacted by the other data.

Object builders:
Since input files may not contain sufficient information in their preamble (depending on the software extraction version), we have implemented specific *read-file* functions, one for each of the supported type of array:

- *readAgilent()*: for Agilent FE files, in a text format.
- *readAffySNP6()*: for Affymetrix SNP6.0, cychp.txt, cnchp.txt or probeset.txt, exported from CEL files, through ChAS or Affymetrix Power Tools.
- *readAffyCytoScan()*: for Affymetrix CytoScanHD, cychp.txt, cnchp.txt or probeset.txt, exported from CEL files, through ChAS or Affymetrix Power Tools.

This step reads a file - possibly in a compressed format, e.g. .gz or .bz2 -, with respect to platform format specificities, saves array information, when exist, and renames each items in order to get a standardized output format. Note that any useful information can be added at any moment during the process.

In order to increase the flexibility, rCGH also supports custom arrays, provided the file to read is a data frame with the following mandatory columns:

- *ProbeName*: Character strings. Typicaly the probe ids.
- *ChrNum*: numeric. The chromosome numbers. In case Chr X and Y are used and named as "X" and "Y", these notations will be converted into 23 and 24, respectively.
- *ChrStart*: numeric. The chromosomal probes locations.
- *Log2Ratio*: numeric. The corresponding Log2Ratios.

The dedicated function to read such files is *readGeneric()*.

Signal adjustment:

When Agilent dual-color hybridization are used, GC content and the cy3/cy5 bias are necessary adjustments, which have to be carried out before segmenting a genomic profile. This step takes care of these adjustments, before computing the LRR. In both cases, a local regression is applied[1]. Note that by default, the cyanine3 signal is used as the reference.

Since Affymetrix cychp or cnchp files contain already computed Log_2 relative Ratios (LRR), this signal adjustment simply rescales the LRR values, if specified by the user, then stores the following quality scores: the derivative Log Ratio Spread (dLRs) and the LRR Median Absolute Deviation (MAD). For all platforms, LRRs are finally smoothed to remove outliers, and to reduce the noise.

In case of custom arrays, rCGH expects LRR already computed from data preprocessed according to user's specifications.

Profile centralization:

The centralization step is crucial since it defines a neutral level (potentially 2-copies) from which gains and losses will be estimated. As mentioned in the main manuscript, we implemented here the method we discussed in a previous paper. Briefly, the vector of LRR is considered as a mixture of several gaussian populations: $g(x,\Theta) = \sum_{k=1}^{K} p_k f(x,\theta_k)$ where p_k and θ_k are the proportion and the parameters distribution of the k^{st} population, respectively. The p_k and θ_k are estimated using an EM algorithm applied on multiple subsets of LRR, and the mean of the lowest density higher than a given proportion (specified by the user) of the maximum density, when

exists, is chosen as the centralization value. Here again, the user has full control on what should be the optimal centralization for a given profile (supplementary figure 1).

Segmentation:

This steps aims to identify breakpoints within the LRR continuity, each possibly defining the end of a DNA region and the start position of a new region with a different mean LRR value. As mentioned in the main manuscript, this step uses the CBS algorithm with slight modifications aiming to facilitate its use. The DNAcopy R package involves several parameters, which have to be set before the segmentation process (see DNAcopy R package for more details). We particularly focused on 2 parameters: the *alpha* value, which specifies a significance level for the test to accept change-points, and the *undo.SD*, which specifies the number of SDs between segment means to keep a split (when “sdundo” method is chosen as the “undo.split” procedure). We simplified the use of *DNAcopy* functions by embedding the different steps into one unique procedure, and by fixing *alpha = 1e-4*. Given this fixed alpha value, the user can either set a undo.SD value (typically between 0.5 and 1.5), or let rCGH estimating an optimal value from the MAD. The decision rule was constructed as follow:

80 genomic profiles (40 Agilent 180K, and 40 Affymetrix CytoScanHD) were manually checked in order to estimate their optimal undo.sd values, given *alpha = 1e-4*. In each case, we manually evaluated the consistency between the segmentation and the LRR signals, given several undo.sd values. The optimal values were finally modeled as a function of MAD (supplementary figure 2):

$$undo.sd = 0.48MAD^{\frac{1}{2}}$$

Genes table:

The goal of this final processing step is to report all the genes included in each segment, with their annotations and segmentation values. To do so, rCGH uses official gene annotations, called internally, according to the genome build specified by the user. Supported genome builds are hg18, hg19 (default setting) and hg38.

For a higher flexibility, the *byGeneTable()* function do not takes as argument a rCGH object itself, but its segmentation table. Hence any other segmentation table can be converted into an annotated genes table, provided data of the same format is passed, and the genome build to use is specified.

Default annotations include 'SYMBOL', 'ENTREZID', 'GENENAME' and 'MAP', but any other item can be exported using the *'columns'* argument. See the AnnotationDbi documentation for an exhaustive list of supported items.

Plot functions:
With this workflow we provide functions for several graphical outputs:

- *plotDensity()* shows the EM centralization..
- *plotProfile()* displays a static visualization of the genomic profile.
- *plotLOH()* displays a static visualization of the LOH profile.
- *multiplot()* combines the genomic and the LOH profiles in a unique graphical report.
- *view()* displays the profile in a web browser, and allows the user to interact with the graph in multiple ways through a command panel, as described below.

Interactive visualization functionalities:
The interactive visualization comes with a command panel and two tabs: the genomic profile is displayed in the first "CGH profile" tab, the gene table is available in "Gene table".
Below are described the functionalities available through the control panel.

- *Gene Symbol*: to display any existing gene, providing its official HUGO symbol.
- *Show chromosome*: to display the entire profile (default is 'All'), or one specific chromosome.
- *Merging segments shorter than*: to merge segments shorter than the size specified by the user, in Kb.
- *Gain/Loss colors*: to change the gain/loss colors. Segments are colored according to the specified thresholds.
- *Recenter profile*: to recenter the profile on-the-fly. The gene values will be also updated in the 'Gene table' tab.
- *Rescale max(y)*: adjusts the top y-axis (0<y) using a proportion of the maximum value.
- *Rescale min(y)*: adjusts the bottom y-axis (y<0) using a proportion of the minimum value.

- *Gain threshold (Log2ratio)*: defines the gain threshold. Segments higher than this value are colored in ‘gain’ color, and the 'Gene table' is filtered, consequently.
- *Loss threshold (Log2ratio)*: same as 'Gain threshold' for losses. Segments lower than this value are colored in ’loss’ color, and the 'Gene table' is filtered, consequently.
- *Download - Profile*: to download the profile as it is displayed on the screen, including modifications.
- *Download - LOH*: to download the LOH plot as it is displayed on the screen, including modifications.
- *Download - Table*: to download the 'Gene table', including modifications.

Notice that the “*Gene table*” also reports copy numbers in the “ApproxCN” column. However, we warn users that these values are approximations from the Log2Ratios, given the user’s current settings, and may not reflect precise integer chromosomal duplications or losses.

The web server version:

The web server version, aCGH-viewer, is a free web application which doesn’t require any R expertise. This application is designed for scientists (clinicians, biologists,…) who need to review and share profiles individually, e.g. during tumor board committees. This web version takes as input a segmentation table generated through either the rCGH package, or any other workflow, provided the data is of the same format as the standard CBS outputs.

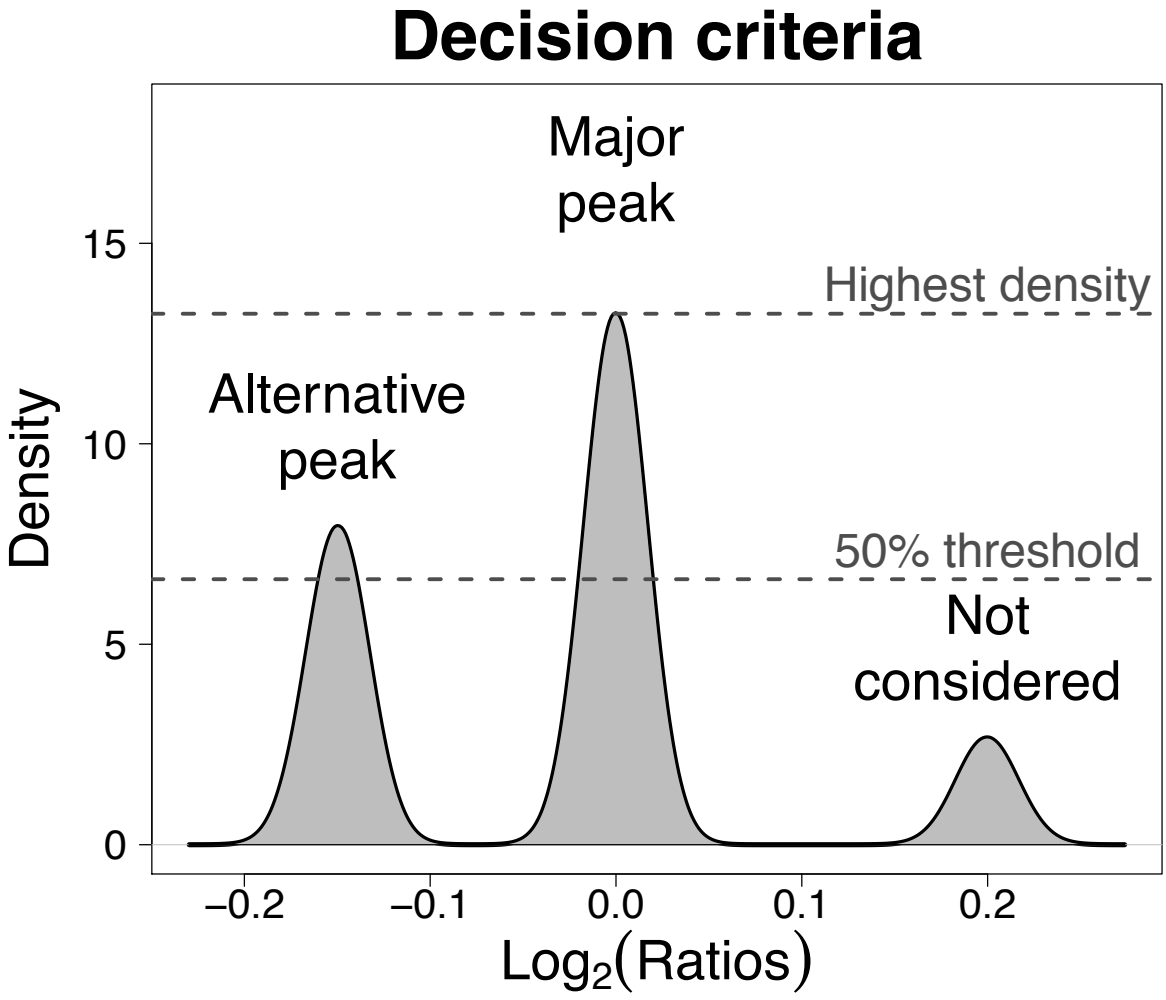
aCGH-viewer resumes the same functionalities provided by the interactive rCGH *view()* function, plus the genome build to use. This latter information is used internally to extract the genes contained in each segment, with respect to the corresponding chromosome, segment start and segment end. The generated genes table is displayed in the *Gene Table* tab, and stored internally to add to the plot any gene called by the user, at its appropriate genome location. Supported genome builds are hg18, hg19 (default setting) and hg38.

Validation

To validate our workflow, we ran the analysis of 995 cell lines from the Cancer Cell Lines Encyclopedia[2] (CCLE), and compared our results with those already published

and available at Gene Expression Omnibus (GEO, GSE36138). Affymetrix SNP6 CEL files were downloaded from http://www.broadinstitute.org/ccle/home, and processed using APT version 1.16.1: 942 out of the 995 CEL files matched the data available at GEO. The cychp.txt output files were processed using rCGH with the default parameters, and the newly generated profiles were compared with the corresponding GEO data by computing Pearson correlations on gene LRRs. We defined a high/low correlation threshold, as the quantile $q_{1e\text{-}2}$ of a normal distribution of same mean and standard deviation as the rho values, after Fisher transformation.

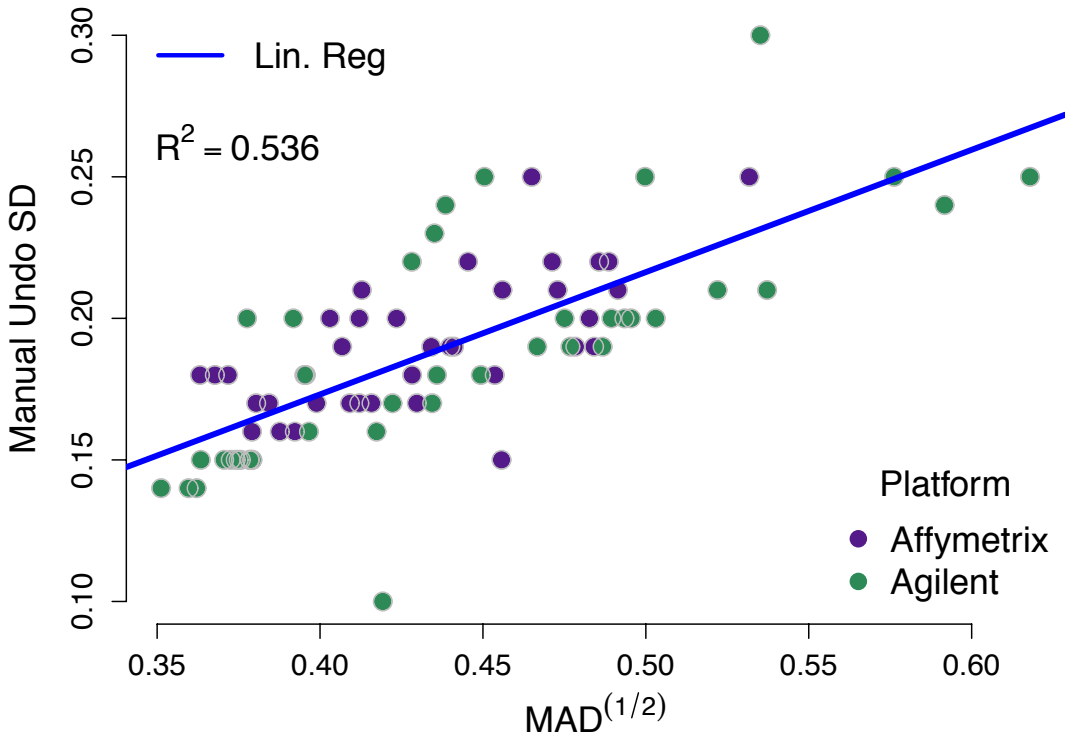
The median of the correlation values was 0.948, and the $q_{1e\text{-}2}$ low-correlation cut off was estimated as $q_{1e\text{-}2}$=0.723. According to this threshold, 910 out of 942 profiles (96.6%) were considered as well correlated with the published data (rho≥0.723), while the remaining 32 profiles showed low correlations (supplementary figure 3). The low-correlated profiles were characterized by a small number of short alterations (supplementary figure 4). A higher sensitivity of such profiles to slight variations in the segmentation parameters, may explain these relative discrepancies.



Supplementary figure 1

Centralization decision rule

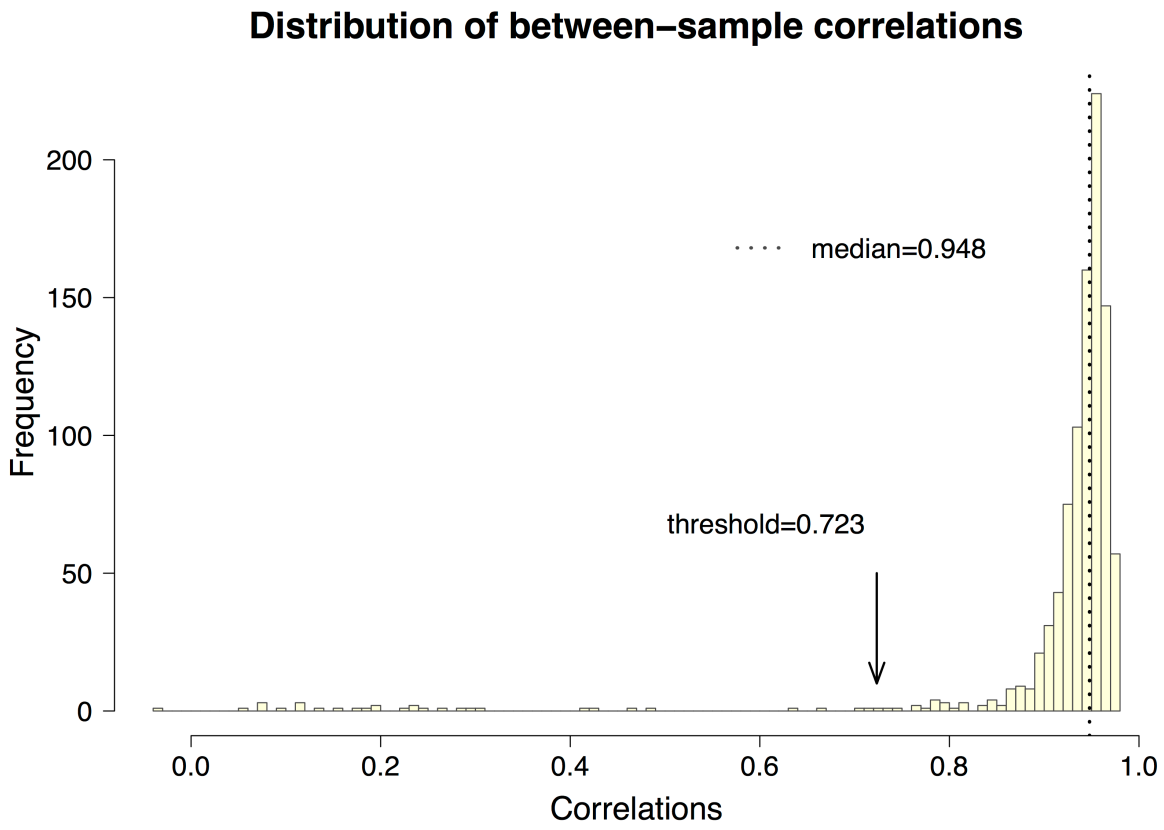
Since the major density peak may not always correspond to a neutral two-copies state, the centralization step allows a tolerance, expressed as a proportion of the highest density peak. When specified (default is 0.5), a peak with a density higher than proportion threshold can be used to centralize the entire profile.



Supplementary figure 2

Parameters optimization

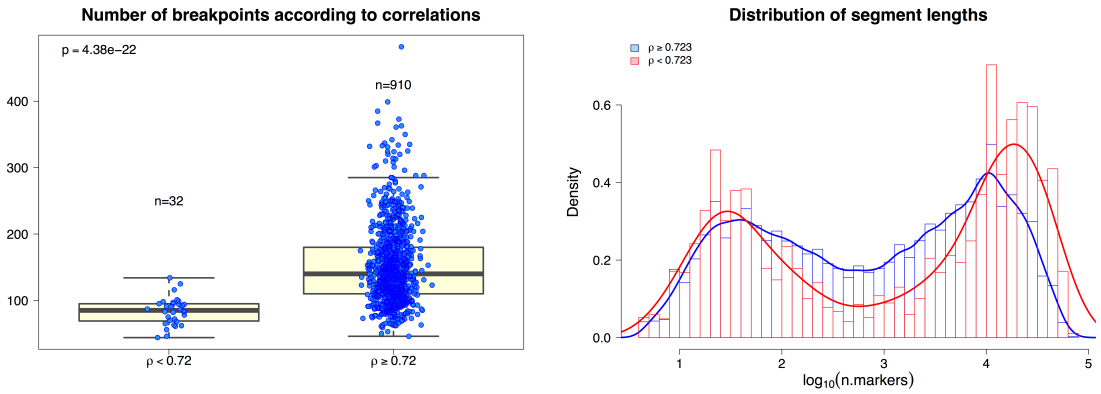
80 genomic profiles (40 Agilent 180K, and 40 Affymetrix CytoScanHD) were manually checked in order to estimate their optimal undo.sd values, given *alpha = 1e-4*. Optimal values were then modeled as a linear function of $MAD^{1/2}$.



Supplementary Figure 3

Between-profiles correlations

995 CCLE Affymetrix SNP6 CEL files were downloaded from Broad, then preprocessed using APT (version 1.16.1). Finally, cychp.txt files were used to reprocess the genomic profiles using rCGH workflow. Results were compared with the original data (GSE36138), using Pearson correlations of genes LRR, on 942 matched cell lines. The median of cell-cell profile correlations was 0.948, and 96.6% of the profiles (910/942) had a correlation greater than 0.723.



Supplementary Figure 4

Non-correlated profiles

When analyzing in details the 32 out of 942 samples with low correlations (rho < 0.723), we noted that 1) these profiles were characterized by a significantly lower number of breakpoints (p<2e-16) (left panel), and 2) breakpoints defined essentially very large and very short segments, while the correlated profiles showed copy number alteration of intermediate size (right panel). In case of short segments, it can be challenging to distinguish real altered DNA regions from noise, and profiles are generally more sensitive to small changes in analysis parameters. This may explain the discrepancies between the profiles generated through rCGH, and the original data on these particular cases.

Supplementary references

1. Smyth, G. K. & Speed, T. *Methods* **31,** 265–73 (2003).

2. Barretina, J. *et al. Nature* **483,** 603–7 (2012).