4 Development of CGHPRO

Until now, array CGH technology has been used primarily as a research tool. With the further technical optimisation, it will enter the realm of clinical application, where it will have a profound impact on the screening and genetic counselling of patients with genomic rearrangements. Combined with RNA and protein analysis, array CGH will substantially enhance our understanding of the relation between disease, phenotype and the underlying molecular defects, and there is reason to believe that array CGH will also be a clue to the identification of risk factors for common diseases, which have been so difficult to find by other approaches.

In this study, to facilitate the application of array CGH in research and as a diagnostic tool, a comprehensive data analysis software called 'CGHPRO' was developed. The program contains a whole set of packages for statistical analysis and visualisation of array CGH data.

4.1 Software development

CGHPRO was programmed in Java and MySQL was used as the back-end database. The decision to use Java and MySQL was based on their public availability, their platform independence and the fact that MySQL can handle large data files with high throughput. The "R" packages from Bioconductor (http://www.bioconductor.org), DNAcopy and aCGH, were implemented in our software, which enable a platform-independent characterization of genomic profiles. Up to now, CGHPRO has been tested in a Linux, Windows 2000 and windows XP environment.

4.2 Notation of array CGH data

In this section, before discussing the development of CGHPRO in detail, a few aspects of the array CGH terminology will be introduced.

4.2.1 Intensities

In BAC array CGH experiments, the data acquisition process (scanning of the slide) results in at least four parameters for each spot, the foreground and background intensities of red and green fluorescence (Rf, Rb, Gf, Gb). If no background correction is applied, the foreground intensities, Rf and Gf are used as the input (which are simply represented by R and G) for normalization and data visualization. Otherwise, the (Rf-Rb) and (Gf-Gb) are taken as input.

4.2.2 Ratios

The data for each spot is usually also represented as the ratio between red and green signal intensities. The ratio X for the ith spot is simply

$$X_i = \frac{Ri}{Gi}$$

Ratios provide a direct measure of DNA copy number changes. Compared with a normal diploid sample, heterozygous duplicated regions in a test sample have a theoretical ratio of 1.5, whereas regions with heterozygous loss have a ratio of 0.5.

4.2.3 Log-intensities and log-ratios

Usually, the intensity and ratio values of spots across a slide differ within a range covering several orders of magnitude, which is difficult for data visualization. This problem is usually solved by a logarithmic transformation that produces a continuous spectrum of values and spreads the values more evenly across the data range. In addition to that, a logarithmic transformation tends to make the variability of data more constant over the intensity range.

4.3 Overview of the data analysis process in CGHPRO

CGHPRO has been designed to analyse array CGH data in a comprehensive way. Users are guided through the analysis process, as shown in Figure 6. Once the back-end database is set up and chromosome positions of clones are stored, the Results file of the image analysis software can be imported and the features of each hybridisation can be checked by a variety of graphic representation tools.

Depending on the hybridisation characteristics of each experiment, the most suitable normalization method can be chosen. Normalization effects can be evaluated again by graphical representation. After an appropriate normalization, the characterization of individual genomic profiles can be performed using various methods. Finally, all results can be visualized in an interactive interface, stored in the back-end database, and used for comparative analysis.

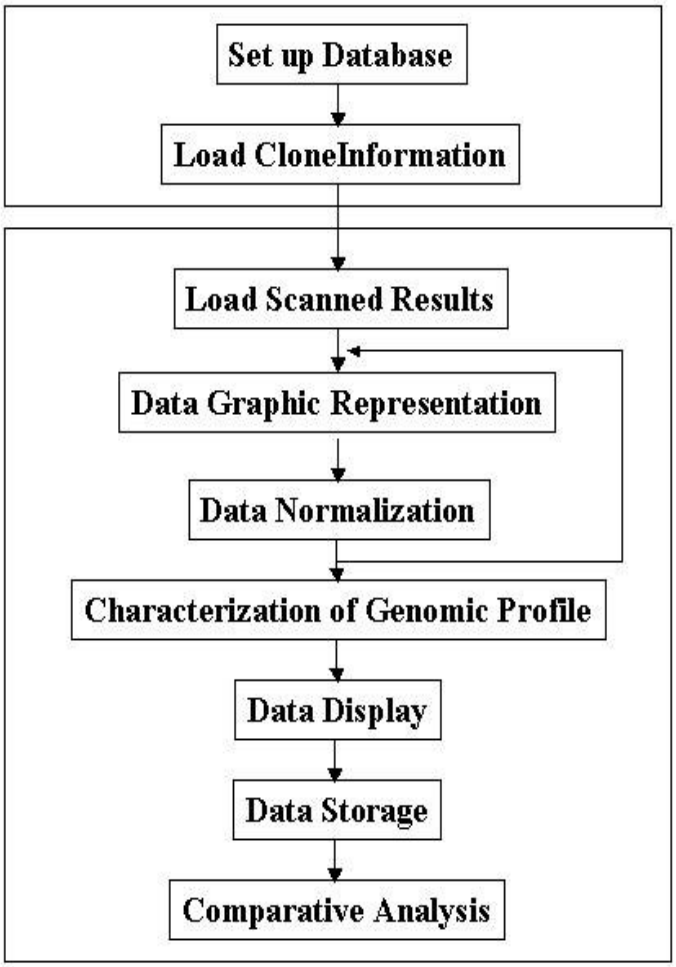

Figure 6: Overview of data analysis process in CGHPRO

4.4 Database design

In CGHPRO, a back-end database that uses MySQL has been implemented. The database stores the description of each analysed chip (glass slide) as an entry in the table 'AnalysedChips'. This description includes all essential information about the experimental and data analysis procedure, e.g. the number of spots that

have been excluded and the normalization algorithm applied. A separate table named according to the Chip ID saves the original data from the image analysis software as well as the results from data analysis. A table called 'ClonePosition' is used to store the user-derived mapping information for each clone. The information comprises data that might influence the reliability of the clone's hybridisation characteristics, such as content of repetitive sequences and most importantly, its involvement in segmental duplications, which can be visualized by a colour code, as discussed below. In addition, a table called 'Aberration' is used to save the aberrations determined by users and in this table, the chromosomal position, the characteristics (gain or loss) and the patient phenotype are described for each aberration.

4.5 Data input

In CGHPRO, mapping information for each clone, based on a specific version of UCSC Genome Browser, has to be provided by user. This can be done simply by loading a file containing the mapping information of clones into the back-end database. The tab-delimited file must include six fields for each clone, the unique identifier, the respective chromosome, the positions of the first and last base pair, the source of the clone, and the user-specified comments of the clone. For the complete tiling path from BACPAC Resources Centre, the mapping information based on the April 2003 (NCBI build 33), July 2003 (NCBI build 34), and May 2004 (NCBI build 35) assembly of UCSC Genome Browser are distributed with the software.

The way this information is acquired differs from other recently published programs like ArrayCGHbase (Menten et al., 2005) or CAPweb (http://bioinfo-out.curie.fr/Capweb), both of which provide these data by directly accessing the respective genome browser. This may be an advantage when looking for the most recent update, but it may pose problems for diagnostic and related applications, where patient confidentiality is important and precludes online data analysis. Offline analysis also speeds up the procedure, as it is not dependent on server capacities or data transfer rates.

CGHPRO allows the import of result files from GenepixPro5.0, Agilent and Imagene, but users can also customize the program to support their own tab delimited data format by specifying which column corresponds to what data field. After importing the result files, essential data are extracted and spots, flagged as "poor" by the image analysis software, are excluded automatically. Mapping information and related annotations for each clone are fetched from the back-end database.

4.6 Graphical analysis of hybridisation characteristics

Visualization of hybridisation characteristics helps to assess the success of the experiment and can guide the choice of normalization method and analysis tool. Therefore, CGHPRO provides a variety of graphical data representation tools to visualize the data before and after normalization.

4.6.1 Scatter plot

In a scatter plot, CGHPRO plots the log-intensity of the red dye against the log-intensity of the green dye: $\log_2 R$ versus $\log_2 G$. This helps to identify the relationship between two dyes and allows for estimates of the noise within a given data set (Figure 7).

4.6.2 MA-plot

An MA-plot is a scatterplot with transformed axes. The X-axis conforms to the logarithm of the average intensity value of the two dyes; the Y-axis shows exactly the log-ratio of the two dyes (Figure 8).

$M = \log_2 R - \log_2 G$ and $A = 1/2(\log_2 R + \log_2 G)$

MA-plots are especially useful for the detection of the intensity-dependent effects in log-ratios (Yang et al., 2002).

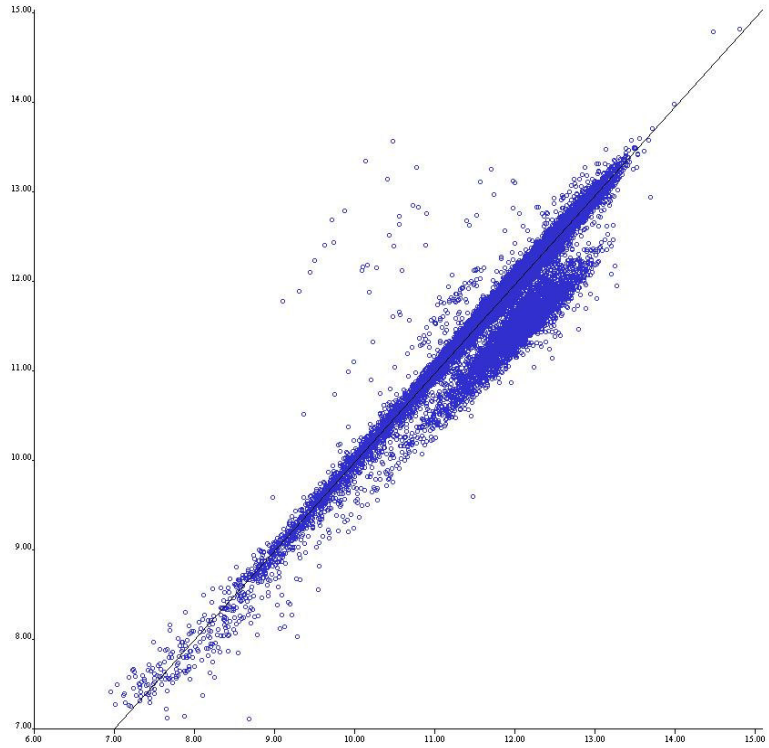

Figure 7: Scatterplot. In a scatterplot, the log-intensity of the red dye (Y axis) is plotted against the log-intensity of the green dye (X axis).

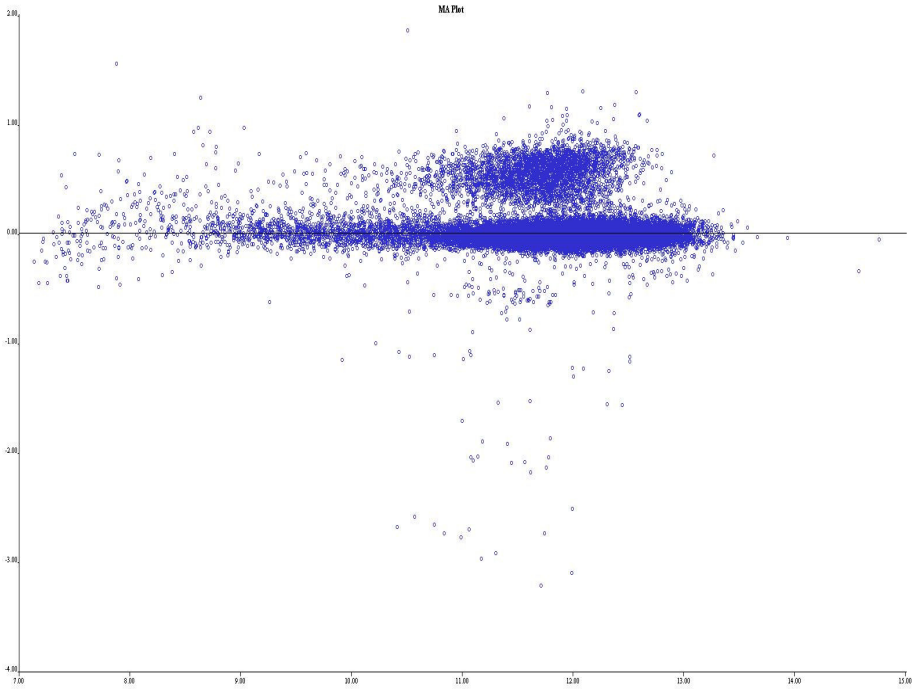

Figure 8 : MA-plot. The X-axis conforms to the logarithm of the average intensity value of the two dyes; the Y-axis shows exactly the log-ratio of the two dyes

4.6.3 Boxplot

A boxplot diplays the central tendency and variability of the data. The box in the middle represents the interquartile range (IQR). The median is marked as line in the middle of box while the whiskers show the spread of the data. In spotted microarray platform, one slide usually consists of a number of different subgrids, where each subgrid is printed with a same print-tip. In order to compare the log-ratios between different subgrids and thus detect the spatial dependency of log ratios, CGHPRO draws boxplot for each subgrid (Figure 9).

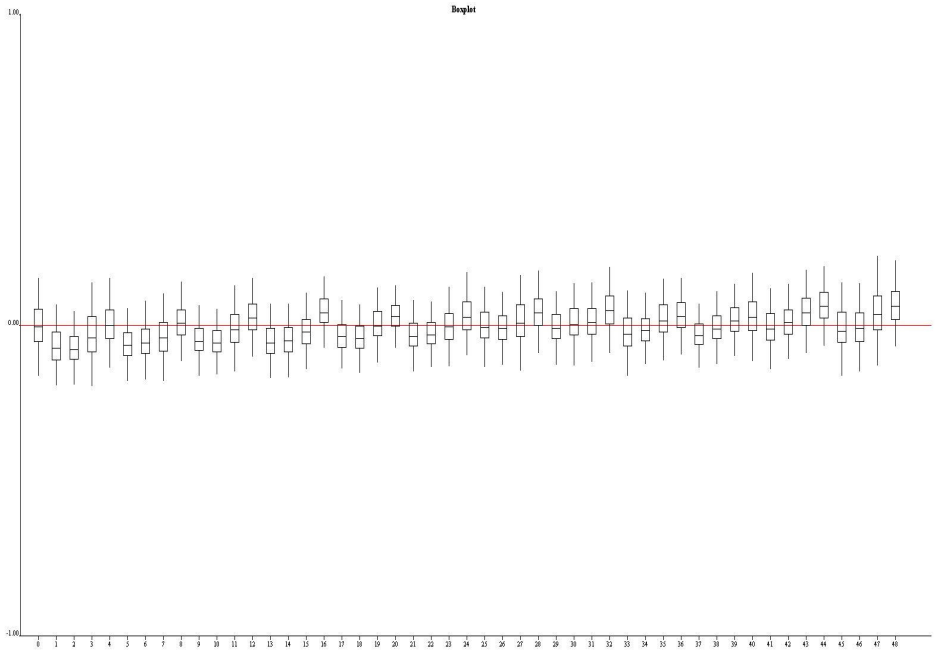

Figure 9: Boxplot. The box in the middle represents the interquartile range (IQR). The median is marked as line in the middle of box while the whiskers show the spread of the data. Here, each boxplot is plotted for one subgrid.

4.6.4 Histogram

A histogram showing the distribution of log-ratios for a single slide provides an overview about the distribution of the data points are distributed and therefore can assist with the choice of the more suitable normalization method or the more appropriate statistical analysis (Figure 10).

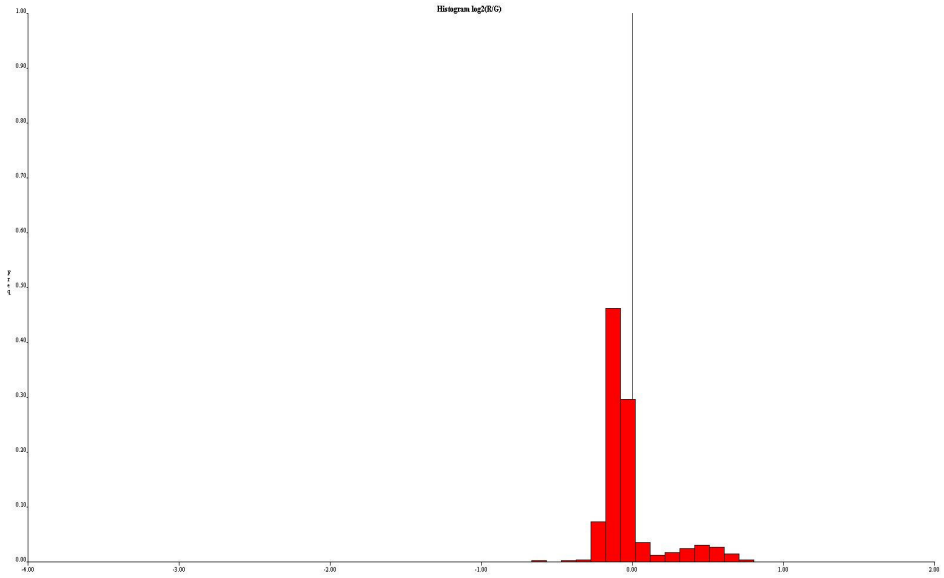

Figure 10: Histogram. The distribution of log-ratios for a single slide is shown here.

4.7 Data filter

CGHPRO enables the user to exclude individual spots based on the following criteria: signal intensity; standard deviation; number of replicates; involvement in segmental duplication; clone source and user-specified comments of the clone. Visual inspection can help to identify clones that should be excluded from further analysis.

4.8 Data normalization

The goal of normalization is to remove any systematic bias in the measured fluorescence intensities. Such systematic bias can originate from different labelling efficiencies of the used fluorochromes, different scanning parameters, and spatial or other effects. Depending on the experimental design and hybridisation characteristics, different normalization methods should be applied. Therefore, CGHPRO offers several options to perform normalization. Generally, these normalization methods can be classified as within-slide normalization and paired-slides normalization for dye-swap pairs.

4.8.1 Within-slide normalization

4.8.1.1 Global normalization

Global normalization methods assume that the red-green bias is constant across the array and the red and green intensities are related by a constant factor, i.e. R = kG. The goal is to estimate a constant factor c and correct the log ratios by simply subtracting c, so that the mean (or median) of the resulting log ratios is 0.

$$\log_2 \frac{Ri}{Gi} \rightarrow \log_2 \frac{Ri}{Gi} - c = \log_2 \frac{Ri}{kGi}$$

A widely used choice for parameter c = log2 k is the mean or median log ratio of the particular slide.

4.8.1.2 Intensity dependent normalization

Several reports have shown that ratio values can depend systematically on the overall spot intensities. The global normalization approaches does not account for this bias. Locally weighted scatter plot smoothing (lowess) or other robust linear regression methods can be used to remove such intensity-dependent effects. An easy way to visualize intensity-dependent effects is to generate a MA-plot for each slide to be normalized. It can be seen in the plot that the majority of points lie on a curve, showing that the red-green bias depends on the intensity of the spot. Lowess estimates this curvature and smoothes the MA-plot by subtracting the values of the estimated function from the original M-values.

$$\log_2 \frac{Ri}{Gi} \rightarrow \log_2 \frac{Ri}{Gi} - c(Ai) = \log_2 \frac{Ri}{k(Ai)Gi} \quad \textit{where} \quad \mathbf{A_i = 1/2(\log_2 R_i + \log_2 G_i)}$$

Here $c(A_i)$ is the lowess-fit to the MA-plot for the ith spot,i = 1, ...,N, and N is the number of spots.

4.8.1.3 Printing tip specific normalization

Every subgrid (or block) is printed with the same printing-tip. There may exist systematic differences between the tips, like differences in length or tip-opening and abrasion. These variations can cause spatial effects on the slide, which can be visualised by Boxplot. Previously explained methods (global and intensity dependent) can be adapted to account for this problem, simply applying them to every single subgrid of one slide.

$$\log_2 \frac{R_i}{G_i} \rightarrow \log_2 \frac{R_i}{G_i} - c_j(A_i) = \log_2 \frac{R_i}{k_{j(A_i)} G_i}$$

where $c_j(A_i)$ is the lowess-fit to the MA-plot for the jth subgrid and the ith spot. i = 1, ...,N, and N is the number of spots. j = 1, ...,M, and M is the number of subgrids.

4.8.2 Dye-Swap normalization

A dye-swap pair consists of two slides. Each hybridisation is done twice, with reverse dye assignment in the second hybridisation.

Slide1 $M_i = \log_2 \frac{R_i}{G_i} = \mu_i + c_i$

Slide2 $M'_i = \log_2 \frac{R'_i}{G'_i} = \mu'_i + c'_i$

where μ_i and μ'_i are the true log-ratios, c_i and c'_i the dye-effects. Because of reversed dye assignments one can expect:

$$\mu_i = -\mu'_i$$

Assuming that the dye biases in the two slides are similar, the log2-ratios for the two slides are combined:

$$\frac{1}{2}(M_i - M'_i) = \frac{1}{2}(\mu_i - \mu'_i + c_i - c'_i) = \mu_i \text{ if } c_i = c'_i$$

The normalized log2-ratios will then be

$$M_i = \hat{\mu}_i = \frac{1}{2}(\log_2 \frac{R_i}{G_i} + \log_2 \frac{R'_i}{G'_i}) = \log \sqrt{\frac{R_i G'_i}{R'_i G_i}}$$

Another possibility is to correct the single intensity values. Calculating

$$k = \sqrt{\frac{R_i G'_i}{R'_i G_i}}$$

and correcting the intensity values with this factor k

$$R_{i,corr} \rightarrow R_i \text{ and } G_{i,corr} \rightarrow kG_i$$

will lead to the same results as correcting the $\log_2$-ratios directly, but gives the opportunity to visualize the effects of normalization (e.g. with scatterplot, MA plot). This step is called self-normalization. To verify the assumption of c = c', the lowess-fits from both slides could be compared. If both fits show similar trends, self-normalization should provide reasonable results.

4.9 Replicate spots handling

If one clone is spotted more than once on the chip, CGHPRO will identify the replicate spots automatically because of their common ID. After normalization, the normalized ratios for the replicates are averaged and the standard deviations calculated. These average ratios will later be used to represent the ratios for the different clones. In subsequent analysis, users can set a threshold based on the number of replicas and standard deviation, such that clones exhibiting inconsistent results can be excluded.

4.10 Characterization of genomic profiles

The eventual goal of array CGH is the characterization of the individual genomic profile. Up to now, the common method is to use fixed thresholds, which should be dependent on the variability of the data. CGHPRO allows users to set a threshold either directly, or smooth the data first and then set a threshold based on the smoothed results. For smoothing, CGHPRO provides two options. When using the option “moving average”, which is applied to each chromosome separately, a window of adjustable size moves along the clones, which are ordered according to their base pair positions on the chromosome. The smoothed ratio of the clone at the centre of a window will be the average ratio of the clones within the window.

The second smoothing strategy is to segment the clones, which are ordered along the chromosome, into sets with equal copy numbers. Then the data can be smoothed via averaging within the sets.

CGHPRO includes two optional methods for the segmentation of chromosomes into regions with identical copy numbers, namely ‘Unsupervised Hidden Markov Partition’ created by Jane Fridlyand (Fridlyand et al., 2004) and ‘Circular Binary Segmentation’ first published by Adam Olshen (Olshen et al., 2004). The two methods were implemented by linking the two R packages, aCGH and DNAcopy, to the program. Based on the smoothed ratios generated by one of these two

algorithms, the Median Absolute Deviation (MAD) has been introduced as an objective measurement of data scattering.

4.11 Data display

4.11.1 Genomic display

The graphical interface of CGHPRO allows to explore the results in an interactive interface (Figure 11). In the Genome Display, the window consists of 24 sub-panels, each containing one chromosome. The 24 sub-panels are arranged as a 6 by 4 grid. In each sub-panel, the ratios of clones are plotted in a size-dependent manner along the ideogram. As described below, several display parameters can be modified.

In each sub-panel, there are three lines along each chromosome. The yellow line represents a log ratio of zero; the individually adaptable green and red lines mark the negative and positive log ratios, respectively. The smoothed log ratios calculated either by moving average, DNAcopy or HMM, can also be displayed as a black line called “Smooth Line”. Optionally, the original data can be blanked out.

Each clone is colour-coded according to its involvement in segmental duplications, as defined by the following formula: (∑Length of Duplication * Copy Number)/ Length of Clone. Based on the factors determined this way, the clones are grouped into seven classes that can be viewed separately by clicking on the button with the corresponding colour in the top right corner.

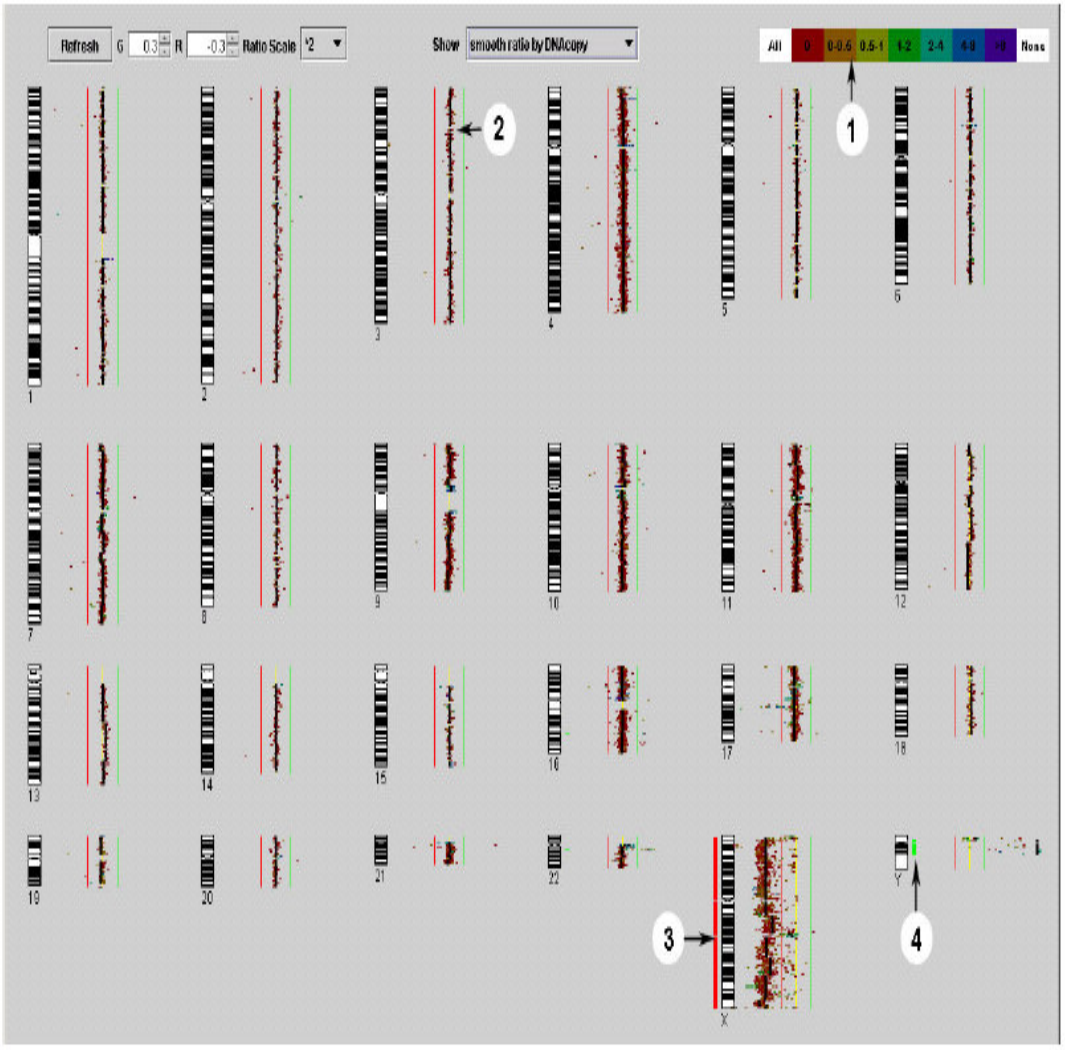

Figure 11: Genome Display exemplified by male versus female hybridization on a 14000 BAC DNA array. Circles 1-4: (1) Colour coding table indicating the involvement of clones in segmental duplication (2) Black line representing the smoothed ratios calculated by DNAcopy (3) and (4) red and green bars to the left and right side of the ideogram highlighting regions of losses and gains, respectively.

Segmental duplications, which comprise ~5% of the human genome, are copies of genomic DNA with >90% sequence identity that range in size from 1 to >200 kb and are present in at least two locations in the human genome (Bailey et al., 2002). Highlighting segmental duplications is useful for the recognition of clones that may show misleading ratio scores (Locke et al., 2004). Moreover, this feature also allows to relate chromosomal rearrangements to duplicated genomic regions. It has already been shown that segmental duplications increase the chances of non-allelic homologous recombination and that genomic regions flanked by these

duplications are particularly prone to rearrangements (Stankiewicz and Lupski, 2002).

A comprehensive understanding of the structural genome variation is essential for proper interpretation of array CGH data and their clinical significance. Special attention should be paid to aberrations that overlap with known variants. If the same aberration is found in individuals with and without the phenotype, very likely, the functional relevance of the aberration, if any should be quantitative rather than qualitative.

To facilitate the comparison of experimental data with the known copy number variants, CGHPRO includes the relevant information from the Database of Genomic Variants (http://projects.tcag.ca/variation/). According to the physical position and size, polymorphic regions are marked by transparent rectangles along the chromosome ideogram. Users can choose to view all known copy number changes or only those from specific sources, such as individual publications.

Clicking on each sub-panel will open a separate window and allow zooming in on a specific chromosome, as shown in Figure 12.

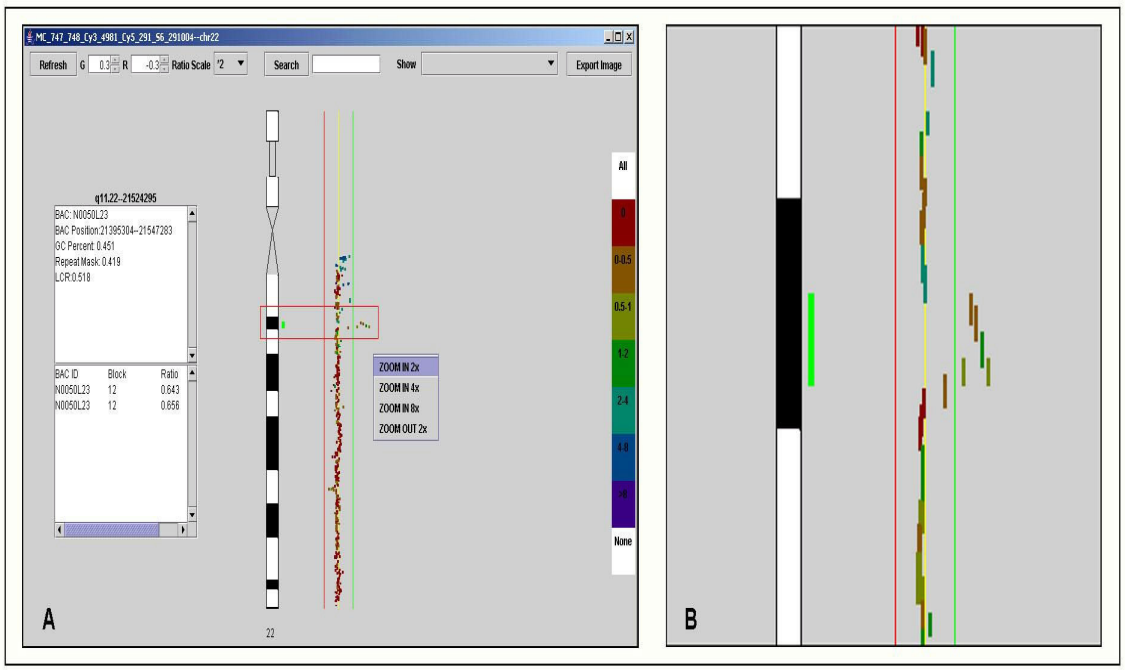

Figure 12: Chromosome Display (A) Detection of a duplication which encompasses about 300kb. (B) Zoom-in view of the relevant region (red rectangle in (A)).

4.11.2 Chromosome display

Chromosome Display provides a detailed view of the selected chromosome (Figure 12). In addition to the features provided by the Genome Display, the Chromosome Display allows to search for clones, to zoom in or out, and to export images. Upon clicking on a clone, information about its exact localization, simple repeats content, its involvement in segmental duplications, as well as information on number, position and ratio of the present replicas will be displayed in a text box. A key feature added to the Chromosome Display is a right-click mouse event, which will open a pop-up menu, offering several zoom options. Finally, Chromosome Display can be exported as an image file in Portable Network Graphics (png) format.

4.12 Comparative analysis of different chips

Once stored in the database, all entries can be used for comparative analysis at the genomic, chromosomal and clone-by-clone level. The feature “Genomic View” offers a summarizing display of chromosomal aberrations in a series of cases. In this mode, the absolute frequencies of aberrations within a study group are displayed alongside the chromosome ideograms ordered in a 6x4 grid. Upon clicking on the chromosomes of interest in the list located at the left side of the screen, the program switches to the Chromosome View and zooms in on the respective chromosome. In addition to the absolute frequencies of aberrations, the relative frequencies can also be shown, which makes it easier to compare study groups of different size (Figure 13).

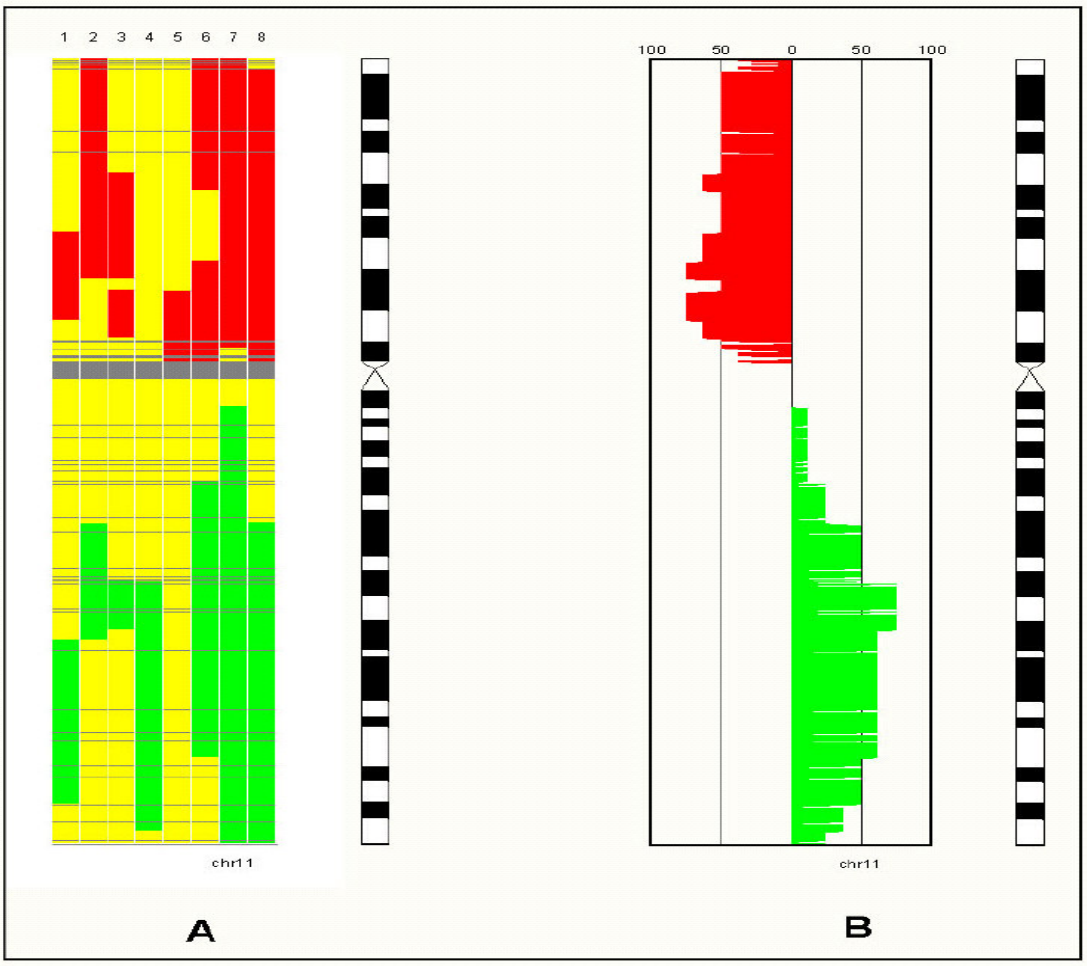

Figure 13:Comparative Analysis: CGHPRO supports the visualization of absolute (A) and relative (B) frequencies of chromosomal aberrations in a series of cases. Results can be displayed simultaneously for all or for single chromosomes, as shown here for chromosome 11.

For detailed analysis, the clone-by-clone view can be used. This mode supports "mouse over functionality", which displays further clone information in the bottom text field when the mouse is moved into the box representing a specific clone. As in all other view modes, balanced regions are indicated in yellow, while deleted and gained regions are shown in red and green, respectively. This option to simultaneously display results from several experiments is useful in the definition of shortest regions of overlap, can help to reveal patterns of chromosomal aberrations, and facilitates the identification of odd clones.

4.13 Application of CGHPRO in sub-array design

With the tiling path BAC array, the highest resolution that can be obtained is around 70 kb, which is not enough for some purposes. To further narrow down the borders of deletions or duplications, so-called 'sub-arrays' can be employed. These sub-arrays carry amplified probes that are distributed evenly along the breakpoint-spanning BAC clone. To facilitate the design of such arrays, I implemented a function called 'Subarray Design'. With this function, every specific breakpoint-spanning fragment can be divided into evenly distributed sub-regions, and for each of these, primer pairs will be designed.

4.14 Batch analysis

The segmentation step by either CBS or HMM is quite time-consuming. For 36K array, it takes DNAcopy about 1 hour to analyse one case on a normal PC. In order to make the analysis more efficient, I implemented a batch analysis tool in CGHPRO. Using this tool, all parameters for the different analysis steps can be set and then employed for the analysis of all relevant hybridisation results. The output is automatically stored in the database. Moreover, with the batch analysis running in the background, the computational task can be performed by several computers that belong to a network.

4.15 Availability of CGHPRO

CGHPRO is freely available for use under the terms of the GNU General Public Licences (GPL) at http://www.molgen.mpg.de/~abt_rop/molecular_cytogenetics/ArrayCGH/CGHPRO/. The open design of CGHPRO allows the easy adaptation to specific needs and the future incorporation of new features.