3 Molecular cytogenetics

Until the advent of molecular cytogenetic techniques, the analysis of genome rearrangements solely relied on the study of chromosome bands. Conventional high-resolution chromosome banding techniques as used in cytogenetic laboratories can yield up to 1000 bands per genome. At such resolution, banding patterns allowed the detection of aberrations greater than about 5 Mb and led to the description of deletion in several syndromes, such as DiGeorge syndrome and Prader-Willi syndrome.

However, the vast majority of disease-associated aberrations and structural variations result from submicroscopic chromosome rearrangements, which cannot be detected by chromosome banding. Moreover, using these techniques, it was often difficult to identify the origin of the chromosome fragments involved in complex translocations.

3.1 Fluorescence in situ hybridisation (FISH)

To improve the resolution of chromosome analysis, the development of FISH in the 1980s was an important step. FISH is based on the use of DNA probes labeled with fluorescent dyes, which can hybridize to their complementary sequences on the chromosomes, where they produce a fluorescent signal (Van Prooijen-Knegt et al., 1982). With probes designed to target specific regions of the genome, abnormalities could even be detected at the level of single genes. In many cases, the duplication, deletion or disruption of a single gene was subsequently found to be the cause of genetic diseases, the paradigm for this being hereditary neuropathy with liability to pressure palsies (HNPP), Charcot-Marie-Tooth (CMT1A) and hemophilia A.

Although FISH is a useful technique, the application of this technique requires prior knowledge about the type and location of expected aberrations and usually, only a limited number of chromosomal loci can be analyzed simultaneously.

3.2 Comparative Genomic Hybridisation (CGH) and array CGH

CGH is a molecular cytogenetic method for the detection of chromosomal imbalances, which does not depend on the availability of chromosome spreads and is not confined to the analysis of growing cells (du Manoir et al., 1993; Kallioniemi et al., 1992). The development of CGH yielded the first efficient approach to screen the whole genome for DNA copy number variations. Upon classical chromosome CGH, the genomic DNAs isolated from test (patient) and reference (control) samples are differentially labelled with two fluorescent dyes and are co-hybridized to normal human metaphase chromosomes on a microscope slide (see Figure 1 (McNeil and Ried, 2000)). Subsequently, CCD images of several metaphase spreads are captured and digital image analysis is used to quantify signal intensity for both fluorescent dyes. The signal intensity ratios of the test and reference hybridization are then calculated for a minimum of 5 metaphase spreads. Finally, an average ratio profile is plotted along the length of each chromosome, as shown in Figure 2 (McNeil and Ried, 2000). For deleted regions, the ratio will be below one, while it will be above 1 for amplified regions. Because conventional CGH allows detection and mapping of DNA sequence copy differences between two genomes in a single experiment and does not require dividing cells, it has become one of the most popular genome scanning technique.

Unfortunately, conventional chromosome CGH has a low resolution, which at best is in the order of 3 Mb (Kirchhoff et al., 1999). Since its development in 1990s, a great deal of effort has been devoted to improving the resolution of the technology. Recently, a major improvement could be achieved by the introduction of array CGH, a high-resolution variant of this technique, where differentially labelled test and reference DNA are co-hybridized onto microarrays of several thousand evenly spaced DNA clones or oligonucleotides representing specific regions of the human genome (Pinkel et al., 1998; Solinas-Toldo et al., 1997).

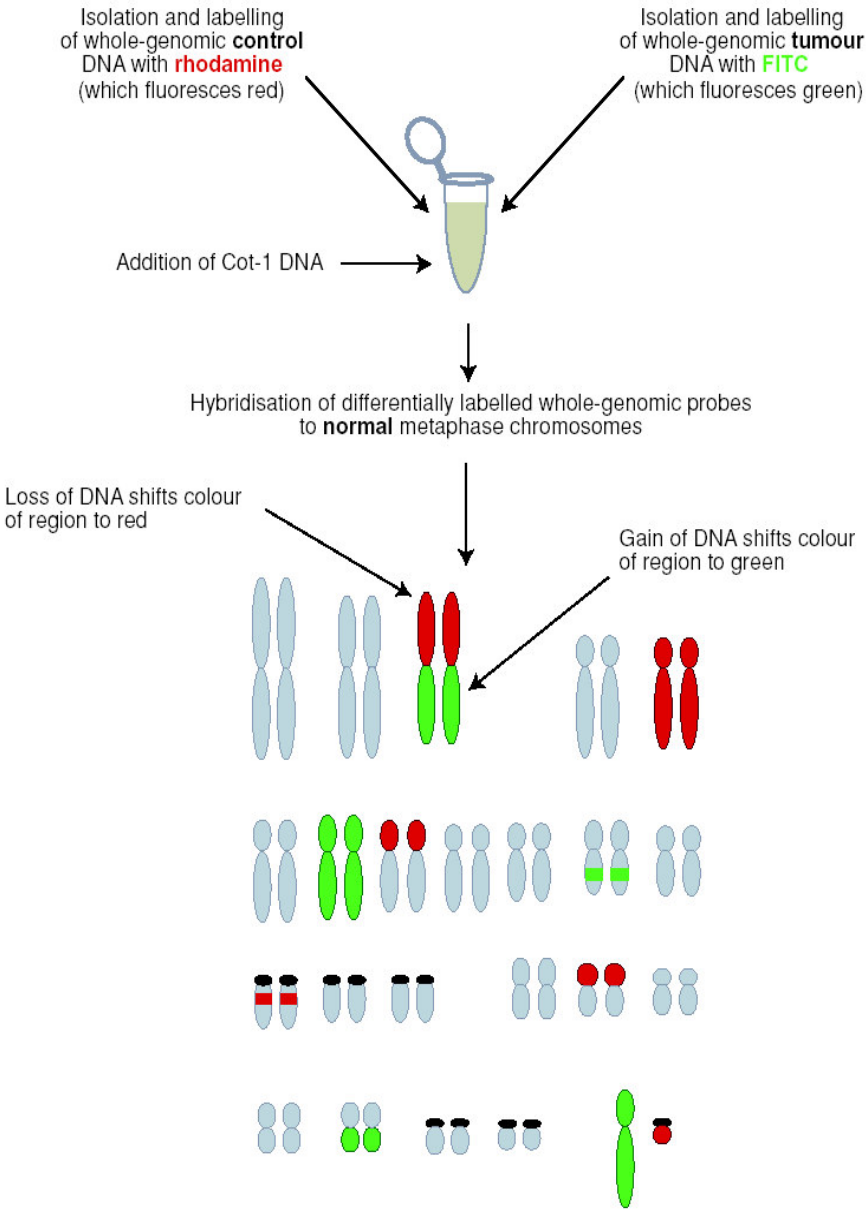

Figure 1: General procedure of Comparative Genomic Hybridisation (CGH) (McNeil and Ried, 2000). For classical CGH, the DNA from test sample and reference sample are differentially labelled. In this case, the test sample is labelled with a green fluorescent dye while the reference DNA is labelled red. Then these labelled DNA samples, are hybridised to normal metaphase chromosome, together with an excess of unlabelled Cot-1 DNA to suppress repetitive sequences. The relative intensities of the green and red fluorochromes reflect copy-number changes in the genome of test sample. DNA losses and gains are indicated by red and green fluorescence, respectively.

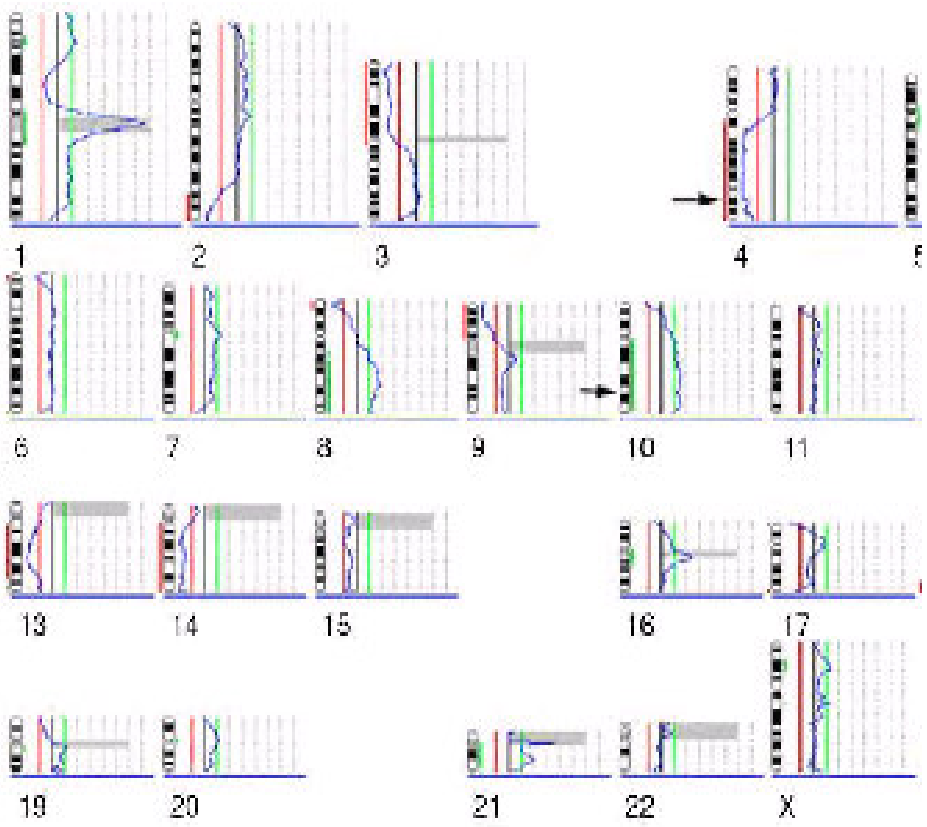

Figure 2: Comparative genome hybridisation (CGH) analysis of a lymph node metastasis from a renal cell carcinoma (McNeil and Ried, 2000). Tumour and reference sample were labelled with green and red fluorochrome respectively. Average ratios between tumor and reference sample were plotted along the ideogram of each chromosome. Red, gray and green vertical lines represented negative, zero and positive ratios. A chromosomal gain in the tumour was reflected by a stronger intensity of the green fluorescence, whereas a loss was indicated by a stronger intensity of the red fluorescence. The grey boxes in the profile represented chromosomal regions that were rich in heterochromatin, which could not be interpreted owing to the abundance of highly repetitive DNA. The prominent gains were at chromosome 10q, 3p, 9p and the most prominent losses could be seen at chromosome 4q and 13q.

The improved resolution as compared with chromosome CGH is based on replacing the metaphase chromosomes with DNA sequences spotted on the glass slides as the hybridisation target. Thus, the resolution of array CGH is only limited by the size and density of the spotted sequences. Theoretically, arrays can be constructed to cover any region of interest with any desired resolution. The general principle of array CGH is shown in Figure 3 (Oostlander et al., 2004).

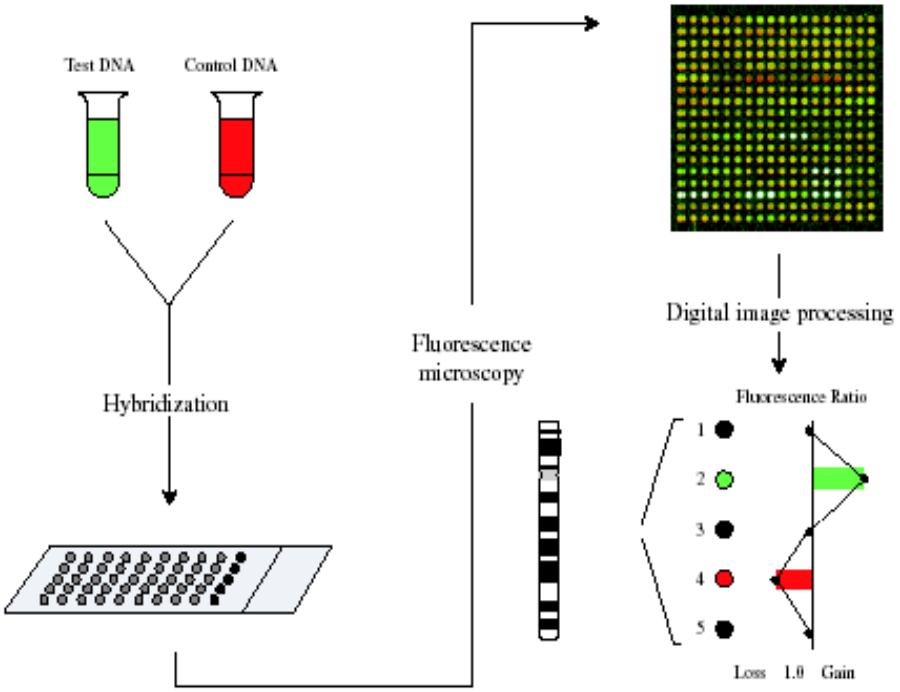

Figure 3: General principle of array CGH (Oostlander et al., 2004). The DNA from test (green) and reference sample (red) are differentially labelled and then hybridised to cloned DNA fragments spotted on the glass slide. Images of the fluorescent signals are captured and analysed. As a result, the gain region in the test sample will show high green signal and the loss region will have high red signal, while yellow spots indicate the presence of equal amount of test and reference DNA.

3.2.1 Experimental platforms for array CGH

Array CGH has been implemented using a wide variety of techniques. While their principle, i.e. detecting copy number differences between two samples, is the same, these platforms vary in terms of the size of the spotted elements and their coverage of the genome.

Originally designed for gene expression studies, cDNA microarrays can also be used in the analysis of copy number changes at the genomic level (Pollack et al., 1999). The first array CGH analysis of human cancer was performed using a cDNA microarray containing 3195 unique cDNA clones distributed throughout the genome (Pollack et al., 2002). A new generation of cDNA arrays have been spotted with exon-specific targets, allowing the detection of aberration in single exons (Dhami et al., 2005). Since the platform was originally designed for gene

expression studies, one advantage of this technique is that genomic aberration can be directly correlated to expression.

However, cDNA arrays do have several disadvantages. Firstly, cDNAs only cover the exonic region and thus alterations in other functional sites such as promoter region are not detectable. Secondly, the number of probes on the chip is limited to the genes that are encoded on the chromosomes; therefore these arrays do not provide continuous and even coverage of the genome. Finally, due to the smaller target size of cDNA clones compared with large-insert genomic clones, cDNA arrays usually have a low signal-to-noise ratio. Consequently, cDNA arrays perform poorly in detecting single copy number changes.

To obtain more intense hybridization signals, arrays spotted with DNA from large-insert genomic clones such as bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs) were used (Pinkel et al., 1998; Solinas-Toldo et al., 1997). The major advantages of the BAC/PAC arrays are the increased complexity of spotted DNA, which can improve the intensities of hybridization signals. Thus, the BACs/PACs platforms allow highly sensitive and reproducible detection of single-copy changes and accurate localization of the boundary of aberrations. Moreover, compared to cDNA arrays, BAC arrays are not limited to loci with annotated genes. Recently arrays carrying a overlapping set of BACs that cover the entire human genome have been constructed (Ishkanian et al., 2004; Krzywinski et al., 2004; Li et al., 2004). By using these 'tiling path' BAC arrays, imbalances of about 70 kb can be detected. The disadvantage of BAC/PAC arrays is that the preparation of sufficient DNA with adequate purity from BAC/PAC is rather laborious. Since the initial DNA yields of isolated BAC clones are low, an amplification step is necessary. Several amplification techniques have been explored, such as ligation-mediated polymerase chain reaction (PCR) (Snijders et al., 2001), degenerate oligonucleotides primer PCR (Telenius et al., 1992); (Hodgson et al., 2001) and rolling circle amplification (Smirnov et al., 2004). A further drawback of using a BAC/PAC platform is that inaccurate mapping information for some BAC/PACs can cause difficulties in data interpretation.

The latest approach is using arrays spotted with oligonucleotides such as the Affymetrix single nucleotide polymorphism (SNP) genotyping platform (Genechip human Mapping 10K/100K arrays) that has been applied in array CGH studies by Bignell et.al. (Bignell et al., 2004). The inherent problem of such arrays lies in the cross hybridisation of oligonucleotides (25 bp in length) to multiple genomic loci. To overcome this, the complexity of sample genomic DNA needs to be reduced before hybridization, which is achieved by a method called whole-genome sampling assay (WGSA). The WGSA assay is based on linker-mediated PCR of XbaI (or EcoRI or BglII)-digested genomic DNA, which only amplifies short restriction fragments (Figure 4) (http://www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf) and results in an enrichment of small restriction fragments throughout the genome (Kennedy et al., 2003). The strength of this platform is its ability to correlate copy number and allelic status at each locus. However, the resolution of such SNP genotyping platforms is limited by the uneven genomic distribution of SNPs that are targeted by the array. This results in an incomplete coverage of the genome. In addition, the necessary amplification of sample DNA may negatively influence the reproductivity of these experiments.

In order to improve hybridization specificity, oligonucleotides with increased length have been introduced (Barrett et al., 2004; Brennan et al., 2004; Carvalho et al., 2004). The representative oligonucleotide microarray analysis (ROMA) method initially used such an oligonucleotide array consisting of 85000 70-mers (Lucito et al., 2003). ROMA probes are designed to target the genomic representation created in a similar way as in WGSA. Assuming an even genomic distribution of the restriction sites used by the technique, ROMA can attain a resolution of 30Kb. Recently, two commercial platform with long oligonucleotide arrays have been introduced by Agilent (http://www.agilent.com/) and Nimblegen (http://www.nimblegen.com/). The Agilent platform consists of up to 200,000 60mer oligonucleotides which are synthesized *in situ*. The arrays provided by Nimblegen contain 385,000 oligonucleotides whose lengths are adjusted (45mer - 85mer) to equalize the melting temperature across the entire set. In theory, the resolution can be greatly improved using such high density oligonucleotide arrays. However, due to the low signal-to-noise ratio, these array platforms

usually require the calculation of a moving average to call single copy changes, which can decrease the effective resolution. Therefore, the merits of such platforms still await thorough experimental evaluation.

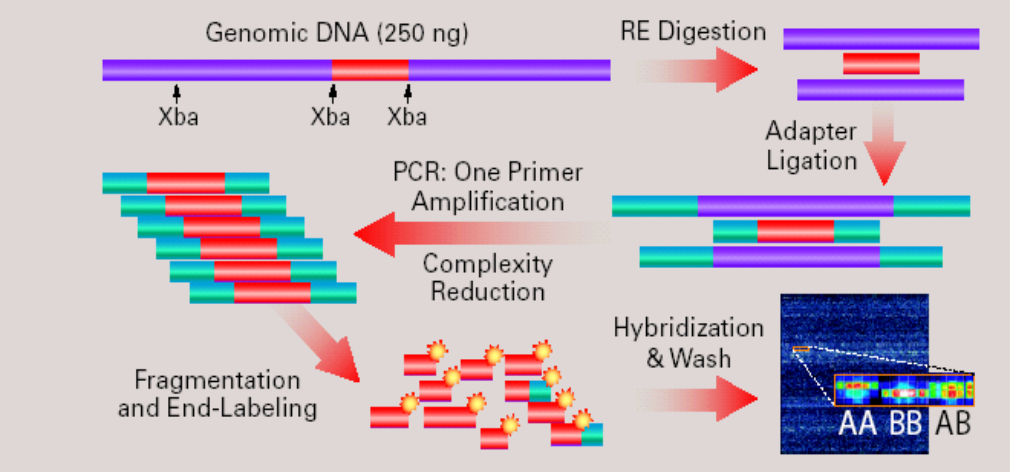

Figure 4: Genechip® mapping array overview.
(http://www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf)

The different array CGH platforms all have certain advantages and disadvantages, and even different implementations of the "same" array CGH approach may yield different levels of performance. So, the technical specification should be chosen carefully depending on the magnitudes of the copy number changes expected, their genomic extents, the state and composition of the specimen, amount of DNA available for analysis, and the required resolution. For example, DNA quantity may be limiting when analysing small biopsies, while DNA quality may be compromised in formalin-fixed, paraffin-embedded pathological samples. In such situations, large insert clone arrays, such as BAC arrays, have the advantage that they will produce readable signal even in samples of low DNA quantity and/or quality. When DNA quantity and quality are not limiting, arrays spotted with oligonucleotides or small PCR fragments may permit higher resolution than those carrying large insert clones.

3.2.2 Data Analysis of array CGH

In a typical array CGH study, after hybridisation, the slides need to be scanned at two wavelengths corresponding to emission spectral of the two fluorescent dyes, in this way, two monochromatic digital images are obtained, one for each dye. These images need to be further processed in order to estimate the copy number changes of test sample versus reference sample.

To extract an intensity for each spot on the array, the images have to be analysed by an image processing software such as GenePix Pro (http://www.moleculardevices.com/pages/software/gn_genepix_pro.html). A basic image analysis consists of three steps. First, each spot needs to be identified. This is usually accomplished by aligning a grid to the spots, because on an array, the spots are arranged in a grid of columns and rows. Once the spots are identified, they can be separated from background by using segmentation methods. Finally, the signal intensity is extracted for each spot and its surrounding background.

The raw signal intensities extracted by the software then need to be normalized. The goal of normalization is to remove any systematic bias in the measured fluorescence intensities such as differential labelling efficiencies, different scanning parameters, spatial bias, and print tip effects. Depending on the experimental design, a variety of normalization methods can be applied. Finally, the normalized data are used to identify the regions showing gains and/or losses. Although the major aberrations are frequently evident by visual inspection, many approaches to improve interpretation in the face of experimental noise have been developed. The common method used is to set thresholds, which are dependent on the variability of the data. If the distribution of the ratios falls into a few well-spaced intervals, the threshold can be easily chosen (Hodgson et al., 2001; Knuutila et al., 1998). However, sample heterogeneity and measurement noise often render the choice of a threshold not straightforward. Smoothing by averaging the ratios of neighbouring targets can alleviate the effect of noise, but at the same time this reduces the resolution and is sensitive to ‘outliers’.

Two important characteristics of array CGH data made the application of more sophisticated algorithms necessary. First, the copy number changes involve chromosome segments. Therefore, when determining copy numbers along the chromosome one should observe segments of equal copy numbers with sudden jumps and occasional single-probe outliers (Bredel et al., 2005). Second, chromosomal proximity and/or overlap as in the case of BAC clones, contributes to correlations of true copy numbers for successive sites. Therefore, the major algorithm problem to be solved in array CGH data analysis is how to segment the array elements which are ordered along the chromosome as shown in Figure 5, into sets with equal copy numbers and to assess the status of each element in the context of its neighbours. The approaches resulting from prior work include Hidden Markov Model (Fridlyand et al., 2004), change point analysis (Olshen et al., 2004), adaptive weights smoothing (Hupe et al., 2004), Bayesian maximum *a posteriori* probabilities (Daruwala et al., 2004) and ratio clustering (Wang et al., 2005)

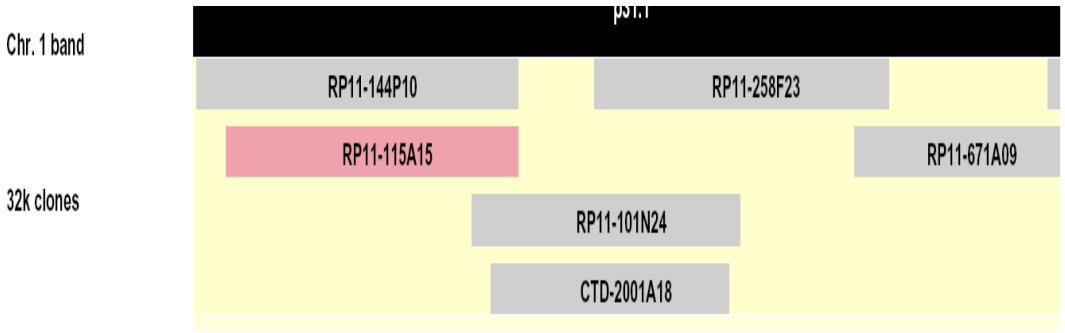

Figure 5: Possible configuration of BACs on a chromosome.

Up to now, array CGH has been predominately used in highly specialized laboratories, and most of the data analysis programs currently available are not able to process the output of array CGH experiments in an easy and comprehensive way. For example, the two R packages from Bioconductor (http://www.bioconductor.org), aCGH and DNAcopy, can identify copy number transitions on chromosomes by using an Unsupervised Hidden Markov Model and Circular Binary Segmentation, but the application of these tools requires basic programming skills in the R language. CGH-Plotter is a MATLAB toolbox with a graphic user interface (Autio et al., 2003; Chi et al., 2004). It detects the regions of amplifications and deletions using k-means clustering and dynamic

programming. However, like aCGH and DNAcopy, CGH-Plotter can only be used to analyse already normalized array data in a specific format**.** In addition, these programs can display the results only in a non-interactive plot. SeeGH, on the other hand is a tool which displays the data in a user friendly interface (Chi et al., 2004). It allows users to explore the results in a conventional karyotype diagram with annotation. However, without the essential statistical methods for characterizing the genomic profile, seeGH is not particularly useful for array CGH data analysis. ArrayCGHbase (Menten et al., 2005) and CAPweb (http://bioinfo-out.curie.fr/Capweb) are two web-based applications that consist of the routines to cover the process from normalization to aberration characterization, but the use of these online analysis tools is heavily dependent on server capacities and the speed of data transfer. In addition, in diagnostic and related applications, online data analysis is precluded due to privacy requirements.