

Software

Open Access

M-CGH: Analysing microarray-based CGH experiments

Junbai Wang^{*1,2}, Leonardo A Meza-Zepeda^{1,2}, Stine H Kresse^{1,3} and Ola Myklebost^{1,2}

Address: ¹Department of Tumor Biology, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway, ²Department of Molecular Biosciences, University of Oslo, Oslo, Norway and ³Faculty of Medicine, University of Oslo, Oslo, Norway

Email: Junbai Wang* - junbaiw@radium.uio.no; Leonardo A Meza-Zepeda - leonardm@radium.uio.no; Stine H Kresse - stinek@radium.uio.no; Ola Myklebost - olam@radium.uio.no

* Corresponding author

Published: 09 June 2004

Received: 19 March 2004

BMC Bioinformatics 2004, 5:74 doi:10.1186/1471-2105-5-74

Accepted: 09 June 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/74>

© 2004 Wang et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: Microarray-based comparative genomic hybridisation (array CGH) is a technique by which variation in relative copy numbers between two genomes can be analysed by competitive hybridisation to DNA microarrays. This technology has most commonly been used to detect chromosomal amplifications and deletions in cancer. Dedicated tools are needed to analyse the results of such experiments, which include appropriate visualisation, and to take into consideration the physical relation in the genome between the probes on the array.

Results: M-CGH is a MATLAB toolbox with a graphical user interface designed specifically for the analysis of array CGH experiments, with multiple approaches to ratio normalization. Specifically, the distributions of three classes of DNA copy numbers (gains, normal and losses) can be estimated using a maximum likelihood method. Amplicon boundaries are computed by either the fuzzy K-nearest neighbour method or a wavelet approach. The program also allows linking each genomic clone with the corresponding genomic information in the Ensembl database <http://www.ensembl.org>.

Conclusions: M-CGH, which encompasses the basic tools needed for analysing array CGH experiments, is freely available for academics <http://www.uio.no/~junbaiw/mcgh>, and does not require any other MATLAB toolbox.

Background

In cancer, gene amplification and deletion frequently contribute to alterations in the expression of oncogenes and tumour-suppressor genes, respectively. Thus, detection and mapping of these DNA copy number changes are important for both the basic understanding of cancer and its diagnosis [1]. Comparative genomic hybridisation to DNA microarrays (array CGH) allows efficient, genome-wide analyses of relative genome copy number in a single experiment. In array CGH [1,2], copy numbers would be related to the Cy3:Cy5 fluorescence ratios (hereafter called

CGH ratios) of the microarray targets bound to each probe spot. There are some public available tools for array CGH analysis, but they either only run in Excel [3] or the software does not support the pre-processing (filtering or normalization) of array CGH data [4]. Therefore, there is a need for tools, preferably platform independent, which are capable of assessing the quality of CGH arrays as well as identifying the DNA copy number changes and link these with relevant genome information. We describe here the development of M-CGH, a MATLAB toolbox for

analysing CGH ratios, which has the ability to reliably locate the copy number changes.

Implementation

M-CGH can directly load clone chromosomal location files and data export files from microarray image analysis packages, such as QuantArray or GenePix, in tab-delimited text format. Microarray experiments contain many sources of errors, i.e. human operations, array printer performance, labelling and hybridisation efficiency [5]. Therefore, reliable pre-processing of CGH ratios is essential for the later detection of copy number changes. In M-CGH, we first exclude all empty and manually flagged spots, and then filter out spots with intensities lower than background in both channels and net signal intensities below two times standard deviations of the local background intensities (figure 1a). Multiple ratio normalization functions (simple normalization, intensity dependent normalization and sub-array position normalization) are available in M-CGH, which help to correct various errors that affect CGH ratios [5]. We assess the quality of spot reproducibility by calculating the mean and standard deviation for the CGH ratios of repeated spots, and eliminate probes for which standard deviation exceed 0.2, or if the ratio measurement is based on a single spot.

Following this analysis, we use data from within each experiment to determine the upper and lower threshold for scoring amplified and deleted chromosomal segments. This is not trivial, since some aneuploid samples may have many copy number deviations, complicating the determination of the "normal" ratio. In particular, dedicated arrays may be enriched for genes that are targets of amplification or deletion, thus making normalization difficult. We accomplish this by using a maximum likelihood method to fit a mixture of three Gaussian distributions (representing amplifications, normal and deletion) to a histogram of normalized \log_2 CGH ratios from the array CGH analysis [3]. The fitted distribution is

$$f(y) = \sum_{i=1}^3 P_i \left(\frac{1}{\sigma_i} \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y - \mu_i}{\sigma_i} \right)^2}$$

where P_i , μ_i and σ_i are the relative proportion, mean and standard deviation of the i^{th} distribution respectively (figure 1b) and $\sum_{i=1}^3 P_i = 1$. CGH ratios representing normal

DNA copy numbers are centered at a mean closest to zero, decreased copy numbers less than zero, and increased copy numbers greater than zero. We obtain the fit by first applying the Expectation Maximization (EM) algorithm [6] to estimate P_i , μ_i and σ_i for each Gaussian distribution,

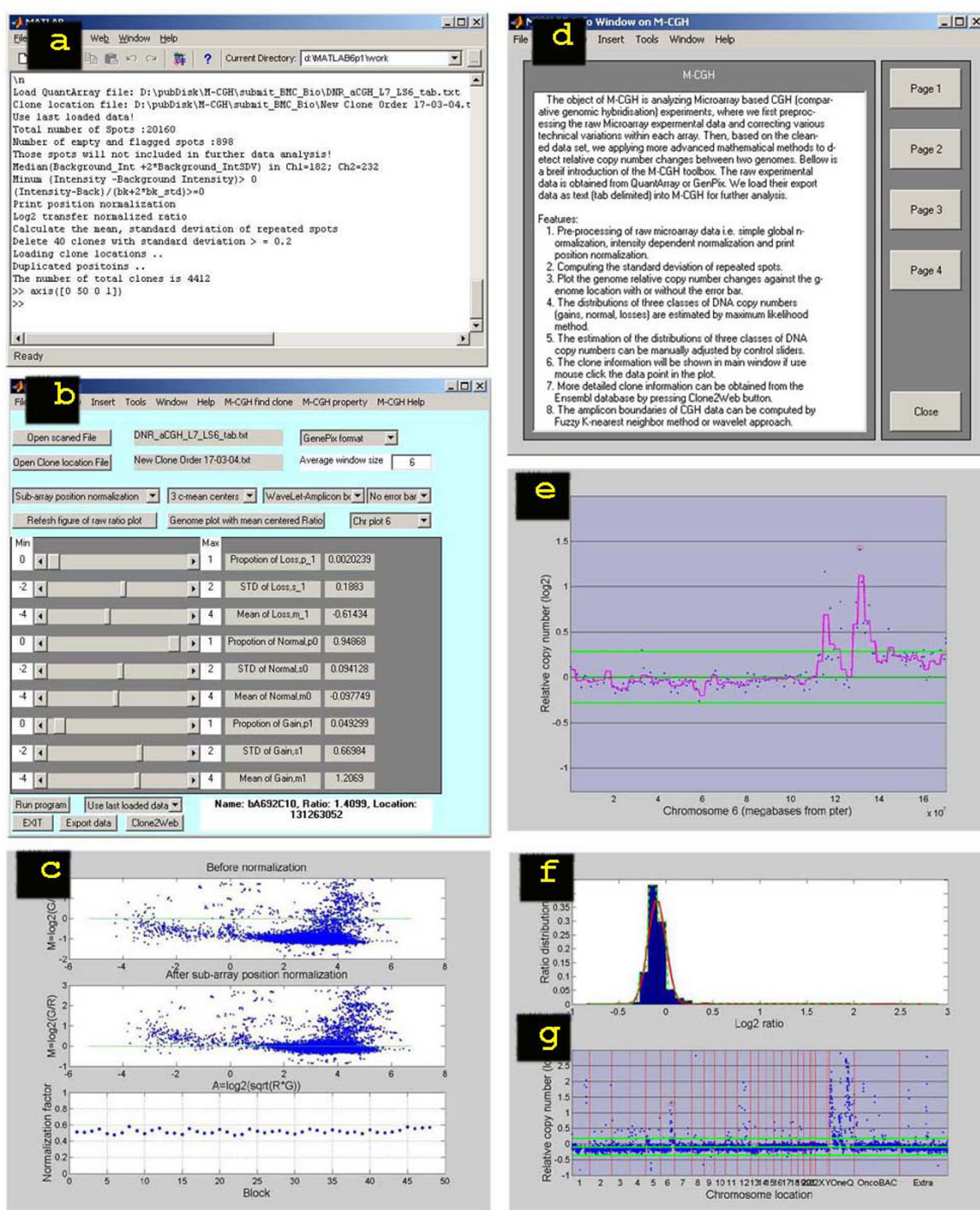
and then plotting an interactive histogram (figure 1f) of normalized \log_2 CGH ratios, with initial fitting of the three Gaussian distributions. The interactive histogram provides visual feedback with a smooth fit line superimposed on the observed data histogram. Users can assess the quality of the initial fitting, and manually adjust the control sliders of P_i , μ_i and σ_i in the M-CGH main window (figure 1b) until the sum of the three Gaussian distributions fits the histogram of CGH ratios (as in figure 1f, a solid red line shows the result of the fit of the sum of the three distributions). After a fit is obtained, we use the distribution fitted to the part of the histogram representing normal DNA copy number to determine the 3σ upper and lower thresholds for determining amplifications and deletions (as in figure 1g, 3σ upper and lower thresholds are represented by two solid green lines). Array CGH experiments in which the central distribution has a σ greater than 0.2 will be considered unreliable, and may be eliminated from further studies. M-CGH also provides an interactive plot of \log_2 CGH ratios as a function of its location in the genome (figure 1g) or chromosome (figure 1e). When a data point is clicked on, the CGH ratio, clone name and location will be displayed in the main window (figure 1b). Detailed clone information, such as gene content, can be obtained from the Ensembl database by pressing the Clone2Web button (figure 1b).

Fuzzy K-nearest neighbour method and wavelet approach

For computing the amplicon boundaries from the CGH array analysis, we first use the nearest neighbour interpolation to smooth observed CGH ratios (users may modify the window size for the interpolation, figure 1b). Then we apply the fuzzy K-nearest neighbour method [7] to classify the data points into three classes (gains, normal and losses) based on estimated means μ_i of the three Gaussian distributions, and the fuzzy membership values will represent the level of copy number changes. Alternatively, we may use the wavelet approach [8] to estimate the amplicon boundaries (as in figure 1e, the pink smooth line is the amplicon boundaries computed by wavelet approach).

Results

The performance of M-CGH was tested on a genomic microarray containing approximately 4000 unique elements (BACs and PACs). The array included a representation of the human genome at 1 Mb resolution, as well as the tiling path of a segment of chromosome 1 (1q12-q25) and 600 genomic clones containing known oncogenes and tumour-suppressor genes. A panel of human sarcomas, malignant tumours of mesenchymal origin, was used to assess the performance of the software (Meza-Zepeda, Kresse, Wang, Myklebost et al., unpublished). Total genomic DNA from approximately 20 tumours was labelled by random priming using Cy3-dCTP, in parallel

**Figure 1**

Illustrations of the M-CGH user interface, 1a) Summary information of each array CGH experiment is listed in the MATLAB command window; 1b) M-CGH main window, showing the parameters used, allowing manual adjustments, and spot information with a link to the Ensembl database (Clone2Web); 1c) Sub-array position normalization of CGH ratios, where the lower panel show the normalization factor used in each sub-array, indicating lack of discrepancies in this case; 1d) M-CGH help documentation; 1e) An interactive plot of log₂ CGH ratios as a function of their relative chromosomal locations, with estimated amplicon boundaries (pink smooth line); 1f) An interactive histogram plot of CGH ratios, with fitted Gaussian distributions (red smooth line); 1g) An interactive plot of log₂ CGH ratios as a function of their genome location, showing all chromosomes in numerical sequence, delimited by red lines.

normal reference DNA was labelled using Cy5-dCTP. Labelled tumour and reference DNA was competitively hybridised to the genome representation present on the array. Hybridisation was performed using an automated hybridisation station, GeneTAC (Genomic Solutions/Perkin Elmer), agitating the hybridisation solution for 48 hours at 37 °C. After hybridisation, slides were washed and scanned using an Agilent G2565BA scanner (Agilent Technologies). Images were analysed using GenePix Pro 4.1 (Axon Laboratories) or QuantArray 3.0 (Packard Biosciences). The spots were automatically segmented and manually adjusted where necessary. Export files were generated, transformed to tab delimited format and imported to M-CGH for further analysis. Figure 1 shows a sarcoma sample analysed using M-CGH.

Conclusions

Overall, we have presented a package for analyzing array CGH experiments, interactive data analysis with a user friendly graphical interface is available, and an on-line genomic information database (Ensembl) is linked. M-CGH is entirely platform independent and only requires MATLAB installed (the student version will suffice).

M-CGH states the number of spots that pass the quality control, allows all plots to be saved as image files, and further export of the filtered dataset, with normalized ratios and clone location, as a text file. Our future development of M-CGH will include applying visualization methods capable of assessing the DNA copy number changes of multiple CGH arrays simultaneously, filtering of misbehaving clones, implementing change-point analysis techniques such as those described by Lucito et al. [9], i.e. representational oligonucleotide microarray analysis, and developing a web tool so that it can be run without MATLAB software.

Availability and requirements

- **Project name:** M-CGH: Analysing microarray-based CGH experiments
- **Project home page:** http://www.mikromatrise.no/array_cgh/index.html
- **Operating system(s):** Platform independent
- **Programming language:** MATLAB
- **Other requirements:** MATLAB Version 6.1
- **License:** GNU
- **Any restrictions to use by non-academics:** Please inform the corresponding author if you are a non-academic user

Authors' contributions

JBW designed and developed software and drafted manuscript. LAMZ attended part of design and tested program, and drafted part of manuscript. SHK provided experiment dataset and tested software. OM conceived of the study, participated in its design and coordination.

Acknowledgements

We thank one anonymous reviewer for constructive comments on the manuscript. This work was supported by the Norwegian Cancer Society <http://www.kreft.no>, the FUGE (functional genomics) program <http://www.fuge.no>, and the University of Oslo (EMBIO). The arrays were produced by the Norwegian Microarray Consortium.

References

1. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet* 1998, **20**(2):207-211.
2. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional programs of human breast tumors.** *Proc Natl Acad Sci USA* 2002, **99**(20):12963-12968.
3. Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, Moore D, Albertson DG, Pinkel D, Collins C, Hanahan D, Gray JW: **Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas.** *Nat Genet* 2001, **29**(4):459-464.
4. Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, Kallioniemi A: **CGH-Plotter: MATLAB toolbox for CGH-data analysis.** *Bioinformatics* 2003, **19**(13):1714-1715.
5. Churchill GA: **Fundamentals of experimental design for cDNA microarray.** *Nat Genet Sup* 2002, **32**:490-495.
6. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *J Roy Stat Soc* 1977, **39**:1-38.
7. Keller JM, Gray MR, Givens JA JR: **A fuzzy k-nearest neighbour algorithm.** *IEEE SMC* 1985, **15**:580-585.
8. Mallat S: **A wavelet tour of signal processing.** 2nd edition. San Diego: Academic Press; 1999.
9. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West AJ, Rostan S, Nguyen KCQ, Powers S, Ye QK, Olshen A, Venkatraman E, Norton L, Wigler M: **Representational Oligonucleotide Microarray Analysis: A High-Resolution method to detect genome copy number variation.** *Genome Res* 2003, **13**(10):2291-2305.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

