Development and Implementation of an Analysis Tool for Array-based Comparative Genomic Hybridization

M. Kreuz[1], M. Rosolowski[1], H. Berger[1], C. Schwaenen[2], S. Wessendorf[2], M. Loeffler[1], D. Hasenclever[1]
[1] Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany
[2]Department of Internal Medicine III, University Hospital of Ulm, Ulm, Germany

Summary
Objectives: Array-comparative genomic hybridization (aCGH) is a high-throughput method to detect and map copy number aberrations in the genome. Multi-step analysis of high-dimensional data requires an integrated suite of bioinformatic tools. In this paper we detail an analysis pipeline for array CGH data.
Methods: We developed an analysis tool for array CGH data which supports single and multi-chip analyses as well as combined analyses with paired mRNA gene expression data. The functions supporting relevant steps of analysis were implemented using the open source software R and combined as package aCGHPipeline. Analysis methods were illustrated using 189 CGH arrays of aggressive B-cell lymphomas.
Results: The package covers data input, quality control, normalization, segmentation and classification. For multi-chip analysis aCGHPipeline offers an algorithm for automatic delineation of recurrent regions. This task was performed manually up to now. The package also supports combined analysis with mRNA gene expression data. Outputs consist of HTML documents to facilitate communication with clinical partners.
Conclusions: The R package aCGHPipeline supports basic tasks of single and multi-chip analysis of array CGH data.

Keywords
Array CGH, DNA copy number, gene expression, gene dosage

Methods Inf Med 2007; 46: 608–613

doi:10.1160/ME9064

1. Introduction

Aberrations in copy number play an important role in various diseases, especially in the pathogenesis of malignant tumors [1]. The length of aberrant segments can range from a few base-pairs to whole chromosomes. Copy number aberrations can affect gene expression: Deletions may lead to tumor suppressor gene inactivation and copy number gains may cause activation of oncogenes [2].

Array comparative genomic hybridization (aCGH) is a method to detect and map these copy number aberrations in the genome. Several thousand known DNA clones or oligonucleotides are spotted on a chip [1]. Each clone represents a specific region of the genome. Resolution and coverage of the analysis depends on the number of spotted clones and their distribution in the genome.

DNA is isolated from test and reference tissue and differentially labeled using fluorescence dyes. A balanced mixture of labeled test and reference DNA is hybridized to the CGH array.

Test and reference DNA compete for free binding sites [3]. Signals of test and reference fluorescence intensity at each clone position are measured, preprocessed [4] and combined as a log2 ratio. This signal is assumed proportional to the log2 ratio of test and reference copy number in the corresponding genomic region. If the reference tissue is chosen euploid, information on copy number changes in the test sample can be obtained. Raw data from one CGH-array experiment consists of several thousand clone-specific log2 ratios.

2. Motivation

Due to data complexity, manual interpretation of array CGH data is time-consuming and error-prone. Automatic methods which facilitate the array CGH analysis are required.

We developed an analysis tool for array CGH data [5] which supports single and multi-chip analysis as well as combined analysis with paired mRNA gene expression data. The methods were implemented using the open source software R [6] and combined as package aCGHPipeline. The package is available upon request. Most of the currently available analysis programs are limited in the capability of multiple-chip analysis. For example, CGH-Plotter [7] and CGHPro [8] provide only a graphical comparison of different chips but there is no support for further statistical investigations of genetic differences. The aim of aCGHPipeline is to overcome these limitations.

Here we describe relevant steps in the analysis of array CGH data that can be performed using functions from our tool. Figure 1 gives a schematic overview over tasks addressed below which are supported by the package. Illustrations are taken from an analysis of array CGH data of 189 aggressive B-cell lymphomas [9]. Each array contained 2799 BAC/PAC clones. 1500 of these clones cover the whole genome at intervals

of approximately 2 Mb, the remaining 1299 clones span regions known to be frequently involved in B-cell neoplasms or contain proto-oncogenes or tumor suppressor genes [9].

3. Methods

3.1 Quality Control

As for other high-throughput methods a quality control step is required to detect and reject invalid data. For every chip and every clone aCGHPipeline calculates the median absolute deviation and the proportion of missing values. Summary statistics, histograms and lists of arrays and clones which are suspicious are summarized in a user-friendly HTML output file for further inspection.

Large-scale copy number polymorphisms also occur as natural variation of the normal human genome. The length of polymorphic segments ranges from kilo- to megabases [10]. To distinguish between normal variation and cancer-specific aberrations, it is important to compare the genomic positions of the used clone probes with known polymorphic sites. Therefore every clone on the array is compared with the "Database of Genomic Variants" (http://projects.tcag.ca/variation/) [11]. Clones which are located in regions with known copy number polymorphisms are marked, enabling the user to interpret the measured values cautiously.

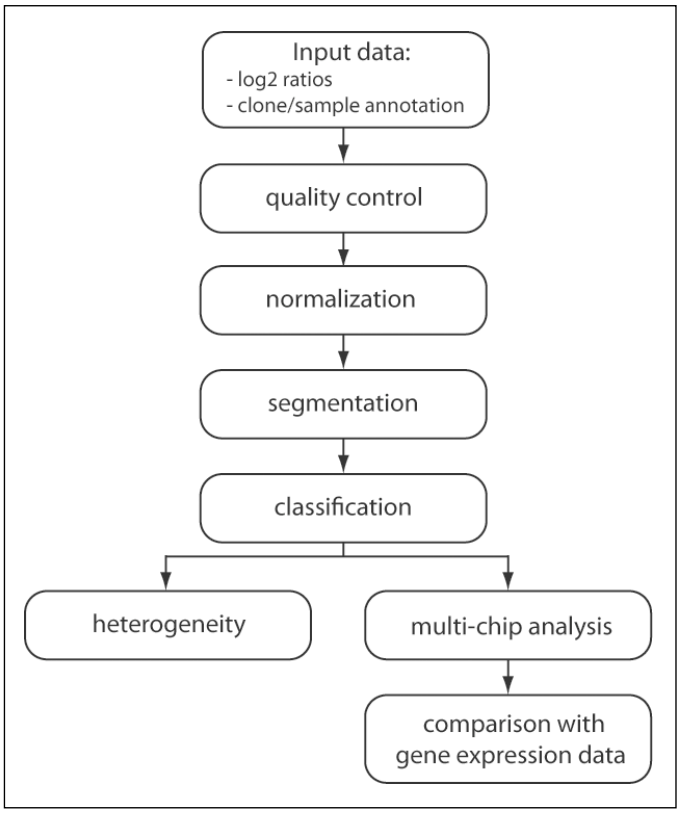

Fig. 1
Steps of analysis in array CGH data

3.2 Normalization

Imbalances in the amount or quality of used test and reference DNA may lead to systematic array-specific bias in the measured log2 ratios. Differences in labeling efficiencies of the fluorescence dyes may be eliminated by using a dye swap design [12]. The aim of normalization is to remove chip-specific bias. aCGHPipeline provides a global normalization for the correction of array CGH data: Assuming that the majority of clone positions is euploid in both tissues most of the log2 ratios should vary about zero. Thus, a measure of location, like mean, median or mode, can be used to estimate the bias individually for each array. The bias is then removed by subtracting the measure of location from all measurements on this array. We take the 50% of the measurements with the highest values in the corresponding density function and calculate the mean of these values weighted with the appropriate density as an estimator for the mode of the distribution. We suggest using this mode estimator for normalization, because based on a simulation study it is more robust in cases with a relevant proportion of aberrant clones.

3.3 Segmentation

The goal of array CGH analysis is the detection of copy number aberrations in tumor DNA. Measured signals have to be classified in chromosomal gains, losses and regions with normal copy number. Copy numbers cannot be directly read off the signal ratios: The signal to noise ratio is rather low [13] and samples of primary tumors are typically contaminated with a certain amount of non-tumor DNA thus attenuating the ratio. A strategy for noise reduction is needed.

Copy number aberrations typically occur in segments comprising several clones. Segmentation methods aim at detecting segments of neighboring clones with the same copy number and to smooth the signal to reduce the noise.

Several segmentation algorithms for array CGH data are described in the literature [2, 7, 14-19]. We have implemented an interface to the Hidden Markov Model (HMM)-based segmentation method of Fridlyand et al. [17] and to the Circular Binary Segmentation (CBS) method of Olshen et al. [16]. Based on the results from other research groups [20, 21], and following our own experience, we currently recommend using the CBS method of Olshen.

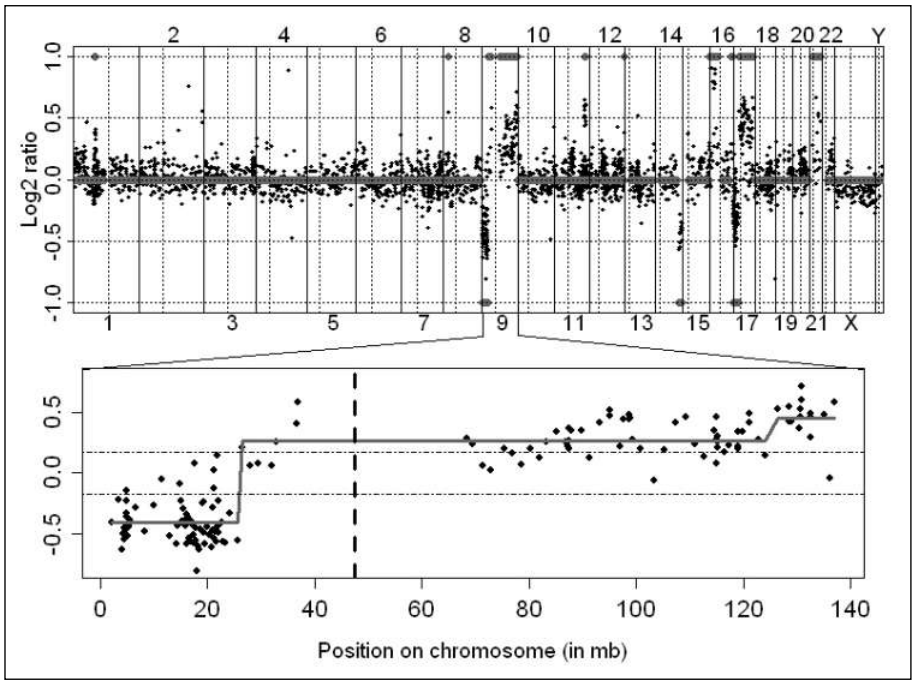

Fig. 2 The upper graphic illustrates an array CGH profile of a single chip. Black points show the normalized log2 ratios in genomic order. Grey bars indicate the classified data (0 = normal copy number; 1 = copy number gain; –1 = copy number loss). The lower graphic illustrates the segmentation for chromosome 9. The bold grey line marks the resulting segments. Thresholds for classification are indicated by dotted black lines.

3.4 Classification

After segmentation each array is represented by a set of segments and smoothed segment-specific signals. To ease interpretation the segments are classified into gains, losses and segments with normal copy number. Note that with current technology further differentiation within the loss or gain category is difficult except for high-level amplification peaks. Thresholds are required to achieve such a subdivision. If a segment level is above the threshold, the segment is classified as a gain. A segment which is below minus the threshold is counted as a loss. This process is also referred to as threshholding. aCGHPipeline supports choosing a fixed or a noise-dependent threshold. A fixed threshold is uniformly specified on the log2 ratio scale for all arrays. A noise-dependent threshold is specified in units of an array-specific noise estimate. For each array aCGHPipeline provides a robust noise estimator based on median absolute deviation of the differences between measured log2 ratios and the corresponding smoothed segment levels. Thus the noise estimate is not inflated by genetic heterogeneity. Choosing the noise-dependent approach leads to an increased threshold for chips with poor quality but not for chips with a high genetic heterogeneity.

Users have to specify the classification threshold manually considering information on the fraction of normal tissue contamination in tumor samples and the sensitivity of the array type used. For a given array platform the sensitivity can be assessed on the basis of data with known copy number aberrations or through the analysis of DNA with different numbers of X chromosomes [13, 22].

An example of classified array CGH data is shown in Figure 2. After classification the genetic heterogeneity of each array can be summarized, e.g. by calculating the number of aberrant segments or clones. These heterogeneity measures help to quantify the instability of the genome in different tumor entities or to analyze the impact of heterogeneity on the prognosis of patients.

3.5 Multi-chip Analysis

When analyzing a large set of tumor samples a major biological question is to detect regions which are commonly aberrant and therefore may be key events for tumor development or proliferation.

A major objective of multi-chip analysis is thus to delineate genomic regions which are characteristically aberrant in a specific type of cancer. These recurrent regions provide lists of putative candidate genes for biological follow-up research. Presence or absence of recurrent regions in single tissues define candidate variables for prognostic factor analyses.

Up to now these recurrent regions were visually determined e. g. by looking at heatmaps (Fig. 3b). We developed an algorithm to automatically propose recurrent regions. We first calculate and plot the frequencies of gains and losses for each clone along the genome (see Fig. 3a for a single chromosome). To delineate recurrent regions as "genomic regions which are characteristically aberrant in the data" we divide clones into segments with the same frequency of aberration and define recurrent regions as those resulting segments that are both above a certain threshold and dominate adjacent segments.

3.5.1 Segmentation of Multiple-chip Frequency Data

For segmentation we use a Hidden Markov Model (HMM) [23]. Each state has a binomial emission distribution with varying underlying probability p. HMMs with up to five states are fitted chromosome-wise to the frequencies of gains and losses of the clones because the number of recurrence levels is initially unknown. The model that fits best is selected via the Akaike information criterion (AIC). Smoothing is achieved through a state transition matrix giving high probability to staying in the same underlying state. The result of the segmentation procedure for gains and losses is illustrated in Figure 3a.

3.5.2 Threshold for Recurrence

As a next step we have to separate "characteristic" segments from sporadic ones. Every segment is therefore compared with a frequency threshold. Regions which show an aberration more frequently than the selected threshold were counted as recurrent segments. Recurrent segments that dominate neighboring segments are selected as recurrent regions.

The user can specify a frequency threshold or use a threshold that is suggested by our algorithm. aCGHPipeline uses a 2-means clustering of the frequencies of gains and losses of the clones to distinguish sporadic form characteristic aberrations. The mean of the center of the "sporadic aberrations" and the center of the "characteristic aberrations " centers is used as the threshold.

Automatically delineated recurrent regions should be checked manually by inspecting a heatmap because theoretically the identity of the cases contributing to a recurrent region may vary along the segment indicating that the region should be split.

3.5.3 Analysis of Recurrent Regions

Having delineated the recurrent regions in the data set one may want to decide on presence or absence of recurrent regions in single samples. We have implemented a voting algorithm for this purpose. A recurrent region consists of a set of clones. A recurrent region is called present in an individual sample if and only if a user-specified proportion (e.g. 50%) of its clone set is classified as concordant aberration.

The algorithm determines a matrix of CGH arrays by recurrent regions indicating presence and absence of the recurrent regions on the respective chip. The horizontal blue bar in Figure 3b indicates the result of the voting algorithm for the recurrent region marked by the vertical grey bar.

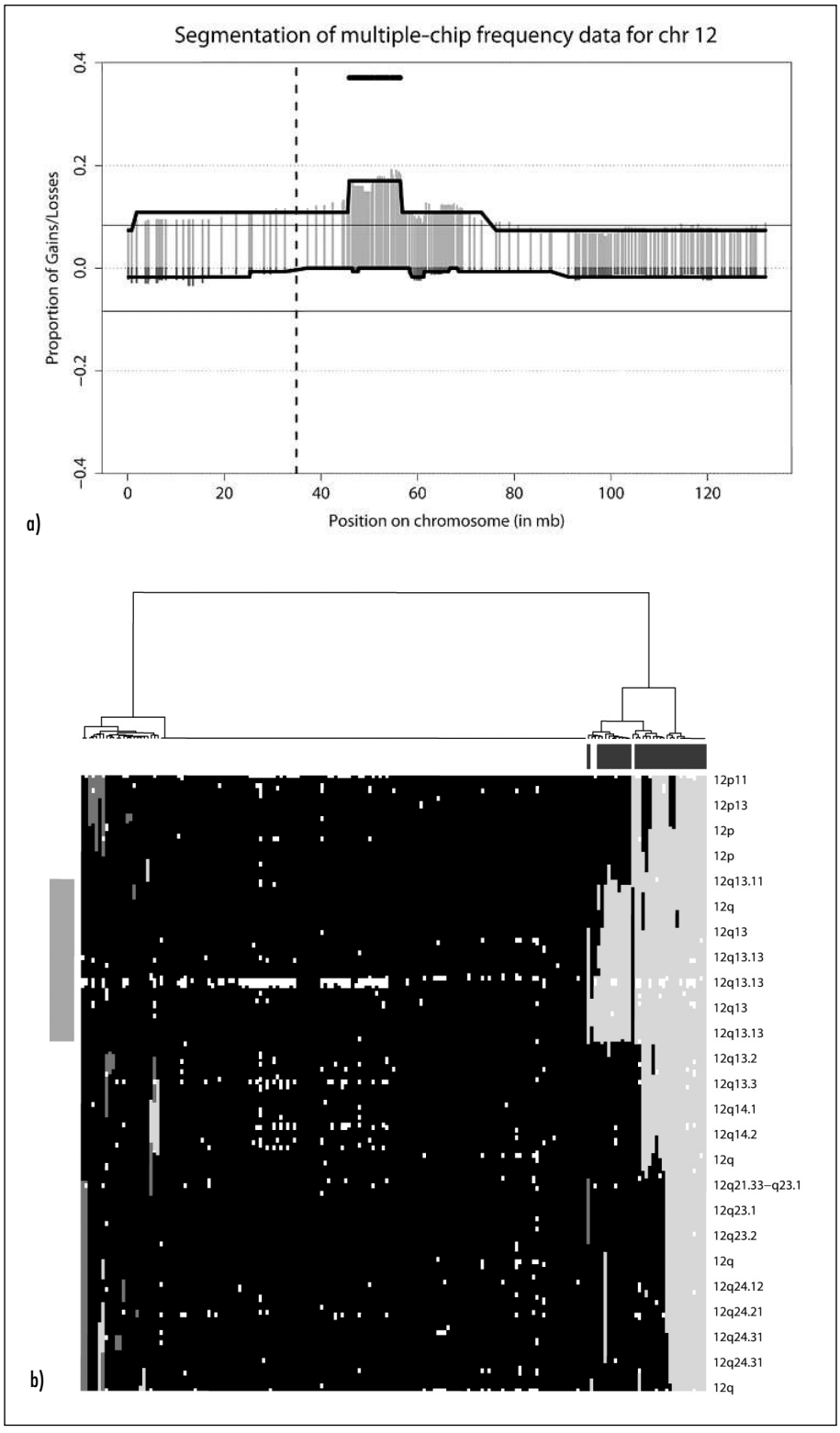

Fig. 3 a) Aggregated data of 183 lymphoma CGH arrays for chromosome 12. Light and dark grey bars indicate the proportion of gains and losses for each clone. Segmentation via HMM is shown by bold lines. Black horizontal lines mark the threshold for recurrence. The recurrent region (45.7-56.9 mb) is highlighted by a bold black bar. b) The heatmap of copy number aberrations on chromosome 12. Light grey lines indicate copy number gains, dark grey lines copy number losses. White points mark missing values. The bold bar on the left side shows the recurrent region. Black marks on the top indicate cases in which the recurrent gain is called present.

3.6 Analysis of Paired Array CGH and Gene Expression Data

If paired array CGH and gene expression data is available one can ask whether a re-

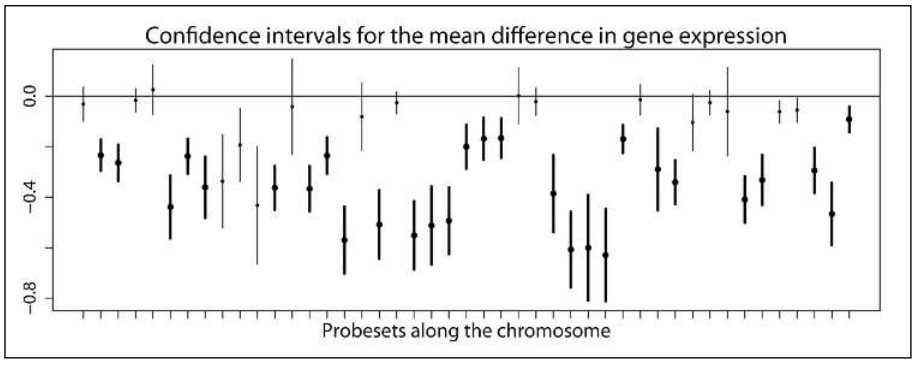

Fig. 4 Gene expression data in a recurrent loss on chromosome 6. Vertical lines illustrate the 95% confidence intervals for the differences in gene expression of cases with and without a loss. Bold lines mark probe sets with a significant gene dose effect.

spective aberration is associated with a gene dose effect on mRNA gene expression. Given the assignment of copy number aberrations in recurrent regions to individual samples as described above, mRNA expression in samples with the aberration can be compared to samples without this aberration.

Figure 4 depicts confidence intervals for the difference in mean gene expression comparing cases with and without a loss for genes mapping into a recurrent region on chromosome 6 in lymphoma data.

If a higher resolution is desired, a mapping assigning copy number information from individual clones to genes in the corresponding genomic region is required.

We have implemented an algorithm that generates a data pair for each gene expression probe set consisting of the respective mRNA expression level and a copy number state derived from the array CGH data. If a probe set maps on one or more CGH clones their copy number state is assigned if equivalent. If the probe set maps in a gap between clones on the CGH array the copy number state of the neighboring clones is returned if equivalent. Otherwise a missing value is assigned.

4. Application

The analysis pipeline was applied to a set of 189 arrays of aggressive B-cell lymphomas. Within the quality control step six arrays were rejected due to high noise. In addition 48 clones were excluded from the analysis because they were missing in more than 50 percent of the cases. 291 of the remaining clones are located in regions with known copy number polymorphisms.

Array CGH data was normalized using global normalization with mode estimator. Normalized log2 ratios were smoothed by applying the Circular Binary Segmentation (CBS) method of Olshen et al. with default parameters.

For a subset of 106 arrays, manual interpretation of the data was available. Best concordance between manual and automatic classification was achieved by using a fixed threshold of 0.17 for the classification of the segments resulting from CBS. In that case 96.8% of all clones were assessed consistently. More than 83% of the clones which were called aberrant by manual interpretation were concordantly detected by the automatic method. Investigation of discrepant results showed that in cases where one method assigned a copy number gain while the other assigned a loss mostly represented sign errors within the manual analysis. The bulk of discrepancies occurred at boundaries of aberrant regions and in cases where the putative aberration is quantitatively weak. In these cases a validation of the results is difficult.

Using the methods for multi-chip analysis 16 regions of recurrent gains and six regions of recurrent losses were delineated for further investigation.

5. Conclusion and Outlook

With aCGHPipeline we developed a comprehensive tool for the analysis of array CGH data. The package covers data input, quality control, normalization, segmentation and classification. The automatic analysis essentially reproduced the manual interpretation. It is less error-prone, less time-consuming and more reproducible, thus further analyses of our group will be based on it.

For multi-chip analysis aCGHPipeline proposes an algorithm for automatic delineation of recurrent regions. This task was performed manually up to now. The package also supports combined analysis with mRNA gene expression data. Results are presented as HTML documents to facilitate communication with clinical partners. aCGHPipeline is implemented in R in which further data analysis can easily be performed. The current version of the program provides no graphical user interface so that users need a basic knowledge of R programming. A planned integration of aCGHPipeline in the gene expression warehouse (GeWare) [24] will add this feature.

Acknowledgments
This work was supported by a grant from the Deutsche Krebshilfe (70-3173-Tr3) within the Molecular Mechanisms in Malignant Lymphomas Network Project. Markus Kreuz is supported by a predoctoral grant (GRK 1034) of the Georg August University of Göttingen.

References

1. Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet 2003; 12: 145-152.
2. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005; 11: 6-27.
3. Repsilber D, Ziegler A. Two-color Microarray Experiments. Methods Inf Med 2005; 44: 400-404.
4. Rahnenführer J. Image Analysis for cDNA Microarrays. Methods Inf Med 2005; 44: 405-407.
5. Kreuz M., Diplomarbeit: Entwicklung und Implementierung eines Auswertungswerkzeuges für Matrix-CGH-Daten. 2006.
6. Ihaka R, Gentleman RR. A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 1996; 5 (3): 299-314.
7. 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.
8. Chen W, Erdogan F, Ropers HH, Lenzner S, Ullmann R. CGHPRO – a comprehensive data analysis tool for array CGH. BMC Bioinformatics 2005; 6: 85.
9. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF, Bernd HW, Cogliatti SB, Dierlamm J, Feller AC, Hansmann ML, Haralambieva E, Harder L, Hasenclever D, Kühn M, Lenze D, Lichter P, Martin-Subero JI, Möller P, Müller-Hermelink HK, Ott G, Parwaresch RM, Pott C, Rosenwald A, Rosolowski M, Schwaenen C, Stürzenhofecker B, Szczepanowski M, Trautmann H, Wacker HH, Spang R, Loeffler M, Trümper L, Stein H, Siebert R; Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. N Engl J Med 2006; 354 (23): 2419-2430.
10. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature 2006; 444 (7118): 444-454.
11. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genetics 2004; 36 (9): 949-951.
12. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002; 30 (4): e15.
13. Bilke S, Chen QR, Whiteford CC, Khan J. Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. Bioinformatics 2005; 21 (7): 1138-1145.
14. Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R. A method for calling gains and losses in array CGH data. Biostatistics 2005; 6 (1): 45-58.
15. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 2004; 20 (18): 3413-3422.
16. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004; 5 (4): 557-572.
17. Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A. Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 2004; 90 (1): 132-153.
18. Jong K, Marchiori E, Meijer G, Vaart AV, Ylstra B. Breakpoint identification and smoothing of array comparative genomic hybridization data. Bioinformatics 2004; 20 (18): 3636-3637.
19. Myers CL, Dunham MJ, Kung SY, Troyanskaya OG. Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics 2004; 20 (18): 3533-3543.
20. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 2005; 21 (22): 4084-4091.
21. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 2005; 21 (19): 3763-3770.
22. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A 2002; 99 (20): 12963-12968.
23. Rabiner LR, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 1989; 77 (2): 257-286.
24. Rahm E, Kirsten T, Lange J. The GeWare data warehouse platform for the analysis of molecular-biological and clinical data. Journal of Integrative Bioinformatics 2007; 4 (1): 47

Correspondence to:
Markus Kreuz
University of Leipzig
Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Haertelstr. 16-18
04107 Leipzig
Germany
E-mail: markus.kreuz@imise.uni-leipzig.de