Aus dem Max Planck Institut für molekulare Genetik

DISSERTATION

Development and application of CGHPRO, a novel software package for retrieving, handling and analysing array CGH data

Zur Erlangung des akademischen Grades
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

Vorgelegt von

Wei Chen

aus Fujian, V. R. China

Juli, 2006

1st Reviewer: _Prof. Dr. H.-Hilger Ropers__

2nd Reviewer: _Prof. Dr. Gerd Multhaup ___

Date of defence: Oct. 27, 2006

Acknowledgment

First, I would like to thank Prof. Dr. Hans-Hilger Ropers for guiding me into the fascinating research field of Human Genetics, and for his supervision throughout my project. I benefit from the discussion with him very much, which is of vital importance for my future career development. I am indebted to Professor Dr. Gerd Multhaup for being my adviser in Free University Berlin.

I would like to thank Dr. Reinhard Ullmann for stimulating discussions and valuable suggestions, which helped in understanding the technical details in molecular cytogenetic experiments, and for giving critical comments on my thesis.

Dr. Andreas W. Kuss deserves my thanks for his expert advice, and for thoroughly reading my thesis.

.

Thanks are also due to members of Dr. Ullmann’s group, especially Dr. Fikret Erdogan, whose experimental work, which produced the basis for my work.

I would like to acknowledge my colleagues, Dr. Lars Riff Jensen, Dr. Andreas Tzschach, Dr. Steffen Lezner, Melanie Wendehack, Masoud Gashasbi, Mohammad Mahdi Motazacker, Lia Abbasi-Moheb, Bettina Lipkowitz, Marianne Schlicht and Marion Amende-Acar, for creating a pleasant working atmosphere throughout my project.

Finally, I want to express my deep appreciation to my families. My wife, Yuhui, her continuous love, care and sweetness has made my life full of joy and interests. My parents, my parents-in-law, and my sister, only your continuous support over all these years enabled me to get this far.

Contents

1 Preface 1

2 Genome rearrangement 2

3 Molecular cytogenetics 4

- 3.1 Fluorescence in situ hybridisation (FISH) 4
- 3.2 Comparative Genomic Hybridisation (CGH) and array CGH 5
  - 3.2.1 Experimental platforms for array CGH 8
  - 3.2.2 Data Analysis of array CGH 12

4 Development of CGHPRO 15

- 4.1 Software development 15
- 4.2 Notation of array CGH data 15
  - 4.2.1 Intensities 16
  - 4.2.2 Ratios 16
  - 4.2.3 Log-intensities and log-ratios 16
- 4.3 Overview of the data analysis process in CGHPRO 16
- 4.4 Database design 17
- 4.5 Data input 18
- 4.6 Graphical analysis of hybridisation characteristics 19
  - 4.6.1 Scatter plot 19
  - 4.6.2 MA-plot 19
  - 4.6.3 Boxplot 21
  - 4.6.4 Histogram 21
- 4.7 Data filter 22
- 4.8 Data normalization 22
  - 4.8.1 Within-slide normalization 23
    - 4.8.1.1 Global normalization 23
    - 4.8.1.2 Intensity dependent normalization 23
    - 4.8.1.3 Printing tip specific normalization 23
  - 4.8.2 Dye-Swap normalization 24
- 4.9 Replicate spots handling 25
- 4.10 Characterization of genomic profiles 25
- 4.11 Data display 26
  - 4.11.1 Genomic display 26
  - 4.11.2 Chromosome display 29
- 4.12 Comparative analysis of different chips 29
- 4.13 Application of CGHPRO in sub-array design 31
- 4.14 Batch analysis 31
- 4.15 Availability of CGHPRO 31

5 Application of array CGH and CGHPRO 32

- 5.1 Impact of segmental duplication on the generation of genome rearrangements . 32
  - 5.1.1 Molecular mechanisms underlying genome rearrangements 32
    - 5.1.1.1 Segmental duplication-mediated nonallelic homologous recombination 33
    - 5.1.1.2 Other genome architectural features 34

5.1.2 Copy number changes in 22 patients with mental retardation................ 35
5.1.3 Overlap of breakpoints in unbalanced aberrations with segmental duplication and CNPs 39
5.1.4 Overlap of breakpoints in balanced translocation with segmental duplication and CNPs 44
5.1.5 Segmental duplication and CNPs in balanced and unbalanced rearrangements.. 46
5.1.6 Segmental duplication could mediate non-allelic homologous recombination (NAHR) 48
5.1.7 NAHR involving segmental duplications cannot explain all unbalanced rearrangements.. 49
5.1.8 NAHR involving segmental duplications cannot explain balanced rearrangements.. 50
5.1.9 Clusters of segmental duplications may cause chromosomal instability 51
5.2 Mapping of balanced translocation breakpoints using array CGH and CGHPRO.. 52
5.2.1 Cytogenetic investigation of the translocation 53
5.2.2 Array CGH investigation of the translocation....................................... 54
5.2.3 Confirmation of the chromosome breakpoints by FISH analysis........... 55
5.2.4 PCR fragment subarray 55
5.2.5 Long-range PCR and sequencing.. 56
5.2.6 Advantage of array painting in mapping balanced translocation breakpoints.. 58
6 Summary 59
7 Zusammenfassung 61
8 Reference 63
9 Supplemental material 70
9.1 Experimental protocols.. 70
9.1.1 Array CGH 70
9.1.1.1 Array production.. 72
9.1.1.2 DNA labelling and hybridisation 74
9.1.1.3 Scanning.. 75
9.1.2 Mapping balanced translocation breakpoint by chromosome sorting and array painting.. 75
9.1.2.1 Cell culture and preparation.. 75
9.1.2.2 Flow karyotyping and sorting 76
9.1.2.3 DNA amplification using GenomePlex 76
9.1.2.4 Array painting using 36K BAC array 76
9.1.2.5 Fluorescence in situ hybridisation.. 76
9.1.2.6 PCR fragment subarray.. 77
9.1.2.7 Long-range PCR and sequencing.. 82
9.1.2.8 Agarose gel electrophoresis 83
9.2 Supplementary tables 84
***Curriculum Vitae* 95**

1 Preface

Comparative genomic hybridisation using arrays of DNA clones as targets (array CGH) is a novel and powerful technique to identify submicroscopic deletions and duplications and to study their roles in genetic disorders. Typically, very high resolution CGH arrays covering the whole human genome comprise >30,000 overlapping Bacterial Artificial Chromosomes (BAC) clones and yield a corresponding number of discrete hybridisation signals. The management and interpretation of these data poses enormous problems, not only because of their quantity, but also because of their variable quality. Recently, CGH arrays comprising 36,000 BAC clones have been generated at Max Planck Institute for Molecular Genetics, which are being employed for deletion/duplication screening in patients with mental retardation and various related disorders. As a prerequisite for these studies, I have developed a comprehensive software package for visualisation, analysis and management of array CGH data. The program, called ‘CGHPRO’, is also designed to support the search for genomic imbalances that are only seen in specific cohort of patients, and even more importantly, it tracks previously reported functional neutral genomic imbalances.

The second part of my project focused on the practical application of high-resolution array CGH and CGHPRO. First, by means of array CGH, copy number changes in 22 patients with mental retardation were analysed. In order to obtain insights into the molecular mechanisms of genome rearrangements, especially the impact of segmental duplications, I investigated the chromosomal breakpoint regions of these 22 patients in more detail. These data were supplemented with fine mapping data of another 41 cases with balanced translocations. Implementation of further features into CGHPRO, which allowed the automatic design of specific sub-arrays, paved the way for high-resolution array-painting. This technique combines chromosome sorting and DNA array technology and enables rapid fine mapping of breakpoints in balanced translocations.

2 Genome rearrangement

Human chromosomes were first observed more than 120 years ago and since then technical innovations have paved the way for progress in this field. In 1923, the study of dividing testicular cells led Thomas Painter to conclude that humans have 48 chromosomes, and the correct number of 46 was only determined in 1956, after Tjio and Levan (Tjio and Levan, 1956) had developed an improved protocol for the preparation and spreading of chromosomes. This innovation was also instrumental in defining a variety of diseases that are due to aberrant number of chromosomes, so-called numerical chromosome aberrations, such as Down Syndrome (trisomy 21), Klinefelter syndrome (47, XXY) and Turner syndrome (45, XO).

In the late 1960s and 1970s, staining protocols were developed, generating specific banding patterns along the length of each chromosome. These banding patterns allowed to distinguish all human chromosomes and greatly facilitated the recognition of chromosome structural rearrangements.

Such structural rearrangements occur when a chromosome breaks and is rejoined to another broken chromosome fragment. They can be confined to a single chromosome, resulting in a loss or gain of material (deletion/duplication), or leading to the inversion of an internal chromosome segment. If fragments of two different chromosomes are exchanged, this will result in reciprocal translocations.

Genome rearrangements play an important role in the etiology of human genetic diseases. The term ‘genomic disorder’ has been coined for a broad spectrum of diseases caused by the rearrangement of specific genomic segments , ranging in size from a few kilobases to several megabases (Lupski, 1998); (Inoue and Lupski, 2002). This group of disorders does not result from single nucleotide substitutions, but is due to recurrent chromosomal aberrations which give rise to DNA copy number changes or disruption of the structural integrity of a dosage sensitive gene(s). Very often, in these disorders, the underlying recurrent genome rearrangements are mediated by nonallelic homologous recombination between highly similar paralogous sequences.

Apart from causing a variety of well-known genomic disorders, such as DiGeorge Syndrome, Williams-Beuren Syndrome and Prader-Willi Syndrome, some of these genome rearrangements are also observed in the normal population and are considered as functionally neutral structural variants. Major types of structural variants consist of copy number polymorphisms (CNPs) and inversion polymorphisms, respectively.

Simple nucleotide substitutions have been implicated in many genetic diseases, but the majority of these are considered as functionally neutral variants. In contrast, small genome rearrangements have only recently been appreciated as an important source of genetic variation. Apart from genome rearrangements directly causing genetic diseases, other may modulate the predisposition for specific disorders. Since the early 1990s, the development of suitable tools for their detection has bridged the gap between karyotyping and molecular genetics and opened the new field of "Molecular Cytogenetics".

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).

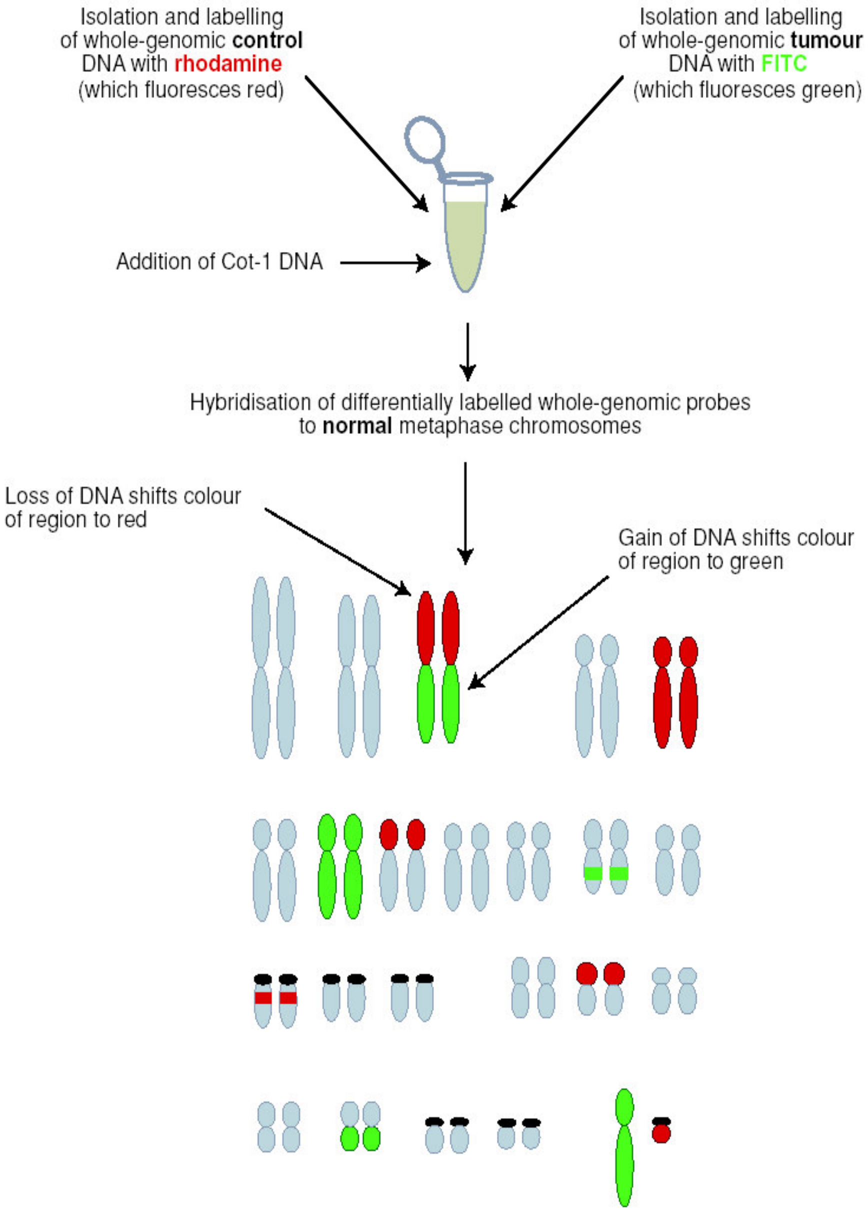

Isolation and labelling of whole-genomic **control** DNA with **rhodamine** (which fluoresces red)
Isolation and labelling of whole-genomic **tumour** DNA with **FITC** (which fluoresces green)
Addition of Cot-1 DNA
Hybridisation of differentially labelled whole-genomic probes to **normal** metaphase chromosomes
Loss of DNA shifts colour of region to red
Gain of DNA shifts colour of region to green

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.

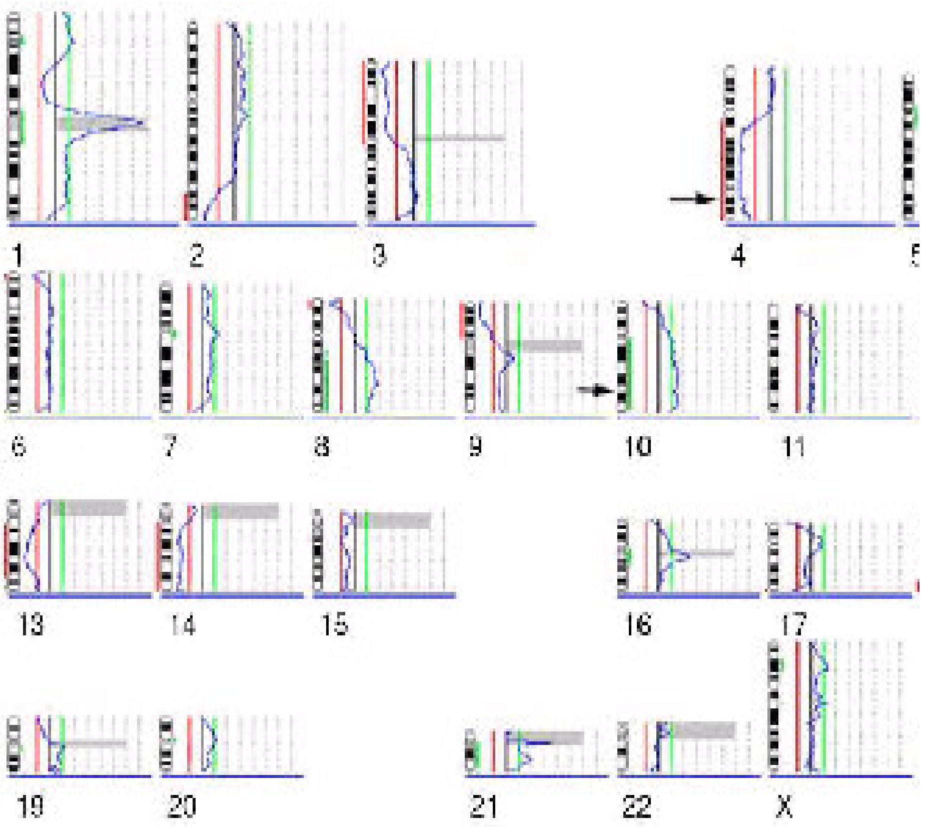


Figure 2: Comparative genome hybridisation (CGH) analysis of a lymph node metastasis from a renal cell carcinoma (McNeil and Ried, 2000). Tumour and reference sample were labelled with green and red fluorochrome respectively. Average ratios between tumor and reference sample were plotted along the ideogram 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).

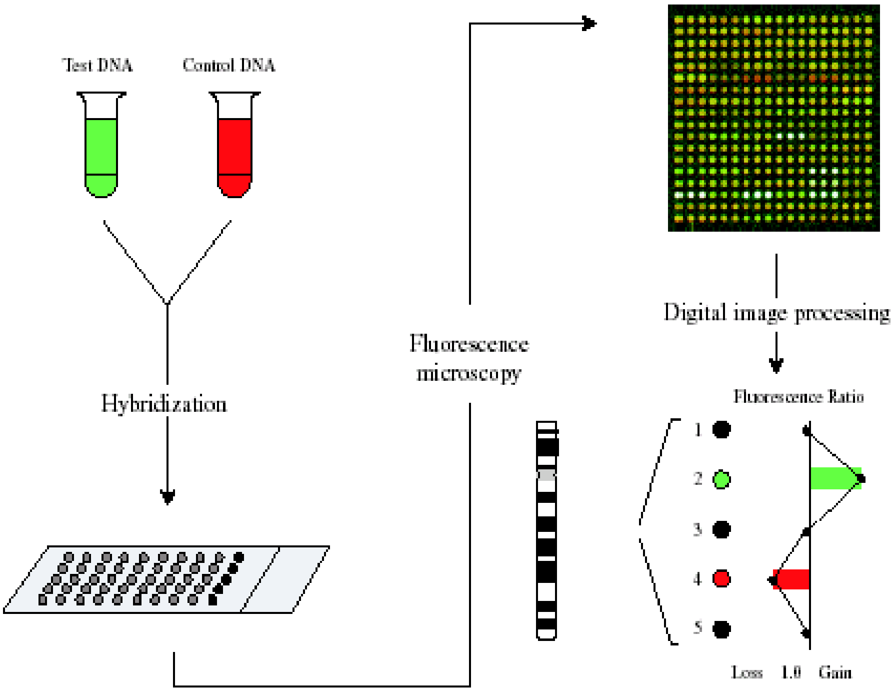


Figure 3: General principle of array CGH (Oostlander et al., 2004). The DNA from test (green) and reference sample (red) are differentially labelled and then hybridised to cloned DNA fragments spotted on the glass slide. Images of the fluorescent signals are captured and analysed. As a result, the gain 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.

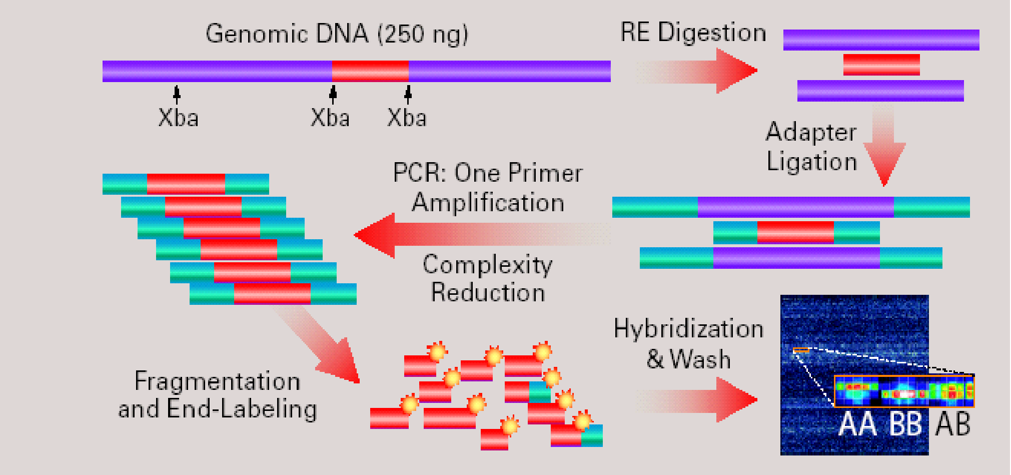


Figure 4: Genechip® mapping array overview.

(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)

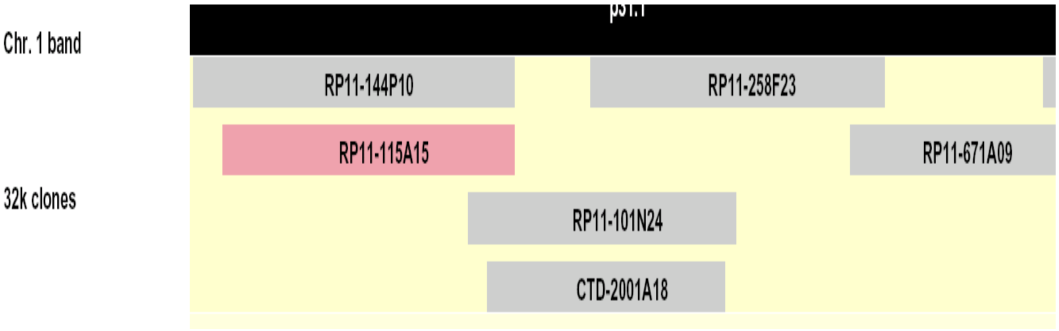


Figure 5: Possible configuration of BACs on a chromosome.

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.

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

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

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.

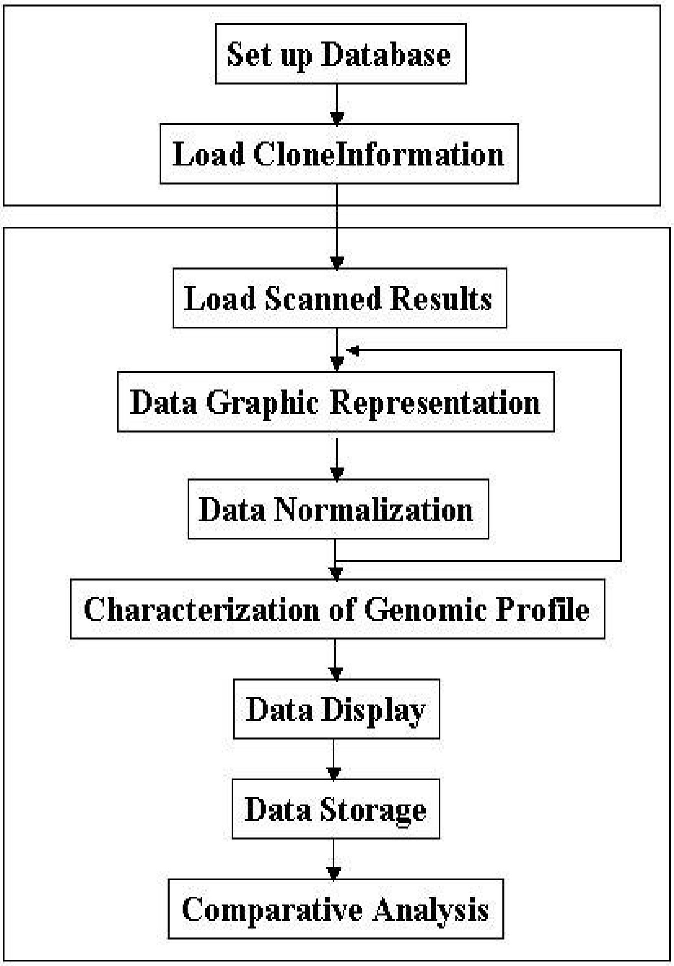


Figure 6: Overview of data analysis process in CGHPRO

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).

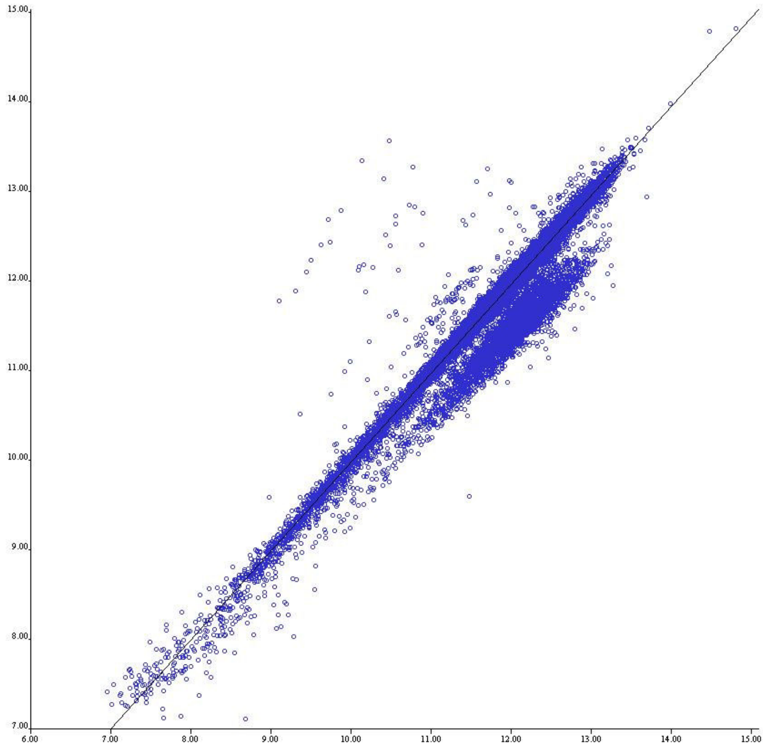


Figure 7: Scatterplot. In a scatterplot, the log-intensity of the red dye (Y axis) is plotted against the log-intensity of the green dye (X axis).

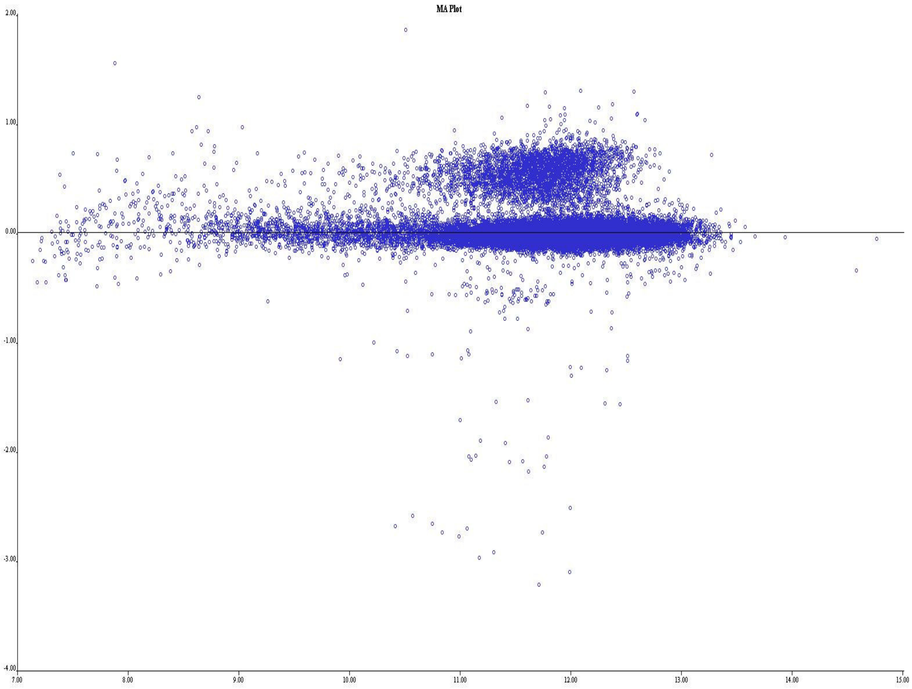


Figure 8 : MA-plot. The X-axis conforms to the logarithm of the average intensity value of the two dyes; the Y-axis shows exactly the log-ratio of the two dyes

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).

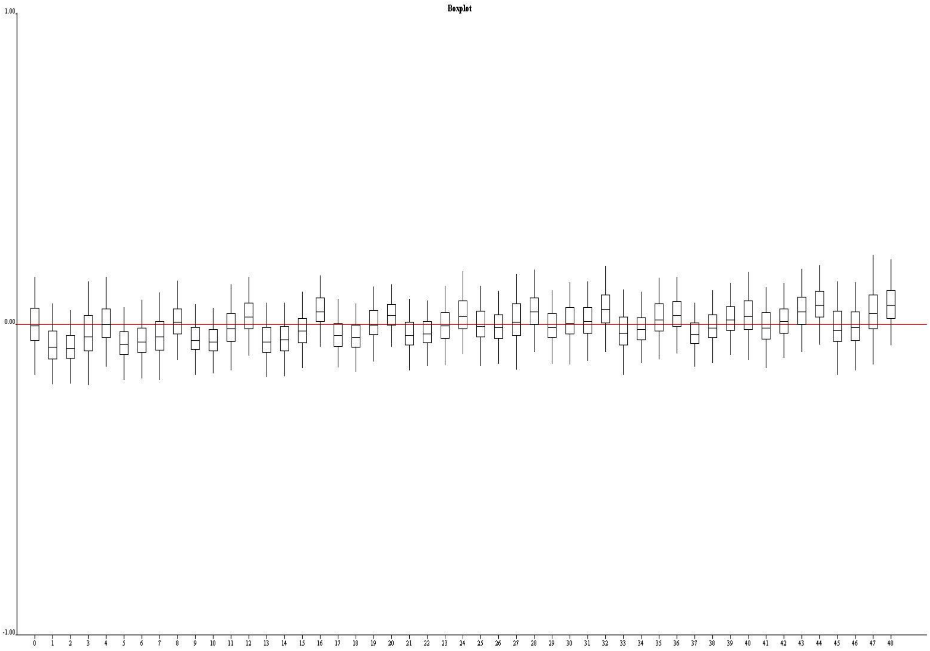


Figure 9: Boxplot. The box in the middle represents the interquartile range (IQR). The median is marked as line in the middle of box while the whiskers show the spread of the data. Here, each boxplot is plotted for one subgrid.

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).

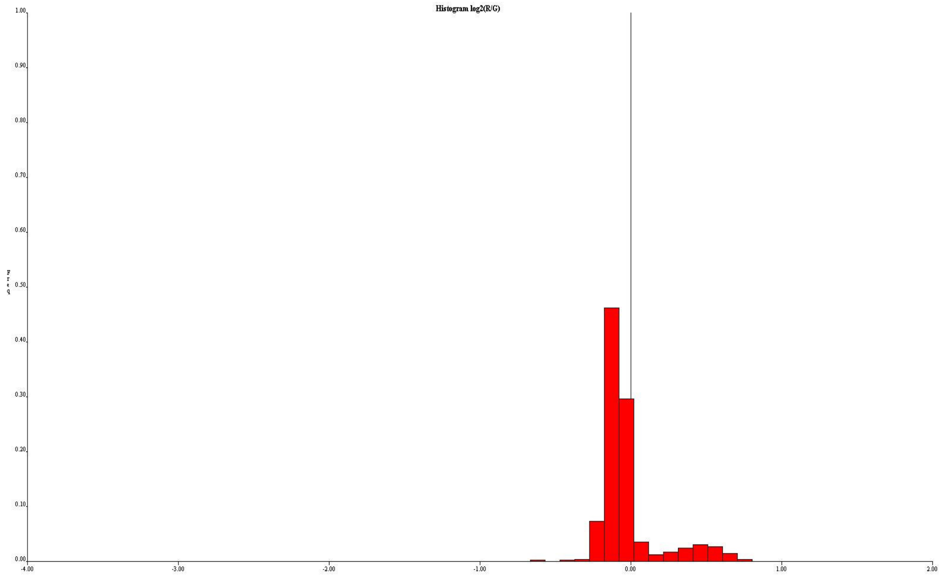


Figure 10: Histogram. The distribution of log-ratios for a single slide is shown here.

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.

$$\log_2 \frac{R_i}{G_i} \rightarrow \log_2 \frac{R_i}{G_i} - c = \log_2 \frac{R_i}{kG_i}$$

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.

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

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

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

Another possibility is to correct the single intensity values. Calculating

$$k=\sqrt{\frac{R_iG'_i}{R'_iG_i}}$$

and correcting the intensity values with this factor k

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

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.

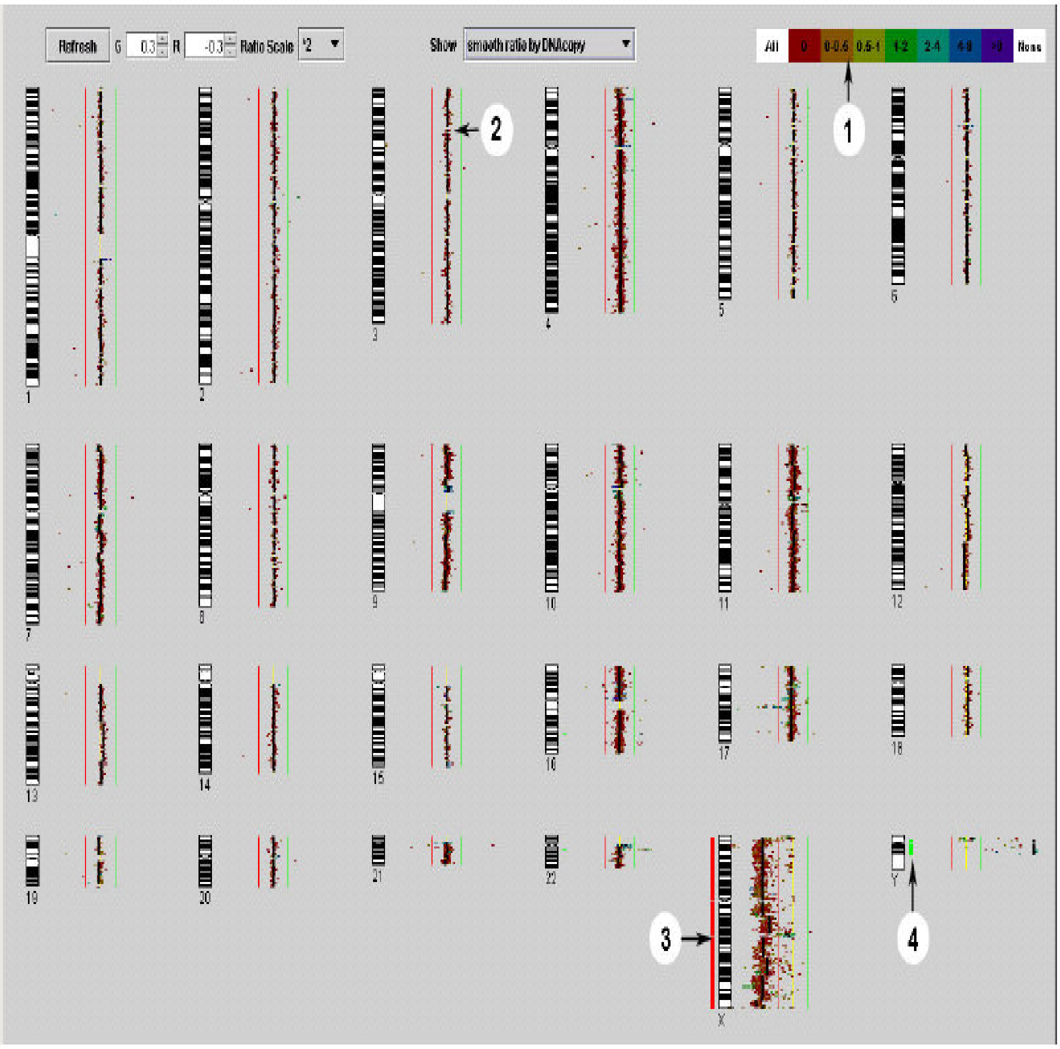
4.11 Data display

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.



G 0.3
R -0.3
Ratio Scale *2
Show smooth ratio by DNAcopy

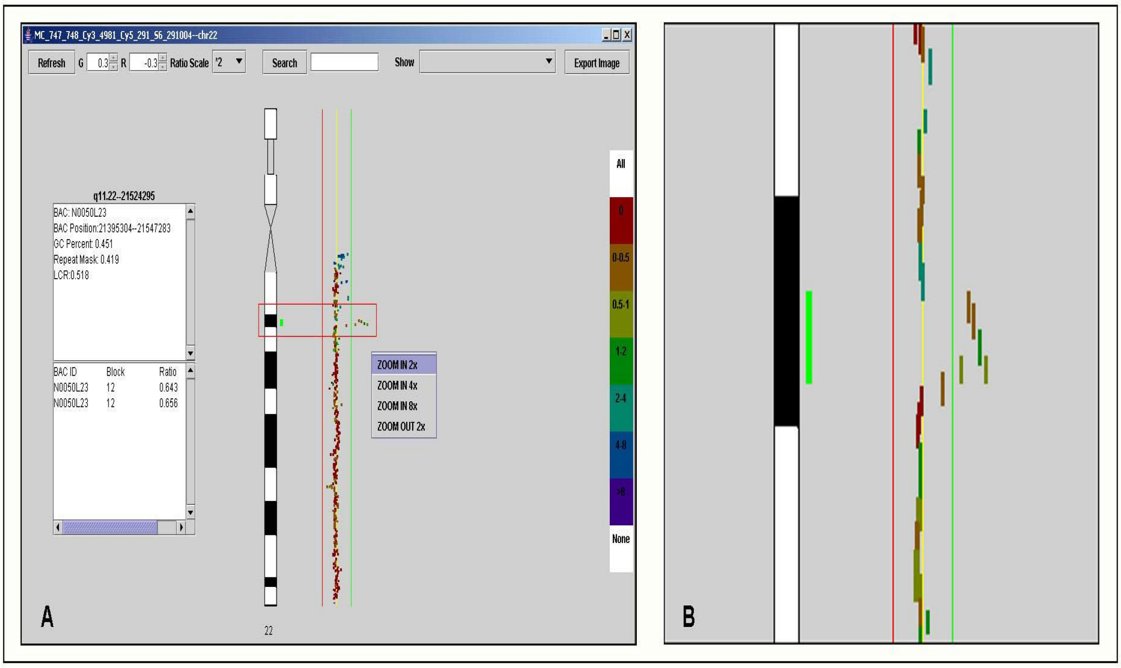
Figure 11: Genome Display exemplified by male versus female hybridization on a 14000 BAC DNA array. Circles 1-4: (1) Colour coding table indicating the involvement of clones in segmental duplication (2) Black line representing the smoothed ratios calculated by DNAcopy (3) and (4) red and green bars to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



G 0.3 R -0.3 Ratio Scale *2

q11.22--21524295

BAC: N0050L23
BAC Position:21395304--21547283
GC Percent: 0.451
Repeat Mask: 0.419
LCR:0.518

BAC ID	Block	Ratio
N0050L23	12	0.643
N0050L23	12	0.656

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).

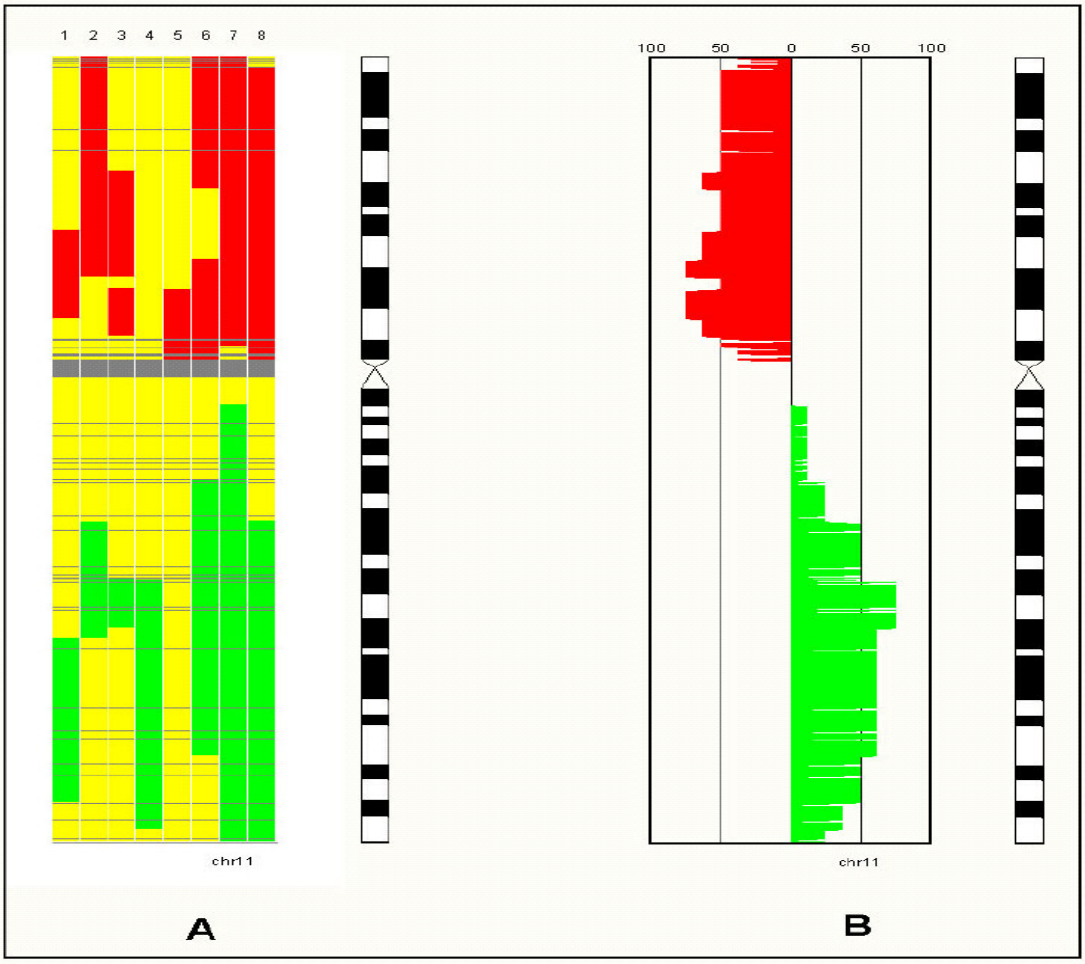


Figure 13:Comparative Analysis: CGHPRO supports the visualization of absolute (A) and relative (B) frequencies of chromosomal aberrations in a series of cases. Results can be displayed simultaneously for all or for single chromosomes, as shown here for chromosome 11.

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.

5 Application of array CGH and CGHPRO

5.1 Impact of segmental duplication on the generation of genome rearrangements

5.1.1 Molecular mechanisms underlying genome rearrangements

Genome rearrangements result from double-strand breaks (DSBs) that arise spontaneously during DNA replication or can be induced by ionizing radiation or chemicals (including anticancer drugs). DSBs are critical lesions which, if not repaired, may be lethal for the affected cell. So a number of cellular DNA repair mechanisms have evolved for the restoration of break sites. In eukaryotes, two major pathways have been identified that differ in their requirements of DNA homology. DSB repair by homologous recombination (HR) requires the presence of homologous sequences elsewhere in the genome (e.g. a homologous chromosome or a sister chromatid). In contrast, non-homologous end joining (NHEJ) fuses the two ends of a DSB through a process that is largely independent of terminal sequence homology and therefore can join ends with diverse chemical and physical characteristics. Both HR and NHEJ have been conserved during evolution, but vary in the contribution to overall DSBs repair in lower and higher eukaryotes. Generally speaking, while HR predominates in lower eukaryotes, DSB in mammals are primarily repaired by NHEJ. Furthermore, their relative contribution varies during development and depends also on the stage of the cell cycle: while NHEJ is active throughout the cell cycle, HR is limited to the late S and G2 phase. Although DSBs repair by either HR or NHEJ is normally efficient and precise, occasional errors can occur in the repair process and thus lead to genome rearrangements.

Regardless of the fact that chromosome rearrangements occur everywhere in the genome, they predominate in the intervals with a complex genomic architecture, such as segmental duplications and AT-rich palindromic repeats. This suggests that genome rearrangements are not random events, but rather result from chromosome instability that is due to the local genomic architecture (Shaw and Lupski, 2004).

5.1.1.1 Segmental duplication-mediated nonallelic homologous recombination

Segmental duplications are large, nearly identical copies of genomic DNA, which range in size from 1 to >200 kb and are present at two or more positions in the human genome. It has been estimated that 5% of the human genome are composed of such duplications, which are clustered in the pericentromeric transition zones, the subtelomers and several interspersed LCR hubs (Bailey et al., 2002; Bailey et al., 2001; Bailey et al., 2002; Cheung et al., 2003; Cheung et al., 2001; Eichler, 2001; Horvath et al., 2001). Many of the segmental duplication in the human genome appear to have arisen during primate speciation. It has been hypothesized that these duplications can drive adaptive evolution by generating new genes. This hypothesis is supported by a variety of studies which have shown DNA copy number changes between human and non-human primates (and Analysis ConsortiumThe Chimpanzee, 2005; Fortna et al., 2004; Fujiyama et al., 2002; Locke et al., 2003; Newman et al., 2005; Wilson et al., 2006; Yunis et al., 1980). Although segmental duplications may be important in an evolutionary sense, their existence poses a risk to the individual human genome, as their highly homologous sequences provide ample substrates for non-allelic homologous recombination (NAHR). As shown in Figure 14, segmental duplication-mediated NAHR can lead to deletions, duplications or inversions, depending on the orientation (direct/inverted) of the duplicated sequences and the involvement of interchromosomal, intrachromosomal or intrachromatid recombination (Stankiewicz and Lupski, 2002). In addition, NAHR between different chromosomes can also result in chromosomal reciprocal translocation. The probability of meiotic misalignment between duplicated sequences may depend on several factors—including length, sequence identity and orientation as well as the distance between duplications.

Recently, segmental duplication-mediated NAHR has been directly implicated in a growing list of recurrent genomic disorders. Similarly, there is increasing evidence that the duplication architecture of the genome may also mediate structural variation in the normal population. In the study of Tuzun et al (2005), more than half of the detected variant sites (163 of 297) map to regions with segmental duplications. The association was most pronounced for the intrachromosmal segmental duplications where the degree of sequence identity exceeds 98% (Tuzun et al., 2005).

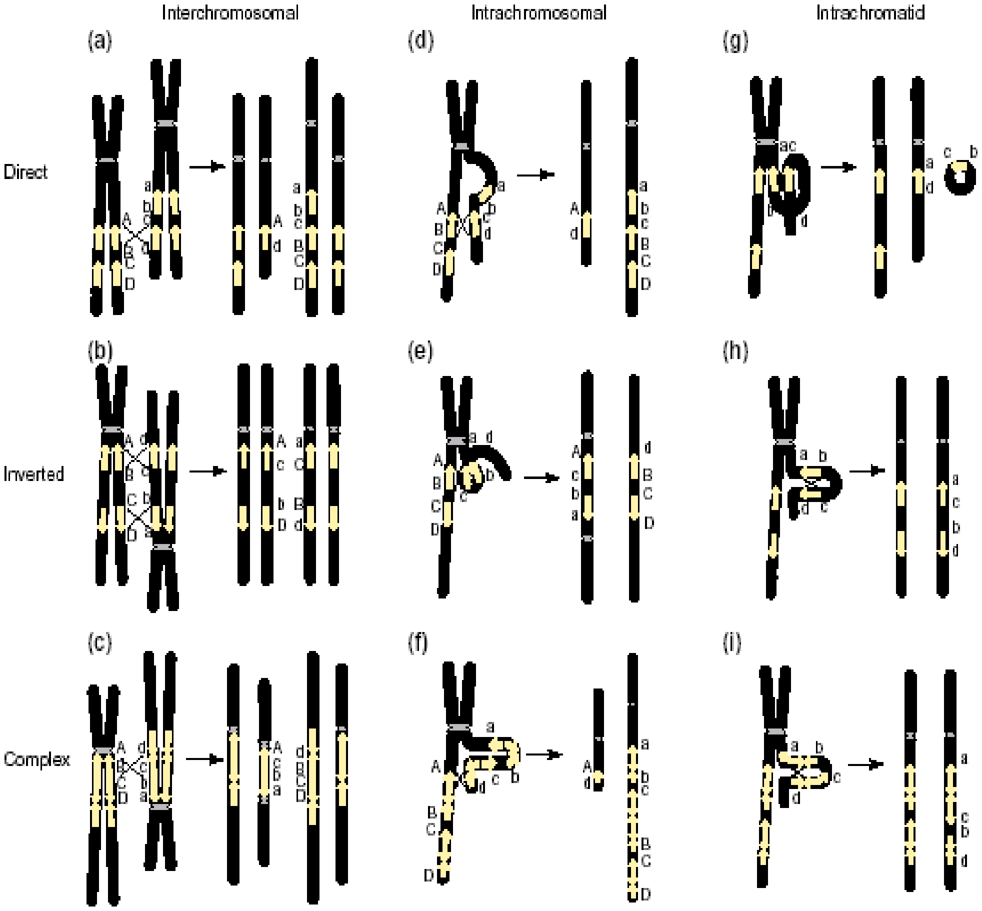


Figure 14: Mechanisms of genome rearrangements resulting from segmental duplications mediated NAHR (Stankiewicz and Lupski, 2002). Chromosomes are shown in black with the centromere depicted in gray line. Yellow arrows represented segmental duplications with specific orientation. All possible rearrangements mediated by segmental duplications are grouped horizontally by orientation and structure of segmental duplications (direct, inverted, complex), and vertically by the mechanisms (interchromosomal, intrachromosomal, intrachromatid).

5.1.1.2 Other genome architectural features

Segmental duplication-mediated NAHR cannot explain all cases of genome rearrangements. Other mechanisms such as NHEJ have been observed, particularly for rearrangements with scattered breakpoints (Roth and Wilson, 1986). Very often, complex genome architectural features are also involved. A systematic study of deletion junctions in the gene for Duchenne muscular dystrophy (DMD), revealed Alu and long tandem repeat (LTR) elements in 3 out of 10 cases (Nobile et al., 2002). The sequence TTTAAA, which is known to bend the DNA molecule (Singh et al., 1997), was found at or near 3 of the junctions examined. In many translocation cases, AT-rich palindromes were found at the break points on the derivative chromosomes (Gotter et al., 2004; Kurahashi et al., 2003; Nimmakayalu et al., 2003). This suggests the possibility of secondary structures based on AT palindromes causing double strand breaks. In addition to that, centromeres, pericentromeric repeats and telomers are often implicated in non-recurrent breakpoints. Their involvement indicates that the chromatin structure can also play a role in genome rearrangements.

In this study, array CGH has been employed to study the impact of segmental duplications on the generation of both balanced and unbalanced genomic rearrangements. For this purpose, a set of 22 mentally retarded patients were examined, which has been pre-selected for the presence of chromosomal aberrations, and the results were compared with FISH mapping data from 41 mentally retarded patients with balanced translocations.

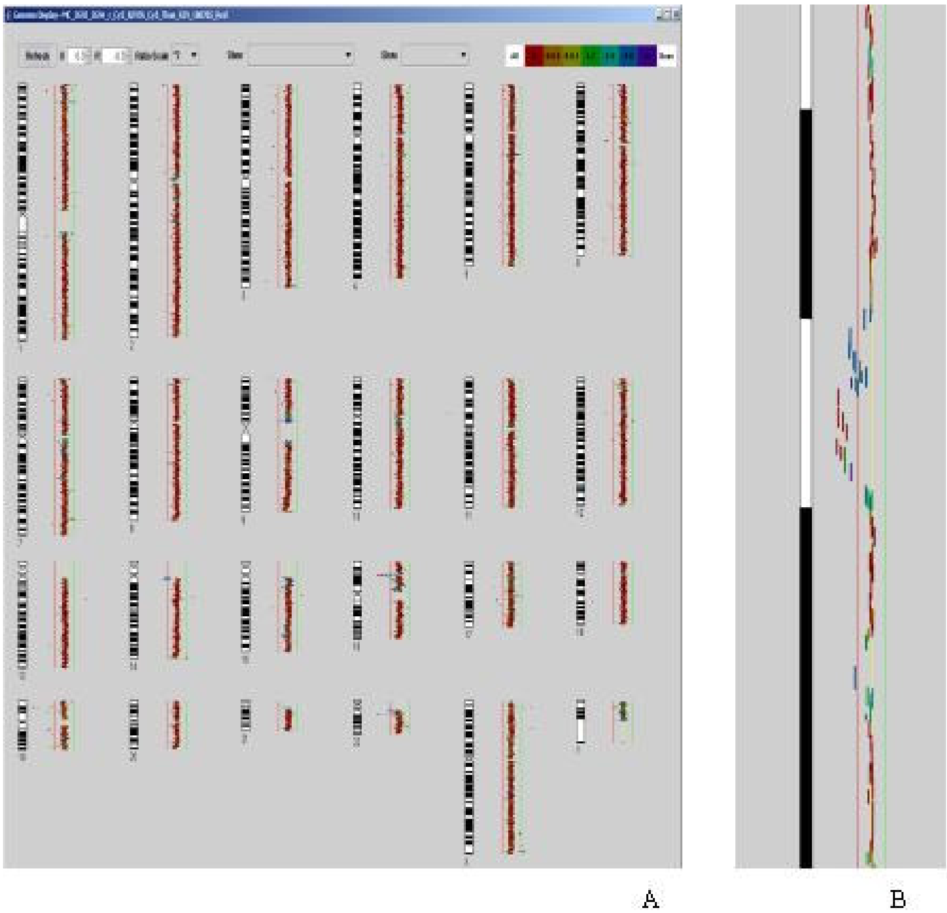
5.1.2 Copy number changes in 22 patients with mental retardation

Array CGH was carried out for 22 patients with mental retardation. In all but four cases, the imbalances have been analysed and verified by HR-CGH (Kirchhoff, et al., 1999, Kirchhoff, et al., 2004). As controls, three patients with the known genomic disorders, Smith-Magenis Syndrome (SMS) (case16), Prader-Willi/ Angelman Syndrome (case15) and 22q11deletion syndrome (case7), were included.

For array CGH, a high resolution tiling path BAC array was used, comprising the human 32k Re-Array set (Krzywinski, et al., 2004; Osoegawa, et al., 2001; Ishkanian, et al., 2004), http://bacpac.chori.org/pHumanMinSet.html: (DNA kindly provided by Pieter de Jong), the 1Mb Sanger set (Fiegler, et al., 2003) (clones kindly provided by Nigel Carter, Wellcome Trust Sanger Center) and a set of 390 subtelomeric clones (assembled by members of the COST B19 initiative: Molecular Cytogenetics of solid tumors). Cases 5, 7, 8 and 16 were hybridised on a 14k array, which provided tiling path resolution only for chromosomes 4, 9, 10, 11, 16, 17, 21, 22, and X. All aberrations discussed here were detected with a sub-megabase tiling path BAC array.

Array CGH data were analyzed by CGHPRO. No background subtraction was applied. Raw data were normalised by “Subgrid LOWESS”. Copy number gains and losses were determined by a conservative log2 ratio threshold of 0.3 and -0.3, respectively. Aberrant ratios involving three or more neighbouring BAC clones were considered as genomic aberrations unless they coincided with a published polymorphism as shown in CGHPRO. In Figure 15, examples are shown of the genomic profiles from case 2 and case 4.

In total, 22 aberrations were identified in 22 patients. The size of aberrations ranged from 651 Kb to 14 Mb. Table 1 lists the aberrations found in each of the patients.



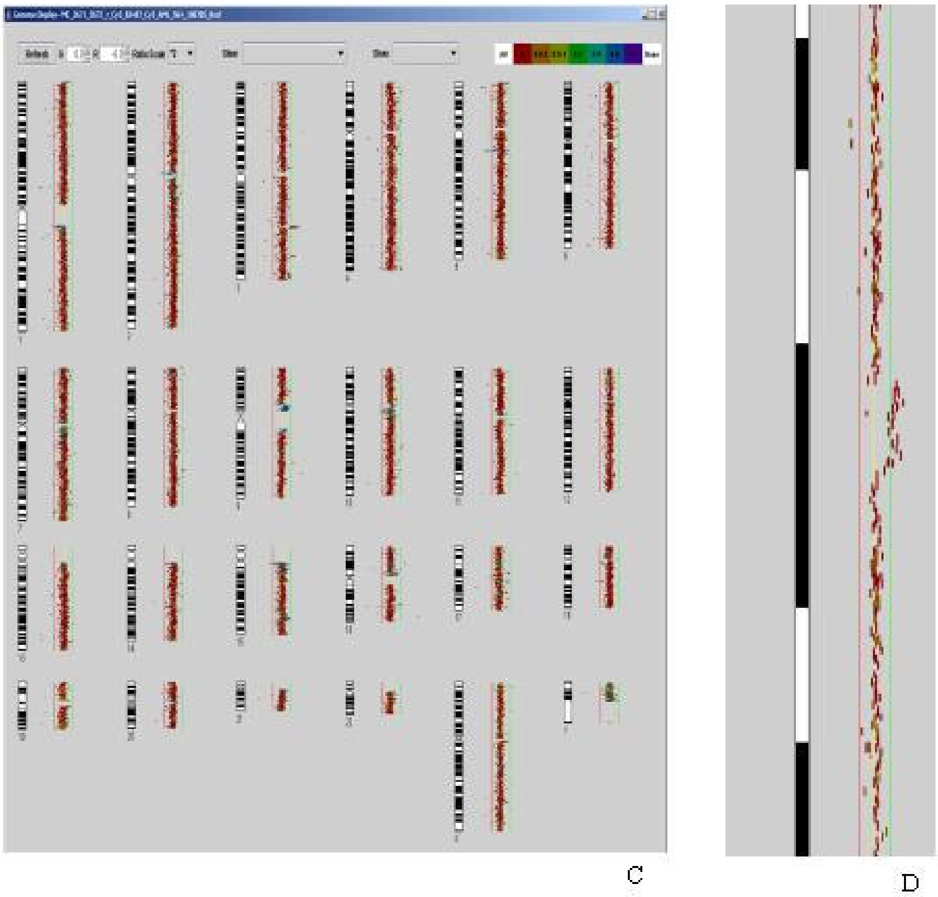


Figure 15: (A) Genome display of case 2. (B) Zoom-in view of the aberration in case 2 (C) Genome display of case 4. (B) Zoom-in view of the aberration in case 4.

Table 1: Array CGH results of 25 patients

case No.	chromosome band	gain/loss	start (kb)	end (kb)	size (kb)
1	16p11.2	loss	29573	30225	652
2	16p13.11	loss	14805	16425	1620
3	2q13	loss	111155	112776	1621
4	3q24	gain	145246	146988	1742
5	17q12	loss	31438	33481	2043
6	17q23.2-17q23.3	loss	55339	57686	2347
7	**22q11.21**	**loss**	**17202**	**20050**	**2848**
8	10q22.2-10q22.3	loss	75118	78473	3355
9	1p34.2-1p34.1	gain	40749	44190	3441
10	10q24.31-10q25.1	loss	102682	106175	3493
11	11q14.1-1q14.2	loss	82029	85900	3871
12	1q25.2	loss	176351	180687	4336
13	8q12.1-8q12.3	loss	60271	64625	4354
14	3q27.1-3q27.3	loss	184933	189324	4391
15	**15q11.2-15q13.1**	**loss**	**21265**	**26123**	**4858**
16	**17p12**	**loss**	**15545**	**20629**	**5084**
17	1p32.1-1p31.3	loss	59040	64557	5517
18	7p22.3-7p21.3	loss	1611	7158	5547
19	1p36.13-1p36.12	loss	17035	23035	6000
20	10q11.21-10q11.23	loss	45432	51611	6179
21	5q14.3-5q21.1	loss	90411	98322	7911
22	7p15.2-7p14.2	loss	26202	35318	9116
23	2q24.1-2q24.3	loss	154773	164127	9354
24	2p25.2-2p24.1	loss	6922	19766	12844
25	1q23.3-1q25.2	loss	160035	174241	14206

*Three patients with previously known genomic disorders are shown in bold. The 25 cases are sorted by aberration size.

5.1.3 Overlap of breakpoints in unbalanced aberrations with segmental duplication and CNPs

To estimate the content of segmental duplications and copy number polymorphisms (CNPs) around the breakpoints of the above 25 cases, a 400kb breakpoint interval, including 200kb proximal and 200kb distal of each breakpoint, was searched against the Segmental Duplication Database (http://humanparalogy.gs.washington.edu/) and the Database of Genomic Variants (http://projects.tcag.ca/variation/, version Dec 13, 2005) , respectively. Breakpoints were defined as the midpoint between end and start of the two neighbouring clones with alternate states, i.e. the one clone with normal and the other one with an aberrant ratio.

When segmental duplications were found to flank both breakpoints, the respective entries in the Segmental Duplication Database were checked for homology and degree of sequence similarity. The same procedure was also applied to the imbalances of an independent cohort of mentally retarded patients published recently (de Vries et al., 2005).

The segmental duplication content and DNA copy number polymorphisms (CNPs) found in the vicinity of breakpoint are shown in Table 2.

Table 2: Segmental duplication content and DNA copy number polymorphisms in 25 patients with unbalanced aberrations

Case No.	LCR content** upper breakpoint	CNP*** upper breakpoint	LCR content** lower breakpoint	CNP*** lower breakpoint	Size of homologous sequence(kb)	Sequence identity
1	3.088	-	2.898	-	146	0.996
2	5.162	-	4.37	-	310	0.957
3	1.218	+	1.615	+	44	0.995
4	0.0	-	0.0	+	0	0.0
5	2.299	+	3.041	+	200	0.987

6	1.454	+	1.769	-	48	0.987
7	**3.256**	**+**	**3.319**	**+**	**332**	**0.979**
8	1.835	-	0.0	+	0	0.0
9	0.0040	-	0.04	-	0	0.0
10	0.0	-	0.0	-	0	0.0
11	0.015	-	0.0	+	0	0.0
12	0.139	-	0.0080	-	0	0.0
13	0.0	-	0.0	-	0	0.0
14	0.0	-	0.0	-	0	0.0
15	**2.631**	**+**	**3.227**	**+**	**72**	**0.980**
16	**0.963**	**+**	**1.708**	**-**	**112**	**0.986**
17	0.0080	-	0.0	-	0	0.0
18	0.076	-	0.041	+	0	0.0
19	0.978	+	0.0	-	0	0.0
20	2.522	-	2.897	+	189	0.951
21	0.0	-	0.0	-	0	0.0
22	0.0080	-	0.0	+	0	0.0
23	0.0	-	0.0	+	0	0.0
24	0.0	-	0.0	-	0	0.0
25	0.0	-	0.0	-	0	0.0

*Three patients with previously known genomic disorders are shown in bold. The 25 cases are sorted by aberration size. Calculation is based on a 400 kb interval centered around the breakpoint.

**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone

*** CNP: DNA Copy Number Polymorphism

The three images displayed in Figure 16a-c represent a case where both breakpoints are flanked by segmental duplications (Figure 16a), a case with a segmental duplication present at only one breakpoint (Figure 16b) and an imbalance without segmental duplication in the breakpoints regions (Figure 16c).

Figure 16d shows the classification and distribution of all 36,000 BAC clones with respect to their segmental duplication content, as described in the Section 4.10.1.

Chromosomal imbalances found in this study can be grouped into four classes:

1. Proximal and distal breakpoints are enriched for segmental duplications with high sequence similarity (6/22; 27,3%)
2. Proximal and distal breakpoints are enriched for segmental duplications, but with low sequence similarity (3/22; 13,6%)
3. Only one breakpoint lies within a segmental duplication (5/22; 22,7%)
4. No segmental duplication lies in the vicinity of both breakpoints (8/22; 36,4%).

When group 1 was compared with the rest groups, a significant difference in aberration size was observed (p=0.018878; Wilcoxon rank sum test).

It is noteworthy that in each of the three patients with known genomic disorders, which are due to non-allelic homologous recombination, homologous segmental duplications were found to flank the respective breakpoints.

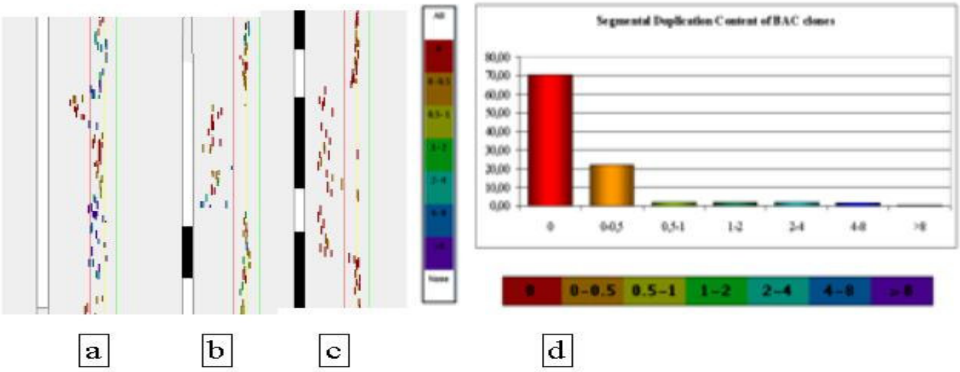


Figure 16: a-c Three example with two/one/no breakpoint covered by segmental duplication enriched clones. d Distribution of BAC clones with regard to their contents of segmental duplication (see text for detail).

When applying the same procedure to the independent dataset from de Vries et.al (de Vries et al., 2005), their aberrations can also be grouped into the same four classes, although the percentages are different. Table 3 lists the chromosome aberrations found in their study and Table 4 summarizes the segmental duplication content as well as the DNA copy number polymorphisms (CNPs) found in the respective breakpoint regions.

Table 3: Genomic imbalances in patients with mental retardation, as determined by array CGH (from de Vries et al., 2005)

Patient	Type of Submicroscopic Aberration and Chromosome Band	Gain or Loss	Start (Mb)	End (Mb)	Length (Mb)
1	1p34.3-1p34.2	Loss	39.22	43.15	3.93
2	2q23.1-2q23.2	Loss	149.17	150.09	.92
3	3q27.1-3q29	Loss	184.43	196.8	12.37
4	5q35.1	Gain	170.52	171.76	1.24
5	9q31.1	Loss	99.74	102.58	2.85
6	9q33.1	Loss	115.3	115.84	.54
7	11q14.1-11q14.2	Loss	78.12	85.61	7.49
8	12q24.21-12q24.23	Gain	114.91	117.21	2.3
9	17p13.2-17p13.1	Gain	4.27	7.16	2.89
9	17p13.1	Gain	7.67	9.10	1.43
9	17p12	Gain	12.65	15.54	2.88
9	17p11.2	Gain	18.55	20.03	1.48
10	22q11.21	Loss	17.1	19.75	2.66
11	1q21.1	Gain	143.25	145.38	2.12
12	3p14.1	Loss	67.59	68.15	.56
13	7q11.21	Loss	64.23	64.58	0.35
14	9p24.3	Gain	.21	.45	.23
15	15q24.1-15q24.2	Loss	72.21	73.86	1.65

Table 4: Segmental duplication content and DNA copy number polymorphisms in patients with unbalanced aberrations* (de Vries et al., 2005)

patient	LCR content** upper breakpoint	CNPs*** upper breakpoint	LCR content** lower breakpoint	CNPs*** lower breakpoint	size of homologous sequence(kb)	sequence identity
1	0.0080	-	0.029	-	0	0
2	0	-	0.01	-	0	0
3	0.013	-	0.613	+	0	0
4	0.038	-	0	-	0	0
5	0.012	-	0	-	0	0
6	0	-	0	-	0	0
7	0	+	0.0040	-	0	0
8	0	-	0.0040	-	0	0
9	0	+	0	-	0	0
9	0	+	0	+	0	0
9	0	+	0.963	+	0	0
9	1.684	-	1.118	-	264	0.992
10	3.07	+	2.261	+	283	0.969
11	0.786	-	4.808	-	0	0
12	0.011	-	0	+	0	0
13	2.624	+	5.824	+	172	0.994
14	1.805	+	0	+	0	0
15	0.9	+	1.484	-	54	0.934

*Calculation is based on a 400 kb interval centered around the breakpoint.
**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone
*** CNP: DNA Copy Number Polymorphism

5.1.4 Overlap of breakpoints in balanced translocation with segmental duplication and CNPs

Compared with unbalanced aberrations, less is known of the impact of segmental duplications on constitutive balanced translocations. Chromosomes 11, 17 and 22, containing regions with AT-rich palindromes embedded in segmental duplications seem to be more frequently involved and it appears that the affected duplications preferentially fuse with the telomeric regions of their translocation partner chromosome (Edelmann et al., 2001; Gotter et al., 2004; Kurahashi et al., 2004; Kurahashi et al., 2003; Kurahashi et al., 2000; Kurahashi et al., 2000; Shaikh et al., 2001; Spiteri et al., 2003; Stankiewicz et al., 2003)

To investigate the effect of segmental duplications on balanced translocations, we analysed the content of segmental duplications and CNPs around breakpoints in 41 balanced translocations, where the breakpoints had been mapped to single BACs by FISH. The procedure was the same as the analysis for chromosome imbalances except that the position of breakpoints in balanced translocations was defined as the midpoint of the respective breakpoint-flanking clone. Table 5 summarized the results.

Table 5: Segmental duplication content and DNA copy number polymorphisms in 41 mentally retarded patients with balanced translocation*

Case No.	LCR content** breakpoint1	CNP*** breakpoint 1	LCR content** breakpoint 2	CNP*** breakpoint 2	Size of homologous sequence(kb)	Sequence identity
1	0.009	-	0.0	-	0	0
2	0.000	+	0.0	-	0	0
3	0.000	+	0.0030	+	0	0
4	0.000	-	0.0	+	0	0
5	0.004	-	0.072	-	0	0
6	0.000	-	0.0	-	0	0
7	0.566	-	0.0	-	0	0

8	0.000	-	0.0	-	0	0
9	0.000	-	0.0	-	0	0
10	0.000	-	0.0	-	0	0
11	0.000	-	0.0	-	0	0
12	0.000	+	0.0	-	0	0
13	0.008	-	0.358	-	0	0
14	0.000	-	0.0	-	0	0
15	0.000	-	0.01	-	0	0
16	0.000	-	0.15	-	0	0
17	0.000	-	0.0	-	0	0
18	0.009	-	0.0	-	0	0
19	0.000	-	0.0	-	0	0
20	0.047	+	0.0	+	0	0
21	1.503	-	3.812	+	0	0
22	0.042	+	1.361	+	0	0
23	0.000	-	0.0	-	0	0
24	0.000	-	0.0070	-	0	0
25	0.000	-	0.0040	-	0	0
26	0.008	+	0.019	-	0	0
27	0.000	+	0.012	+	0	0
28	0.000	-	0.0	-	0	0
29	0.016	+	0.0070	-	0	0
30	0.0	-	0.087	-	0	0
31	6.382	-	0.0	-	0	0
32	0.000	-	0.0	-	0	0
33	1.987	+	0.000	+	0	0

34	0.000	-	0.032	-	0	0
35	0.007	+	0.0040	-	0	0
36	0.000	-	0.0030	+	0	0
37	0.000	-	0.0040	-	0	0
38	0.000	-	0.0030	-	0	0
39	0.0090	-	0.0	-	0	0
40	0.0080	-	0.358	-	0	0
41	0.133	+	0.0	-	0	0

*Calculation is based on a 400 kb interval centered around the breakpoint.
**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone
*** CNP: DNA Copy Number Polymorphism

For the constitutive translocations we have encountered the following distribution:

- Proximal and distal breakpoints are enriched for segmental duplications, but with low sequence similarity: 8/41; 20%.
- Only one breakpoint lies within a segmental duplication: 19/41; 46%).
- No segmental duplications in the vicinity of both breakpoints: 14/41; 34%.

5.1.5 Segmental duplication and CNPs in balanced and unbalanced rearrangements

Contrary to our findings in unbalanced rearrangements, both breakpoints in patients with constitutive balanced translocations were never found to be flanked by highly similar segmental duplications. Figure 17 shows segmental duplication frequency in the breakpoint regions of 41 balanced translations and 22 unbalanced aberrations.

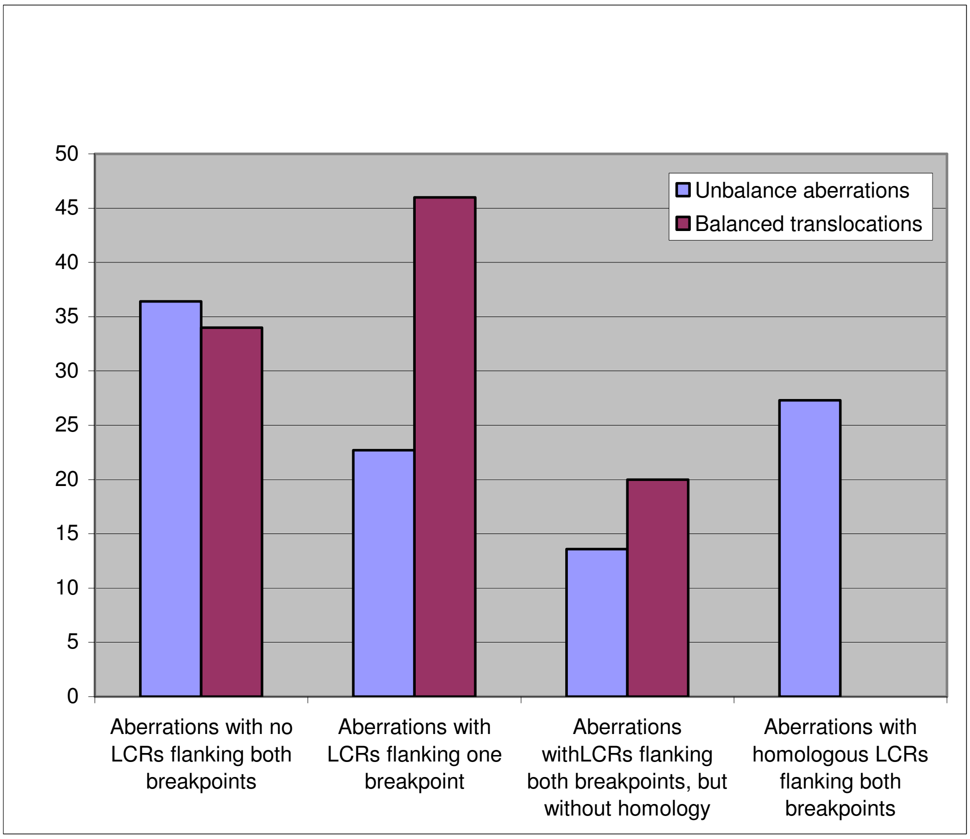


Figure17: Comparison of segmental duplication frequency and homology in breakpoint regions** of balanced and unbalanced aberrations. Data are based on a 400kb interval centered on the breakpoint. *LCR (low copy repeat) is same as segmental duplication. **Defined as 400kb interval centered around the respective breakpoint.

Figure 18 shows the frequency of DNA copy number polymorphisms at the breakpoint regions of 41 balanced translocations and 22 unbalanced aberrations.

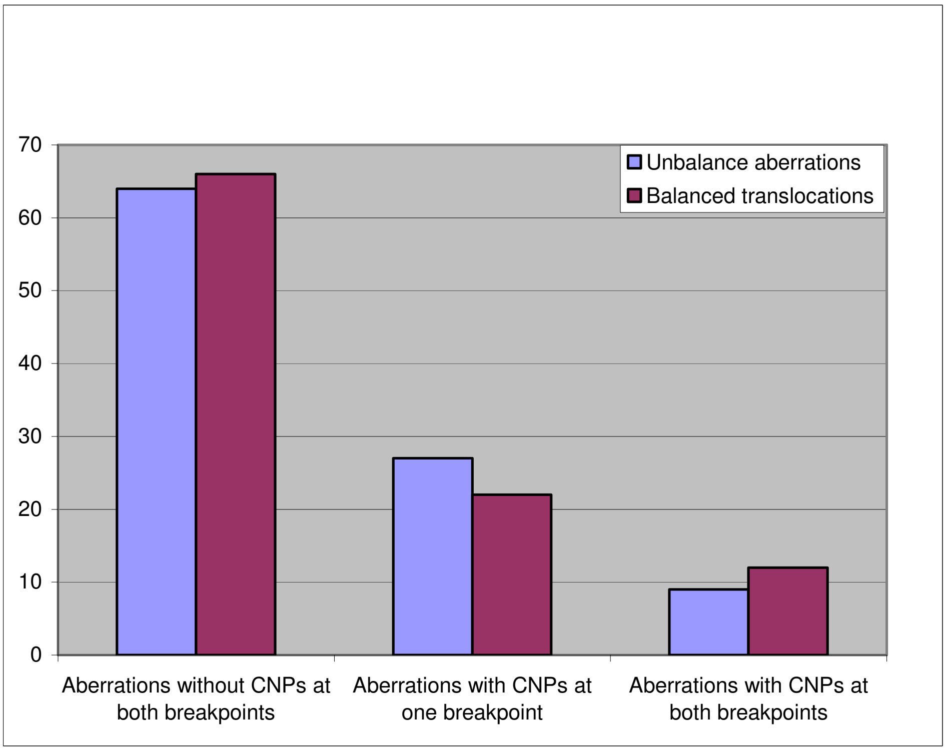


Figure 18: Comparison of CNP frequency in breakpoint regions** of balanced and unbalanced aberrations. Data are based on a 400kb interval centered on the breakpoint. *CNP: copy number polymorphism. ** Defined as 400kb interval centered around the respective breakpoint.

In order to avoid any loss of information due to the way of breakpoint definition, the whole procedure was repeated with breakpoint regions defined as 200kb and 1 Mb interval centered around the respective breakpoint, respectively. In Supplemental Material, Table S1 to Table S5 show the results obtained for the 200Kb interval, whereas Table S6 to Table S10 shows the results for the 1 Mb interval. Although the results differed between the intervals with different length, the underlying trends could still be observed. In this regard, interval size has no significant bearing on the conclusions.

5.1.6 Segmental duplication could mediate non-allelic homologous recombination (NAHR)

In this study, I have found a conspicuous clustering of segmental duplications at or near the borders of sub-microscopic deletions and duplication that were identified in mental retarded patients. In 41% of these imbalances, both breakpoint regions carried segmental duplications and in two third of these cases, sequence comparisons identified NAHR as the most likely cause of rearrangement. Several reasons may account for this unexpected observation. For instance, it could be due to an ascertainment bias: deletions or duplications arising from NAHR were associated with a more pronounced phenotype than those due to non-homologous end joining (NHEJ), which may affect the dosage in fewer genes. Alternatively, this observation may reflect certain flexibility in the choice of the most appropriate DSB repair pathway, and the presence of homologous sequences in the immediate vicinity of the damage can shift the decision in favour of NAHR. Therefore it seems plausible that DSB repair by NAHR is favoured in regions, where segmental duplications are clustered. The similarity of duplicated sequences, which is a prerequisite for NAHR, may be maintained by gene conversion. The reciprocal relationship of aberration size and frequency of NAHR perfectly matches this idea. Finally, if the distances between segmental duplications are small, there will be a low probability for meiotic crossover in the intervening fragments. This increases the relative frequency of NAHR, since such connections can serve as anchors to prevent chromosome slippage (Inoue and Lupski, 2002).

5.1.7 NAHR involving segmental duplications cannot explain all unbalanced rearrangements

NAHR between segmental duplications cannot be the sole mechanism underlying unbalanced rearrangements; after all, only nine out of 22 patients had segmental duplications in both breakpoint regions. Moreover, in some of these cases, the replicated sequences found near the deletion/duplication borders were not similar, and in other patients, segmental duplications were only observed at one of the two breakpoints.

However, this does not strictly rule out NAHR in these cases, as it is possible and even likely that many of segmental duplications are not represented in the human reference sequence because they were overlooked during sequence assembly (Cheung et al., 2003; She et al., 2004). Also, homologous segments that are necessary for NAHR can actually be fairly small, as illustrated by recombination events observed between Alu or Mer5B elements (Deininger and Batzer, 1999; Shaw and Lupski, 2005). Given the limited resolution of this analysis, such small repetitive elements may have been missed. However, even the presence of segmental duplications at one of the breakpoints may be enough to predispose for unbalanced rearrangements. In a recent report on Xq22.2 duplications in Pelizaeus-Merzbacher syndrome (Woodward et al., 2005), a one-sided distribution of segmental duplications at the chromosomal breakpoints has been explained by a mechanism involving homologous and non-homologous recombination (Richardson and Jasin, 2000). The resolution of the BAC array is not high enough to decide whether or not this mechanism accounts for some of the duplications in this study, but it cannot explain the respective deletions since this process is characterised by the generation of duplicated sequence. In two other studies, one with a deletion in a patient with Pelizaeus-Merzbacher like phenotype (Inoue et al., 2002) and the other dealing with Smith-Magenis syndrome (Shaw and Lupski, 2005), segmental duplications were also confined to one of the breakpoints. Detailed sequence analysis provided clear evidence for the involvement of NHEJ in these cases.

5.1.8 NAHR involving segmental duplications cannot explain balanced rearrangements

In patients with balanced translocations, no evidence was found for a major role of NAHR involving segmental duplications. Eight patients displayed low copy repeats in both breakpoint-spanning BAC clones, but the sequence similarity was only minimal. Inter-chromosomal duplication events are thought to have occurred more frequently at early stages of genome evolution, which explain why on average, the similarity between duplicated sequences on different chromosomes is lower than between duplicated sequences on the same chromosome (Zhang et al., 2005). Therefore, the chance to find a highly similar sequence on a heterologous chromosome is relatively low. Moreover, NAHR between duplicated sequences on different chromosomes may be restrained by the larger spatial distance in the nucleus.

Almost half of balanced translocations (19/41) displayed clustering of segmental duplication at one of the breakpoints, similar to the finding in patients with unbalanced rearrangements Thus, segmental duplication seem to be involved, but it is unlikely that NAHR is an important cause of balanced chromosomal rearrangements.

5.1.9 Clusters of segmental duplications may cause chromosomal instability

Given the conspicuous clustering of segmental duplications in breakpoint regions and the various arguments against a major role of NAHR, it is tempting to speculate that segmental duplications decrease the stability of DNA. This instability does not seems to be dependent on recombination events and may not only be caused by the potential of segmental duplications to form secondary structures at replication forks, but may also involve other e.g. epigenetic, mechanisms that are not yet understood.

In summary, we have demostrated an accumulation of segmental duplications at chromosomal breakpoints in mentally retarded patients, and it is possible that their presence predisposes chromosomes to rearrangement. However, it is also possible that segmental duplications are not always the cause of chromosomal instability, e.g. both may be secondary to other genetic or epigenetic factors, which are hitherto unknown (Bailey et al., 2004; Eichler and Sankoff, 2003).

5.2 Mapping of balanced translocation breakpoints using array CGH and CGHPRO

Disease-associated balanced chromosomal rearrangements form a unique resource for bridging genotypes and phenotypes (Bugge et al., 2000) and fine-mapping of chromosomal breakpoints in these patients has led to the identification of many disease genes. The traditional methods of mapping breakpoints in balanced translocations consist of FISH and Southern Blot analysis. Typically FISH mapping and ending up with the identification of a breakpoint-spanning clone requires several rounds of hybridisation starting with widely spaced clones. The whole process is rather labour-intensive and time-consuming, which has limited the large-scale application of this strategy for identifying disease genes.

As introduced in Section 3.2, Array CGH has been mainly used to detect copy number changes. In order to study balanced translocation, a novel technique termed ‘array painting’ has been developed (Fiegler et al., 2003; Veltman et al., 2003), which combines array CGH and chromosome sorting. As array CGH is the high-resolution variant of CGH, array painting is essentially an advanced variant of reverse chromosome painting (Carter et al., 1992). By replacing metaphase chromosomes with DNA sequences spotted on microarrays as hybridization targets, array painting greatly improves the resolution that can be attained by conventional reverse chromosome painting. As shown in Figure 19, array painting involves two steps. First, the two derivative chromosomes are isolated by preparative flow-sorting. The sorted chromosomes are then amplified, differentially labeled, and hybridized onto DNA microarrays. Fluorescence intensities are quantified with scanning device. Plotting the signal intensity ratios can reveal the DNA segment containing the respective junction fragment.

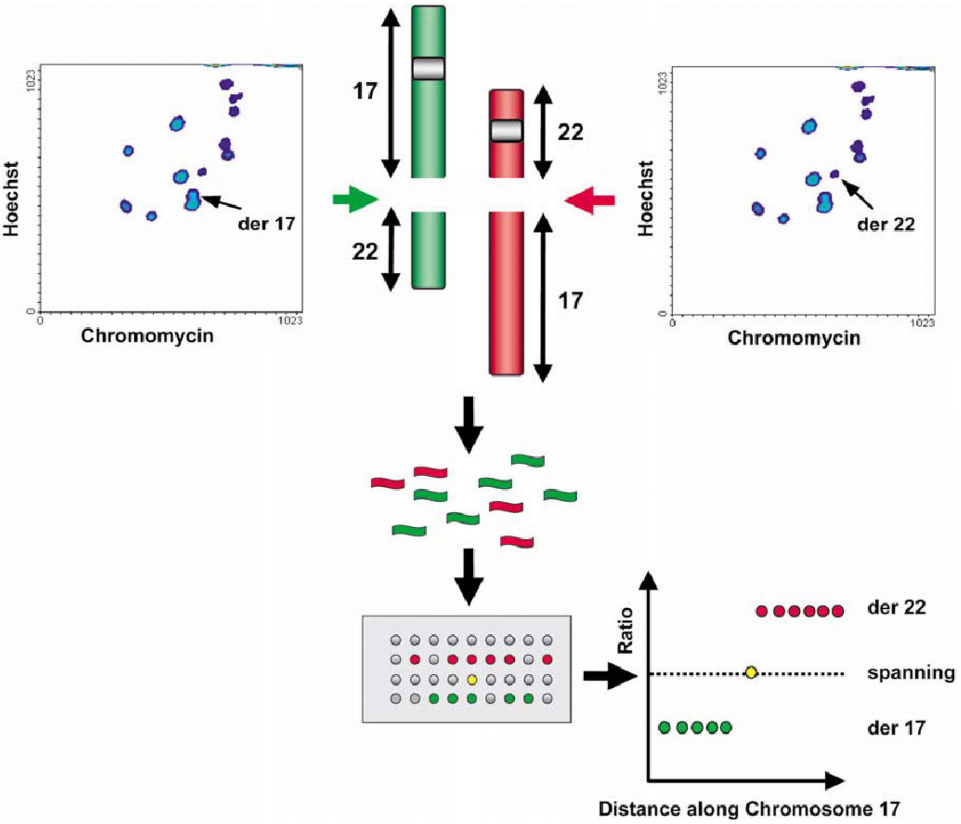


Figure 19: Principle of array painting (Gribble et al., 2004). First, the 2 derivative chromosomes, der 17 and der 22 are flow sorted. Then the sorted chromosomes are amplified, differentially labelled and hybridised onto DNA microarray. Fluorescence intensities are quantified with scanning device. Plotting the signal intensity ratios can reveal the DNA segment containing the respective junction fragment.

We have streamlined this approach even further. Below, I describe the successful application of this novel protocol to fine map the breakpoint in balanced translocation t(1;13). This protocol combining high-resolution array CGH analysis of flow-sorted chromosomes, the generation of DNA subarrays covering the respective breakpoint-spanning BACs, and long range PCR across the breakpoints and eventually sequencing of the junction fragments.

5.2.1 Cytogenetic investigation of the translocation

Chromosome analysis (GTG banding) in the proband had identified a balanced translocation between the short arm of chromosome 1 and the long arm of chromosome 13 (Figure 20). The karyotype was 46,XY,t(1;13)(p22;q14).

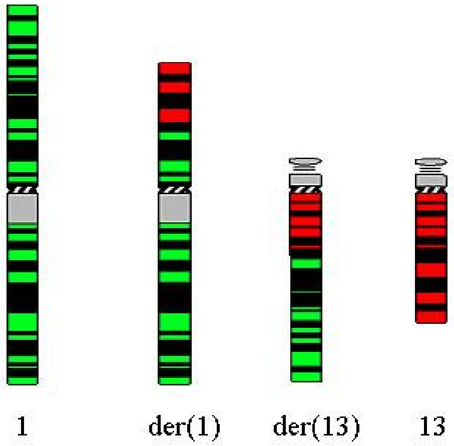


Figure 20: Partial karyotype of the patient with balanced translocation, showing the normal and derivative chromosomes 1 (green) and 13 (red).

5.2.2 Array CGH investigation of the translocation

Co-hybridization of differentially labeled DNA from the flow-sorted derivative chromosomes (der(1) and der(13), respectively) on the BAC array clearly showed the breakpoint regions on both derivative chromosomes (Figure 21 A B C). With the Zoom-in function of CGHPRO, even the breakpoint-spanning BAC clones, RP11-764P11 on chromosome 1 and RP11-339I10 on chromosome 13, could be identified.

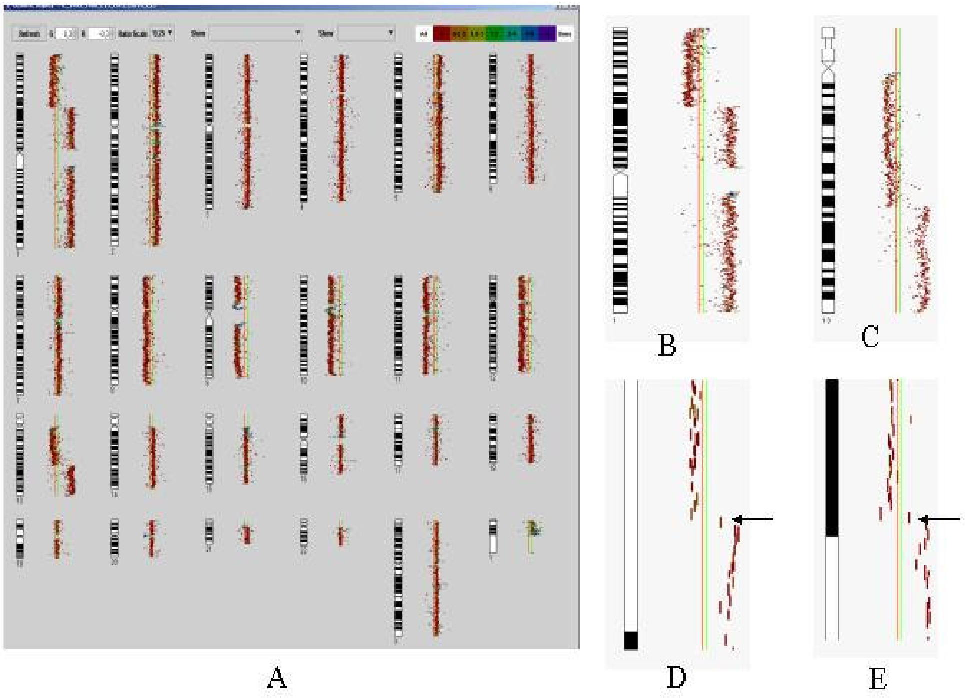


Figure 21: A) Results of the hybridisation of differentially labeled DNA from derivative chromosomes (der(1) and der(13)) on a whole-genome 36k array CGH

chip. B and C) Chromosome display of chromosomes 1 (B) and 13 (C), clearly indicating the translocated regions. D and E) Zoon-in view at the breakpoints reveals breakpoint-spanning BAC clones (arrows): RP11-764P11 on chromosome 1 and RP11-339I10 on chromosome 13.

5.2.3 Confirmation of the chromosome breakpoints by FISH analysis

To confirm the results, FISH experiments were performed. As shown in Figure 22 A, hybridisation with BAC clone RP11-764P11 gave rise to signals on both derivative chromosomes [(der(1) and der (13)], as well as a hybridisation signal on the normal, non-rearranged chromosome 1, thus confirming RP11-764P11 as the breakpoint-spanning clone on chromosome 1. Similarly, on chromosome 13, BAC clone RP11-339I10 was shown to span the breakpoints on both derivative chromosomes (Figure 22B).

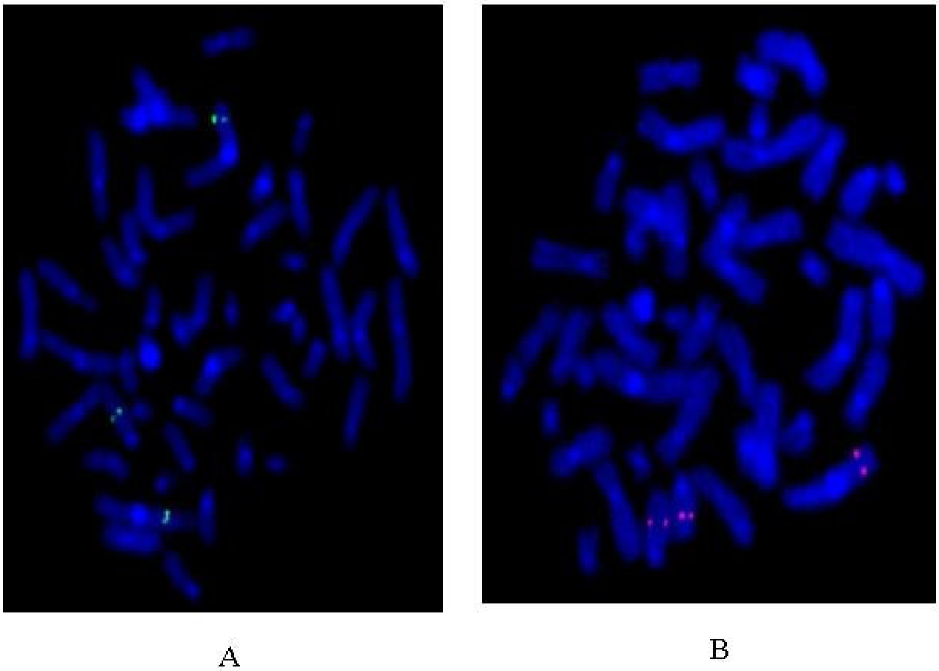


Figure 22: a) FISH result with the breakpoint-spanning BAC RP11-764P11 (green signals, and b) BAC RP11-339I10 (red signals). The three paired hybridisation signals correspond to the breakpoints of the derivative chromosomes and the respective normal counterpart (i.e., chr.1 or 13)

5.2.4 PCR fragment subarray

To further narrow down the breakpoint intervals, a so-called ‘sub-array’ was spotted with PCR products amplified from short segments distributing evenly along the two breakpoint-flanking clones, RP11-764P11 and RP11-339I10.

A customized Perl script, which is part of the CGHPRO package, was used to design the primers for the amplicons spotted on the subarray. The script first divides the BAC clones into evenly distributed intervals of 2 kb. Then, by using the Primer3 software (Rozen and Skaletsky, 2000), it designs primers for the generation of PCR probes within each of these intervals (average probe size: 500-800bp). To confirm that the amplicons are specific for the target region, the script searches the whole human genome for the presence of the amplicon sequences by using BLAST with the default parameter. Amplicons with more than one match in the genome are excluded from PCR amplification and spotting.

In total, 175 PCR products covering the breakpoint spanning BAC clones, RP11-764P11 and RP11-339I10, were generated and spotted on a subarray (for further technical details, see Section 8.1.2.6). Co-hybridisation of labeled DNA from the flow-sorted chromosomes on the subarray further narrowed down the two breakpoints, as shown in Figure 23.

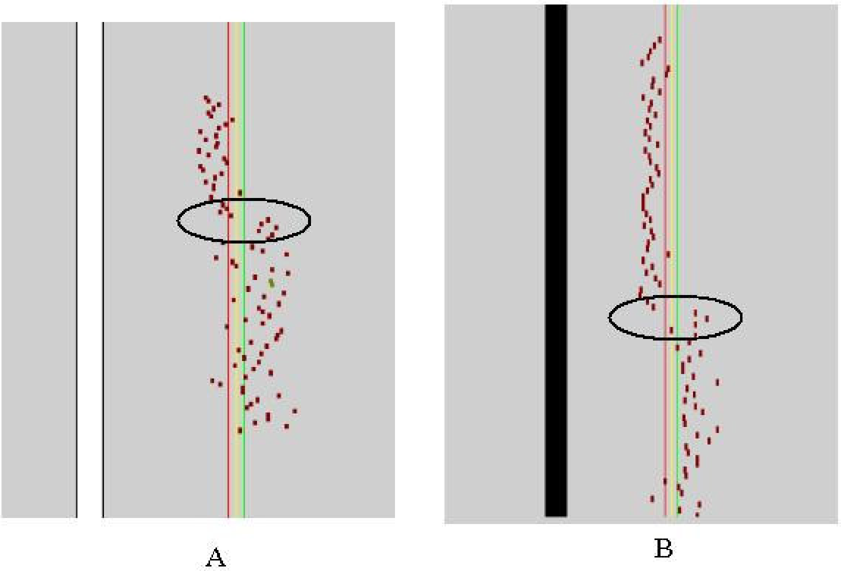
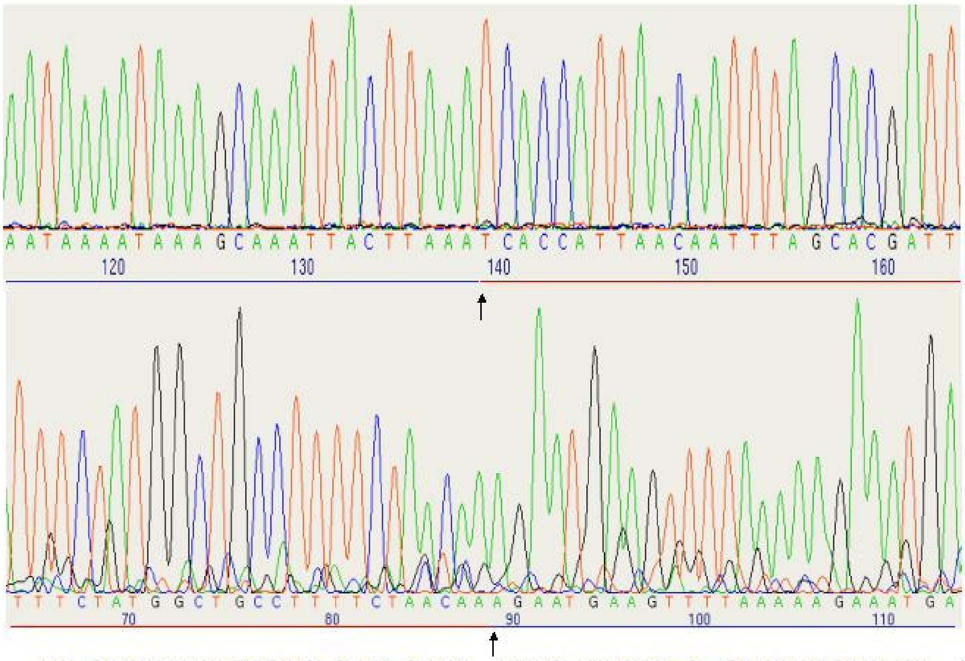


Figure 23:. Flow-sorted DNA from derivative chromosomes 1 (A) and 13 (B) was hybridised to a high resolution subarray spotted with multiple fragments of approximately 800 bp in size for fine mapping of the translocation breakpoint. The breakpoints (encircled) are thus mapped to a region of about 2-4 kb.

5.2.5 Long-range PCR and sequencing

The junction fragments from the derivative chromosomes were amplified by long-range PCR with primer pair 3 and 8 (more details, see Section 8.1.2.7). Sequencing of the specific PCR products and subsequent BLAST search against the human genome reference sequence mapped the chromosome 1 breakpoint to 67776107 bp, the chromosome 13 breakpoint to 71824031 bp (May 2004 UCSC Genome Browser, NCBI build 35). The sequencing chromatograph of the 2 amplified junction fragments and the corresponding genomic sequence from normal chromosome 1 and chromosome 13 are shown in Figure 24.



AATAAAATAAAGCAAATTACTTAAA AGAATGAAGTTTTAAAAAGAAATGA 1

TTTCTATGGCTGCCTTTTCTAACAA TCACCATTAACAATTTAGCACGATT 13

Figure 24: Sequencing chromatograph of the 2 amplified junction fragments and the genomic sequence from normal chromosome1 and chromosome 13. Arrows marked the exact breakpoints.

Genomic sequences around the translocation breakpoints were searched for specific features that may promote chromosome instability. In the junction fragments, microhomology and loss of one adenosine was observed, two characteristic features usually associated with NHEJ. On chromosome 1, a Mer-1 type element, MER5C, was found to span the breakpoint, while the short interspersed element (SINE) ALUJo was found to reside about 60 bps away from the breakpoint on chromosome 13. It is noteworthy that a Mer1-type element has been implicated in two related rearrangements (Abeysinghe et al., 2003) deposited in Gross Rearrangement Breakpoint Database [GRaBD; http://www.uwcm.ac.uk/uwcm/mg/grabd/grabd.html].

5.2.6 Advantage of array painting in mapping balanced translocation breakpoints

The strategy for chromosome breakpoint analysis described here is less laborious, less time-consuming and thus less expensive than previously employed methods such as single clone FISH experiments and Southern blotting. The costs of manual labour and reagents of the traditional method easily outweigh the considerable costs of high-resolution array CGH and chromosome flow-sorting experiments. Even though the application of this protocol is limited to the analysis of derivative chromosomes that can be separated by flow sorting, its implementation promises to pave the way for large-scale breakpoint mapping and gene finding in patients with disease-associated balanced translocations.

Hitherto, chromosome breakpoint analysis has been nearly exclusively restricted to de-novo rearrangements associated with congenital or early-onset disorders. However, such cases represent only a small percentage of carriers of balanced chromosome rearrangements. Balanced chromosome rearrangements were also shown to be associated with complex and late-onset disorders, and the identification of genes affected by the breakpoints could elucidate candidate genes for complex disorders such as schizophrenia, dyslexia, Tourette’s syndrome and psoriasis. Therefore, the availability of fast and cost-efficient methods of breakpoint analysis should also contribute to the identification of genetic factors in the etiology of late onset and complex disorders.

6 Summary

Genome rearrangements contribute significantly to the etiology of genetic disorders but also to human genetic diversity and disease susceptibility. For the detection of submicroscopic deletions and duplications on a genome wide level, a BAC-Array based technique for comparative genomic hybridisation (Array CGH), using a high number of overlapping BACs covering the whole genome is now being applied. The resulting data output however is of a magnitude that requires powerful management tools for handling not only large data quantities but also for coping with data quality variation.

To facilitate the analysis and management of array CGH data, I have developed a comprehensive software package called ‘CGHPRO’. Using the results from the image analysis software, CGHPRO allows hybridisation features to be checked with a variety of graphical representation options, thus enabling the selection of the most suitable normalisation method for individual experiments. A variety of options is then offered to characterize individual genomic profiles from the normalized data sets. All results are visualized in an interactive interface and stored in a database. The database allows the repetitive use of the stored results in comparative analyses, e.g. for investigating chromosomal aberration patterns in specific patient cohorts. In order to take the resolution of ArrayCGH applications beyond the BAC level CGHPRO allows the design of high-resolution specific sub-arrays.

The power of CGHPRO was demonstrated in the analysis of 22 mentally retarded patients with submegabase resolution whole genome tiling path BAC array CGH, which led to the identification of 20 deletions and two duplications. Additionally, as a proof of principle for CGHPRO assisted sub-array design, the breakpoints from a balanced translocation t(1;13) were successfully fine mapped.

When comparing the breakpoint regions for the 22 mentally retarded patients with those from a set of 41 balanced translocation carriers, in 6 of 22 unbalanced aberrations, breakpoint flanking duplications with a high degree of sequence similarity were found, suggesting that unequal crossing over might be one factor in chromosome instability. In all 41 balanced translocations however, even though breakpoint flanking duplications were observed, sequence homology between them never occured. This second finding indicates the existence of additional chromosomal instability factors which depend on or coincide with segmental duplications.

Taken Together, the results presented here demonstrate the powerful enhancement of the Array-CGH technique by the development and application of a versatile data management and ananlysis tool. It can be concluded, that the implementation of the protocols introduced here will, also for studies in large patient cohorts, greatly facilitate the identification and investigation of disease-associated chromosomal aberrations.

.

7 Zusammenfassung

Strukturelle Veränderungen im menschlichen Genom leisten einen signifikanten Beitrag zur Ätiologie genetischer Erkrankungen, aber auch zur genetischen Vielfalt sowie zur Kranheitsdisposition. Zum genomweiten Nachweis von submikroskopischen Deletionen und Duplikationen wird inzwischen vielfach eine BAC-Array basierte Methode zur vergleichenden Genomhybridisierung (array based comparative genomic hybridisation, Array-CGH) eingesetzt. Dabei führt die hohe Dichte an überlappenden, das gesamte Genom abdeckenden BAC-Klonen dazu, dass bei Array-CGH Experimenten Datenmengen generiert werden, deren Umfang und qualitative Heterogenität spezielle Software-Werkzeuge für eine effektive Auswertung erfordern.

Um die Analyse und Verwaltung von Array-CGH Daten zu erleichtern, habe ich das umfassende Software-Paket CGHPRO entwickelt. Dies ermöglicht dem Benutzer nach Übernahme der Daten von der Bildanalyse Software die Hybridisierungscharakteristika der einzelnen Experimente mit einer Reihe von graphischen Darstellungsoptionen zu überprüfen und eine jeweils geeignete Normalisierungsmethode auszuwählen. Für den Umgang mit den normalisierten Daten bietet das Programmpaket eine Auswahl an Methoden zur Charakterisierung individueller genomischer Profile. Alle Ergebnisse werden auf einer interaktiven Oberfläche dargestellt und in einer Datenbank abgelegt. Die Datenbak erlaubt die Verwendung der dort abgelegten Ergebnisse in vergleichenden Analysen wie z.B. der Suche nach Mustern chromosomaler Aberrationen innerhalb spezifischer Patientenkohorten. Um eine Auflösung über das mit BAC-Arrays erreichbare Maß hinaus zu erzielen, erlaubt CGHPRO das Design spezifischer hochauflösender Sub-Arrays.

Die Leistungsfähigkeit von CGHPRO wurde im Rahmen einer Analyse von 22 mental retardierten Patienten demonstriert die, unter Verwendung eines genomweiten BAC-Array mit Auflösung im Submegabasen Bereich,zur Identifizierung von 20 Deletionen und zwei Duplikationen führte. Ausserdem wurden, um das CGHPRO- unterstützte Design von Sub-Arrays experimentell zu überprüfen, die Bruchpunkte einer bekannten balancierten t(1;13) Translokation erfolgreich feinkartiert.

Beim Vergleich der Bruchpunktregionen der 22 mental retardierten Patienten mit den entsprechenden genomischen Bereichen von 41 Trägern balancierter Translokationen jedoch wurden bei 6 von 22 unbalancierten Translokationen bruchpunktflankierende Duplikationen mit einem hohen Grad an Sequenzhomologie beobachtet, was auf ungleiches Crossing-Over als einen Faktor chromosomaler Instabilität hindeutet. Bei allen 41 balancierten Translokationsfällen wurde trotz des Auftretens bruchpunktflankierender Duplikationen zwischen diesen niemals Sequenzhomologie gefunden. Letzteres weist auf das Vorhandensein weiterer chromosomaler Instabilitätsfaktoren hin, die entweder gemeinsam mit, oder in Abhängigkeit von segmentalen Duplikationen auftreten.

Insgesamt wurde in dieser Arbeit demonstriert, wie durch die Entwicklung und Anwendung einer vielseitigen Datenmanagement- und Analysesoftware die Leistungsfähigkeit der Array-CGH stark erhöht werden kann. Die gezeigten Ergebnisse erlauben darüber hinaus die Schlußfolgerung, dass eine Implementierung der vorgestellten experimentellen Ansätze insbesondere auch beim Studium großer Patientenkohorten stark zur Erleichterung der Identifikation und Untersuchung krankheitsrelevanter chromosomaler Abberationen beitragen wird.

8 Reference

Akaike, H.: Fitting autoregressive models for prediction. Annals of the institute of statistical mathematics 21 (1969) 243–

Abeysinghe, S.S., N. Chuzhanova, M. Krawczak, E.V. Ball, and D.N. Cooper. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. Hum Mutat. 22:229-44.

and Analysis ConsortiumThe Chimpanzee, S. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. 437:69.

Autio, R., S. Hautaniemi, P. Kauraniemi, O. Yli-Harja, J. Astola, M. Wolf, and A. Kallioniemi. 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. Bioinformatics. 19:1714-5.

Bailey, J.A., R. Baertsch, W.J. Kent, D. Haussler, and E.E. Eichler. 2004. Hotspots of mammalian chromosomal evolution. Genome Biol. 5:R23.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. Science. 297:1003-7.

Bailey, J.A., A.M. Yavor, H.F. Massa, B.J. Trask, and E.E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 11:1005-17.

Bailey, J.A., A.M. Yavor, L. Viggiano, D. Misceo, J.E. Horvath, N. Archidiacono, S. Schwartz, M. Rocchi, and E.E. Eichler. 2002. Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet. 70:83-100.

Barrett, M.T., A. Scheffer, A. Ben-Dor, N. Sampas, D. Lipson, R. Kincaid, P. Tsang, B. Curry, K. Baird, P.S. Meltzer, Z. Yakhini, L. Bruhn, and S. Laderman. 2004. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci U S A. 101:17765-70.

Bignell, G.R., J. Huang, J. Greshock, S. Watt, A. Butler, S. West, M. Grigorova, K.W. Jones, W. Wei, M.R. Stratton, P.A. Futreal, B. Weber, M.H. Shapero, and R. Wooster. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res. 14:287-95.

Bredel, M., C. Bredel, D. Juric, G.R. Harsh, H. Vogel, L.D. Recht, and B.I. Sikic. 2005. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. Cancer Res. 65:4088-96.

Brennan, C., Y. Zhang, C. Leo, B. Feng, C. Cauwels, A.J. Aguirre, M. Kim, A. Protopopov, and L. Chin. 2004. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. Cancer Res. 64:4744-8.

Bugge, M., G. Bruun-Petersen, K. Brondum-Nielsen, U. Friedrich, J. Hansen, G. Jensen, P.K. Jensen, U. Kristoffersson, C. Lundsteen, E. Niebuhr, K.R. Rasmussen, K. Rasmussen, and N. Tommerup. 2000. Disease associated balanced chromosome rearrangements: a resource for large scale genotype-phenotype delineation in man. J Med Genet. 37:858-65.

Carter, N.P., M.A. Ferguson-Smith, M.T. Perryman, H. Telenius, A.H. Pelmear, M.A. Leversha, M.T. Glancy, S.L. Wood, K. Cook, H.M. Dyson, and et al. 1992. Reverse chromosome painting: a method for the rapid analysis of aberrant chromosomes in clinical cytogenetics. J Med Genet. 29:299-307.

Carvalho, B., E. Ouwerkerk, G.A. Meijer, and B. Ylstra. 2004. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. J Clin Pathol. 57:644-6.

Cheung, J., X. Estivill, R. Khaja, J.R. MacDonald, K. Lau, L.C. Tsui, and S.W. Scherer. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. 4:R25.

Cheung, V.G., N. Nowak, W. Jang, I.R. Kirsch, S. Zhao, X.N. Chen, T.S. Furey, U.J. Kim, W.L. Kuo, M. Olivier, J. Conroy, A. Kasprzyk, H. Massa, R. Yonescu, S. Sait, C. Thoreen, A. Snijders, E. Lemyre, J.A. Bailey, A. Bruzel, W.D. Burrill, S.M. Clegg, S. Collins, P. Dhami, C. Friedman, C.S. Han, S. Herrick, J. Lee, A.H. Ligon, S. Lowry, M. Morley, S. Narasimhan, K. Osoegawa, Z. Peng, I. Plajzer-Frick, B.J. Quade, D. Scott, K. Sirotkin, A.A. Thorpe, J.W. Gray, J. Hudson, D. Pinkel, T. Ried, L. Rowen, G.L. Shen-Ong, R.L. Strausberg, E. Birney, D.F. Callen, J.F. Cheng, D.R. Cox, N.A. Doggett, N.P. Carter, E.E. Eichler, D. Haussler, J.R. Korenberg, C.C. Morton, D. Albertson, G. Schuler, P.J. de Jong, and B.J. Trask. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature. 409:953-8.

Chi, B., R.J. DeLeeuw, B.P. Coe, C. MacAulay, and W.L. Lam. 2004. SeeGH--a software tool for visualization of whole genome array comparative genomic hybridization data. BMC Bioinformatics. 5:13.

Daruwala, R.S., A. Rudra, H. Ostrer, R. Lucito, M. Wigler, and B. Mishra. 2004. A versatile statistical analysis algorithm to detect genome copy number variation. Proc Natl Acad Sci U S A. 101:16292-7.

de Vries, B.B., R. Pfundt, M. Leisink, D.A. Koolen, L.E. Vissers, I.M. Janssen, S. Reijmersdal, W.M. Nillesen, E.H. Huys, N. Leeuw, D. Smeets, E.A. Sistermans, T. Feuth, C.M. van Ravenswaaij-Arts, A.G. van Kessel, E.F. Schoenmakers, H.G. Brunner, and J.A. Veltman. 2005. Diagnostic genome profiling in mental retardation. Am J Hum Genet. 77:606-16.

Deininger, P.L., and M.A. Batzer. 1999. Alu repeats and human disease. Mol Genet Metab. 67:183-93.

Dhami, P., A.J. Coffey, S. Abbs, J.R. Vermeesch, J.P. Dumanski, K.J. Woodward, R.M. Andrews, C. Langford, and D. Vetrie. 2005. Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. Am J Hum Genet. 76:750-62.

du Manoir, S., M.R. Speicher, S. Joos, E. Schrock, S. Popp, H. Dohner, G. Kovacs, M. Robert-Nicoud, P. Lichter, and T. Cremer. 1993. Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Hum Genet. 90:590-610.

Edelmann, L., E. Spiteri, K. Koren, V. Pulijaal, M.G. Bialer, A. Shanske, R. Goldberg, and B.E. Morrow. 2001. AT-rich palindromes mediate the constitutional t(11;22) translocation. Am J Hum Genet. 68:1-13.

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. 17:661-9.

Eichler, E.E., and D. Sankoff. 2003. Structural dynamics of eukaryotic chromosome evolution. Science. 301:793-7.

Fiegler, H., P. Carr, E.J. Douglas, D.C. Burford, S. Hunt, C.E. Scott, J. Smith, D. Vetrie, P. Gorman, I.P. Tomlinson, and N.P. Carter. 2003. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. Genes Chromosomes Cancer. 36:361-74.

Fiegler, H., S.M. Gribble, D.C. Burford, P. Carr, E. Prigmore, K.M. Porter, S. Clegg, J.A. Crolla, N.R. Dennis, P. Jacobs, and N.P. Carter. 2003. Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays. J Med Genet. 40:664-70.

Fortna, A., Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, A. Karimpour-Fard, D. Glueck, L. McGavran, R. Berry, J. Pollack, and J.M. Sikela. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol. 2:E207.

Fridlyand, J., A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N.A.N. Jain. 2004. Hidden Markov models approach to the analysis of array CGH data. 90:132.

Fujiyama, A., H. Watanabe, A. Toyoda, T.D. Taylor, T. Itoh, S.F. Tsai, H.S. Park, M.L. Yaspo, H. Lehrach, Z. Chen, G. Fu, N. Saitou, K. Osoegawa, P.J. de Jong, Y. Suto, M. Hattori, and Y. Sakaki. 2002. Construction and analysis of a human-chimpanzee comparative clone map. Science. 295:131-4.

Gotter, A.L., T.H. Shaikh, M.L. Budarf, C.H. Rhodes, and B.S. Emanuel. 2004. A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. Hum Mol Genet. 13:103-15.

Gribble, S.M., H. Fiegler, D.C. Burford, E. Prigmore, F. Yang, P. Carr, B.L. Ng, T. Sun, E.S. Kamberov, V.L. Makarov, J.P. Langmore, and N.P. Carter. 2004. Applications of combined DNA microarray and chromosome sorting technologies. Chromosome Res. 12:35-43.

Hodgson, G., J.H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, N. Nowak, D.G. Albertson, D. Pinkel, C. Collins, D. Hanahan, and J.W. Gray. 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nat Genet. 29:459-64.

Horvath, J.E., J.A. Bailey, D.P. Locke, and E.E. Eichler. 2001. Lessons from the human genome: transitions between euchromatin and heterochromatin. Hum Mol Genet. 10:2215-23.

Hupe, P., N. Stransky, J.P. Thiery, F. Radvanyi, and E. Barillot. 2004. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics. 20:3413-22.

Inoue, K., and J.R. Lupski. 2002. Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet. 3:199-242.

Inoue, K., H. Osaka, V.C. Thurston, J.T. Clarke, A. Yoneyama, L. Rosenbarker, T.D. Bird, M.E. Hodes, L.G. Shaffer, and J.R. Lupski. 2002. Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. Am J Hum Genet. 71:838-53.

Ishkanian, A.S., C.A. Malloff, S.K. Watson, R.J. DeLeeuw, B. Chi, B.P. Coe, A. Snijders, D.G. Albertson, D. Pinkel, M.A. Marra, V. Ling, C. MacAulay, and W.L. Lam. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet. 36:299-303.

Kallioniemi, A., O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman, and D. Pinkel. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 258:818-21.

Kennedy, G.C., H. Matsuzaki, S. Dong, W.M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M.S. Phillips, M.T. Boyce-Jacino, S.P. Fodor, and K.W. Jones. 2003. Large-scale genotyping of complex DNA. Nat Biotechnol. 21:1233-7.

Kirchhoff, M., T. Gerdes, J. Maahr, H. Rose, M. Bentz, H. Dohner, and C. Lundsteen. 1999. Deletions below 10 megabasepairs are detected in comparative genomic hybridization by standard reference intervals. Genes Chromosomes Cancer. 25:410-3.

Knuutila, S., A.M. Bjorkqvist, K. Autio, M. Tarkkanen, M. Wolf, O. Monni, J. Szymanska, M.L. Larramendy, J. Tapper, H. Pere, W. El-Rifai, S. Hemmer, V.M. Wasenius, V. Vidgren, and Y. Zhu. 1998. DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am J Pathol. 152:1107-23.

Krzywinski, M., I. Bosdet, D. Smailus, R. Chiu, C. Mathewson, N. Wye, S. Barber, M. Brown-John, S. Chan, S. Chand, A. Cloutier, N. Girn, D. Lee, A. Masson, M. Mayo, T. Olson, P. Pandoh, A.L. Prabhu, E. Schoenmakers, M. Tsai, D. Albertson, W. Lam, C.O. Choy, K. Osoegawa, S. Zhao, P.J. de Jong, J. Schein, S. Jones, and M.A. Marra. 2004. A set of BAC clones spanning the human genome. Nucleic Acids Res. 32:3651-60.

Kurahashi, H., H. Inagaki, K. Yamada, T. Ohye, M. Taniguchi, B.S. Emanuel, and T. Toda. 2004. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. J Biol Chem. 279:35377-83.

Kurahashi, H., T. Shaikh, M. Takata, T. Toda, and B.S. Emanuel. 2003. The constitutional t(17;22): another translocation mediated by palindromic AT-rich repeats. Am J Hum Genet. 72:733-8.

Kurahashi, H., T.H. Shaikh, P. Hu, B.A. Roe, B.S. Emanuel, and M.L. Budarf. 2000. Regions of genomic instability on 22q11 and 11q23 as the etiology for the recurrent constitutional t(11;22). Hum Mol Genet. 9:1665-70.

Kurahashi, H., T.H. Shaikh, E.H. Zackai, L. Celle, D.A. Driscoll, M.L. Budarf, and B.S. Emanuel. 2000. Tightly clustered 11q23 and 22q11 breakpoints permit PCR-based detection of the recurrent constitutional t(11;22). Am J Hum Genet. 67:763-8.

Li, J., T. Jiang, J.H. Mao, A. Balmain, L. Peterson, C. Harris, P.H. Rao, P. Havlak, R. Gibbs, and W.W. Cai. 2004. Genomic segmental polymorphisms in inbred mouse strains. Nat Genet. 36:952-4.

Locke, D.P., R. Segraves, L. Carbone, N. Archidiacono, D.G. Albertson, D. Pinkel, and E.E. Eichler. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res. 13:347-57.

Locke, D.P., R. Segraves, R.D. Nicholls, S. Schwartz, D. Pinkel, D.G. Albertson, and E.E. Eichler. 2004. BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. J Med Genet. 41:175-82.

Lucito, R., J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J.A. West, S. Rostan, K.C. Nguyen, S. Powers, K.Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler. 2003. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res. 13:2291-305.

Lupski, J.R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 14:417-22.

McNeil, N., and T. Ried. 2000. Novel molecular cytogenetic techniques for identifying complex chromosomal rearrangements: technology and applications in molecular medicine. Expert Rev Mol Med. 2000:1-14.

Menten, B., F. Pattyn, K. De Preter, P. Robbrecht, E. Michels, K. Buysse, G. Mortier, A. De Paepe, S. van Vooren, J. Vermeesch, Y. Moreau, B. De Moor, S. Vermeulen, F. Speleman, and J. Vandesompele. 2005. arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. BMC Bioinformatics. 6:124.

Newman, T.L., E. Tuzun, V.A. Morrison, K.E. Hayden, M. Ventura, S.D. McGrath, M. Rocchi, and E.E. Eichler. 2005. A genome-wide survey of structural variation between human and chimpanzee. Genome Res. 15:1344-56.

Nimmakayalu, M.A., A.L. Gotter, T.H. Shaikh, and B.S. Emanuel. 2003. A novel sequence-based approach to localize translocation breakpoints identifies the molecular basis of a t(4;22). Hum Mol Genet. 12:2817-25.

Nobile, C., L. Toffolatti, F. Rizzi, B. Simionati, V. Nigro, B. Cardazzo, T. Patarnello, G. Valle, and G.A. Danieli. 2002. Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet. 110:418-21.

Olshen, A.B., E.S. Venkatraman, R. Lucito, and M. Wigler. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 5:557-72.

Oostlander, A.E., G.A. Meijer, and B. Ylstra. 2004. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet. 66:488-95.

Osoegawa, K., A.G. Mammoser, C. Wu, E. Frengen, C. Zeng, J.J. Catanese, and P.J. de Jong. 2001. A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. 11:483-96.

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, and D.G. Albertson. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 20:207-11.

Pollack, J.R., C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. 23:41-6.

Pollack, J.R., T. Sorlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Borresen-Dale, and P.O. Brown. 2002. 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. 99:12963-8.

Richardson, C., and M. Jasin. 2000. Coupled homologous and nonhomologous repair of a double-strand break preserves genomic integrity in mammalian cells. Mol Cell Biol. 20:9068-75.

Roth, D.B., and J.H. Wilson. 1986. Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. Mol Cell Biol. 6:4295-304.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 132:365-86.

Shaikh, T.H., H. Kurahashi, and B.S. Emanuel. 2001. Evolutionarily conserved low copy repeats (LCRs) in 22q11 mediate deletions, duplications, translocations, and genomic instability: an update and literature review. Genet Med. 3:6-13.

Shaw, C.J., and J.R. Lupski. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet. 13 Spec No 1:R57-64.

Shaw, C.J., and J.R. Lupski. 2005. Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. Hum Genet. 116:1-7.

She, X., Z. Jiang, R.A. Clark, G. Liu, Z. Cheng, E. Tuzun, D.M. Church, G. Sutton, A.L. Halpern, and E.E. Eichler. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature. 431:927-30.

Singh, G.B., J.A. Kramer, and S.A. Krawetz. 1997. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res. 25:1419-25.

Smirnov, D.A., J.T. Burdick, M. Morley, and V.G. Cheung. 2004. Method for manufacturing whole-genome microarrays by rolling circle amplification. Genes Chromosomes Cancer. 40:72-7.

Snijders, A.M., N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, and D.G. Albertson. 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet. 29:263-4.

Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer, and P. Lichter. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer. 20:399-407.

Spiteri, E., M. Babcock, C.D. Kashork, K. Wakui, S. Gogineni, D.A. Lewis, K.M. Williams, S. Minoshima, T. Sasaki, N. Shimizu, L. Potocki, V. Pulijaal, A. Shanske, L.G. Shaffer, and B.E. Morrow. 2003. Frequent translocations occur between low copy repeats on chromosome 22q11.2 (LCR22s) and telomeric bands of partner chromosomes. Hum Mol Genet. 12:1823-37.

Stankiewicz, P., and J.R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet. 18:74-82.

Stankiewicz, P., C.J. Shaw, J.D. Dapper, K. Wakui, L.G. Shaffer, M. Withers, L. Elizondo, S.S. Park, and J.R. Lupski. 2003. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet. 72:1101-16.

Telenius, H., A.H. Pelmear, A. Tunnacliffe, N.P. Carter, A. Behmel, M.A. Ferguson-Smith, M. Nordenskjold, R. Pfragner, and B.A. Ponder. 1992. Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes. Genes Chromosomes Cancer. 4:257-63.

Tjio, H.J., and A. Levan. 1956. The chromosome numbers of man. Hereditas. 42:1-6.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M.V. Olson, and E.E. Eichler. 2005. Fine-scale structural variation of the human genome. Nat Genet. 37:727-32.

Van Prooijen-Knegt, A.C., J.F. Van Hoek, J.G. Bauman, P. Van Duijn, I.G. Wool, and M. Van der Ploeg. 1982. In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp Cell Res. 141:397-407.

Veltman, I.M., J.A. Veltman, G. Arkesteijn, I.M. Janssen, L.E. Vissers, P.J. de Jong, A.G. van Kessel, and E.F. Schoenmakers. 2003. Chromosomal breakpoint mapping by arrayCGH using flow-sorted chromosomes. Biotechniques. 35:1066-70.

Wang, P., Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. 2005. A method for calling gains and losses in array CGH data. Biostatistics. 6:45-58.

Wilson, G.M., S. Flibotte, P.I. Missirlis, M.A. Marra, S. Jones, K. Thornton, A.G. Clark, and R.A. Holt. 2006. Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome Res. %R 10.1101/gr.4456006. 16:173-181.

Wirth, J., H.G. Nothwang, S. van der Maarel, C. Menzel, G. Borck, I. Lopez-Pajares, K. Brondum-Nielsen, N. Tommerup, M. Bugge, H.H. Ropers, and T. Haaf. 1999. Systematic characterisation of disease associated balanced chromosome rearrangements by FISH: cytogenetically and genetically anchored YACs identify microdeletions and candidate regions for mental retardation genes. J Med Genet. 36:271-8.

Woodward, K.J., M. Cundall, K. Sperle, E.A. Sistermans, M. Ross, G. Howell, S.M. Gribble, D.C. Burford, N.P. Carter, D.L. Hobson, J.Y. Garbern, J. Kamholz, H. Heng, M.E. Hodes, S. Malcolm, and G.M. Hobson. 2005. Heterogeneous duplications in patients with Pelizaeus-Merzbacher disease suggest a mechanism of coupled homologous and nonhomologous recombination. Am J Hum Genet. 77:966-87.

Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30:e15.

Yunis, J.J., J.R. Sawyer, and K. Dunham. 1980. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science. 208:1145-8.

Zhang, L., H.H. Lu, W.Y. Chung, J. Yang, and W.H. Li. 2005. Patterns of segmental duplication in the human genome. Mol Biol Evol. 22:135-41.

9 Supplemental material

9.1 Experimental protocols

9.1.1 Array CGH

For array CGH a 36k BAC sub-megabase resolution array was used, comprising the 1Mb Sanger set (clones kindly provided by Nigel Carter, Wellcome Trust Sanger Centre)(Fiegler et al., 2003) a set of 390 subtelomeric clones (assembled by members of the COST B10 initiative: Molecular Cytogenetics of solid tumours) and the human 32k Re-Array set, http://bacpac.chori.org/pHumanMinSet.htm; DNA kindly provided by Pieter de Jong) (Ishkanian et al., 2004; Krzywinski et al., 2004; Osoegawa et al., 2001).

The general overview of the process of array CGH experiment is shown in Figure 25. The following sections describe each step in detail.

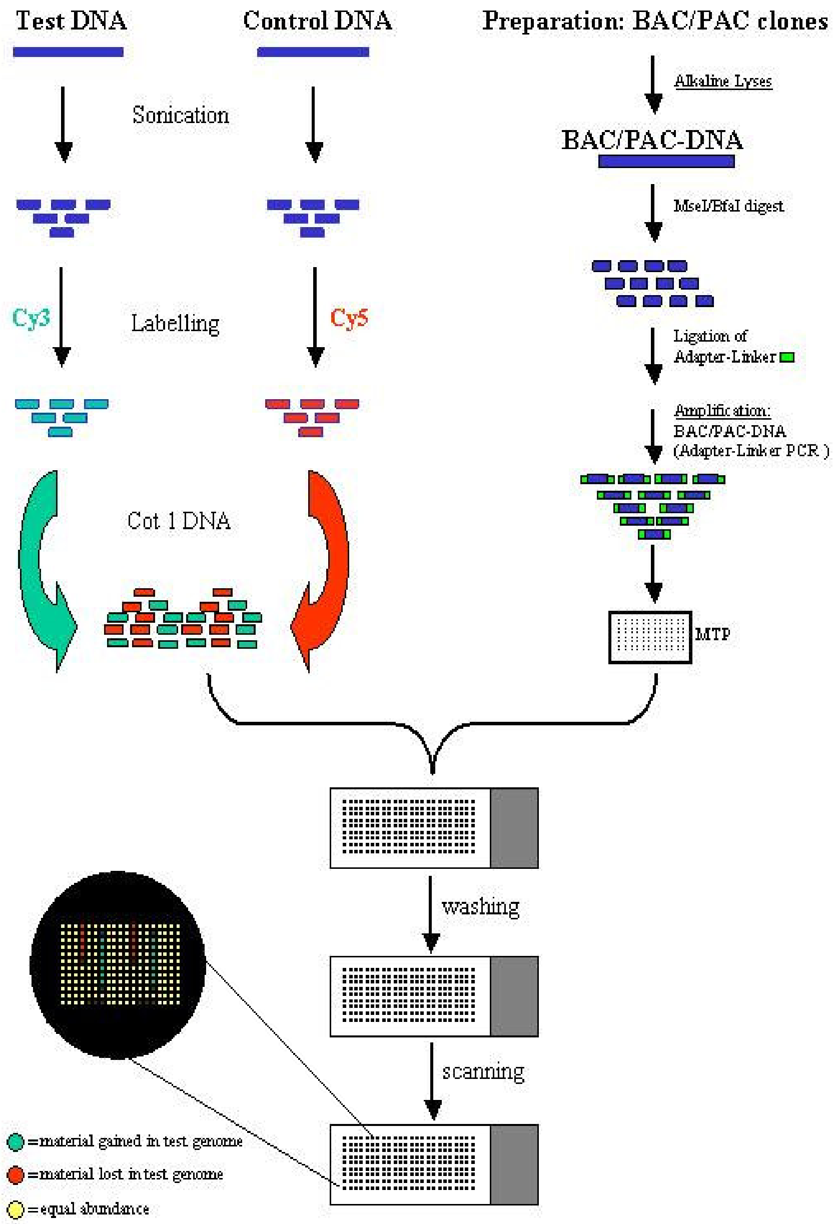


Figure 25: Procedure of array CGH (Courtesy of Dr. Erdogan). Up right part: the DNA from BAC clone inserts is isolated and amplified. The amplified products are spotted on the glass slides. Up left part: the DNA from test sample and reference sample are differentially labelled. The 2 differentially labelled DNA samples and excess unlabeled Cot-1 DNA, which can suppress the repetitive sequence, are then hybridised on the glass slides. Lower part: After hybridisation, the slides are washed and scanned.

9.1.1.1 Array production

9.1.1.1.1 BAC Insert isolation

BAC inserts were isolated in a 96 well format. First, BACs were treated by alkaline lyses according to the standard protocol. Then the DNA was treated with the exonuclease treatments to remove the remaining Ecoli DNA.

9.1.1.1.2 Amplification of BAC DNA by Linker adapter PCR

The isolated BAC DNA was amplified by linker adapter PCR. Linker adapter PCR consists of three steps: the target DNA is first digested with an appropriate restriction enzyme. And then each end is ligated to an adaptor. Finally, the known adaptor sequences are used to uniformly amplify each of the many DNA fragments representing the original samples. The following sections describe the three steps in detail.

9.1.1.1.2.1 Restriction Enzyme Digest of BAC DNA

Restriction enzyme digestion was carried out in a 7.5 µl reaction volume containing: 0.75µl of 10 x NEB1-buffer (New England Biolabs), 0,075µl of 100xBSA, 0,012µl of MseI (50U/µl), 0,15µl of BfaI (5U/µl) and 1,513 ml of H_2O, 5µl of exonuclease digested DNA. The reaction was placed in a PCR machine for 3h at 37°C. After incubation, the reaction was inactivated at 80°C for 20min. The digests were then run on a conventional 1% agarose gel to check fragment length. Restriction sizes should range from 100 bp to 1500 bp.

9.1.1.1.2.2 Ligation of Specific Primers to BAC DNA

The ligation reaction was first set up in a 8 µl reaction volume containing: 0,5µl 100µM primer-21 (5`-AGTGGGATTCCGCATGCTAGT-3´) and 0,5µl 100µM primer-12 (5`-TAACATGCATGC-3`), 0,8 10x ligase buffer (Roche), and 5,2µl of H_20 and 1µl digested BAC DNA (see above). In a thermocycler with heated lid an, the reaction was carried out at 65°C for the first 2 min to make the two oligos single stranded, and then the temperature was shifted down to 15°C, with a ramp of 1.0°C/min, to allow annealing of the two oligos. At 15°C, 0,2µl ligase buffer, 0,2µl T4-DNA-Ligase (5 U/µl; Roche), and 1,6µl of H_2O were added and the reaction was placed in the thermocycler for an overnight incubation at 15°C (18-20h).

9.1.1.1.2.3 Ligation Mediated PCR

The ligation mediated PCR was carried out in a reaction volume of 50 µl containing: 1µl ligation product, 5µl 10x PE buffer, 10µl dNTPs (1mM each), 0,5µl primer-21 (100µM), 32,5µ H_2O. Overlaying the sample with 30µl mineral oil to avoid evaporation at high temperature, the PCR program started at 68°C for 4 min to remove the MseI/BfaI-Lig12-Primer and 1 µl (10 units) of DNA polymerase was added and a 4 min. incubation for the fill-in reaction. After 3 min. at 95°C denaturation step, the PCR cycled at 95°C for 40 sec, 59°C for 30 sec, 90 sec (+2 sec/cycle) for 35 cycles. A 7 min extension at 72°C completed the protocol. Some of PCR products were run on a conventional 1% agarose gel to check fragment length. Size of the PCR product should range from 70 to 1500 bp, with the highest concentration of product around 200 to 800 bp.

9.1.1.1.2.4 Re-PCR of Ligation Mediated PCR

The ligation mediated PCR is used as a template in a Re-PCR reaction to generate DNA for spotting. 1µl of the primary PCR product was amplified under the following condition. After 3 min. at 95°C denaturation step, the PCR cycled at 95°C for 40 sec, 59°C for 30 sec, 90 sec (+2 sec/cycle) for 35 cycles. A 7 min extension at 72°C completed the protocol. Again some of the PCR products were run on a conventional 1% agarose gel to check fragment length. Size of the PCR product should range from 200 to 1500bp.

9.1.1.1.3 Preparation of Spotting Solutions from Re-PCR used for array CGH

The Re-PCR products were precipitated by adding 150 µl pre-chilled 100% ethanol and sodium acetate (pH 5.2). Then the dried DNA pellet was dissolved in 3xSSC/1,5 M Betaine.

9.1.1.1.4 Production of array

The products were robotically spotted onto epoxy coated glass slides (Nunc, Wiesbaden, Germany) using an in-house modified Qarray (originally from

Genetix, new Milton, U.K.) and Pointech (Gibbon, MN) Tungston PTL 2500 slit pins. Here, the microspotting technique was applied, where a spotting roboter disposed the PCR products directly on the slides. Epoxy slides were chosen due to several reasons. First, epoxy slides are especially suitable for covalent immobilization of oligonucleotides (10 to 80 bases), PCR products as well as cDNA molecules. Second, additional amino-modifications of the nucleic acids are not required. Third, their hydrophobic surface allows small spot diameters (100 to 130 µm, depending on the type of pins and spotting buffer) to create high-density arrays. Finally their surface chemistry is very stable and remains active even during very long spotting runs.

9.1.1.2 DNA labelling and hybridisation

9.1.1.2.1 Random Primed Labelling of genomic DNA for array CGH analysis

Genomic DNA samples were sonicated to generate fragments 200–2,000 bp in size. Test and reference DNA was labelled by random prime labelling (BioPrime DNA Labeling System, Invitrogen, Carlsbad, California) with fluorolink Cy3-dUTP and Cy5-dUTP (Amersham Biosciences, Piscataway, NJ). Briefly 1µg of DNA were mixed with 2.5 x random primer solution, incubated at 95°C for 10 min. and then immediately cooled on ice for 5 min. Consequently, 5µl dNTP mix (2mM dATP, 2mM dCTP, 2mM dGTP, and 1mM dTTP, in TE buffer), 3µl 1mM Cy3-dUTP or Cy5-dUTP and 1µl Klenow fragment was added to the reaction mix and incubated overnight at 37°C. The labelling reaction was stopped by adding 5µl stop solution. Probes were then purified by Qiaquick purification kit to remove the unincorporated nucleotides according to manufacture instructions. Two reactions as described above were pooled for each channel.

9.1.1.2.2 Slide Processing

Slides was prehybridised at 42°C for 1h in the blocking solution (200µl heringsperm DNA, 0,1% SDS, 4xSSC, 0,5% BSA). Afterwards, the Slides were immediately rinsed 5 times with Millipore water and air-dried by centrifugation for 5 min. at 150g.

9.1.1.2.3 Hybridisation of labeled genomic DNA

The labelled test genomic DNA and the labelled reference genomic DNA from two purified random priming reactions were pooled with 500 µg of human Cot-1 DNA (Invitogen, Roche). Pooled DNA was then precipitated by adding 2.5 volumes of ice-cold 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.2). The precipitated DNA was dissolved in 6,8µl 10%SDS, 3,4µl yeast tRNA (100µg/µl, Invitrogen), and 24µl master hybridisation mix (70% formamide, 2,8 x SSC, 8% dextran sulphate), and denatured at 70°C for 15 min. After denaturation, the hybridisation mix was incubated at 42°C for 2h to allow the Cot1 DNA to anneal to repetitive sequences on both the sample and reference DNA. The labelled probes were then placed on the slide under a coverslip. The arrays were incubated for 24 hours under humidified conditions using a slide booster from Implen (Munich, Germany). After hybridisation the slides were washed with 50% formamide 2xSSC, 0,1% SDS for 15min. at 42°C, followed by a 10 minute wash in PN buffer (0,2 M sodium phosphate with 0,001% NP40) at room temperature. The slides were then incubated for a 30 seconds in 1xPBS and 2-3sec in millipore water. Finally the slides were dried by centrifugation at 150g for 5min.

9.1.1.3 Scanning

Following hybridisation, slides were scanned at 532 nm (Cy3) and 635 nm (Cy5) using a GenePix 4000B laser scanner (Axon Instruments, Union City, CA) in order to read out the fluorescence signal intensities in each channel. The resulting 16 bit TIFF images were analysed employing Genepix Pro 5.0 software (Axon Instruments).

9.1.2 Mapping balanced translocation breakpoint by chromosome sorting and array painting

9.1.2.1 Cell culture and preparation

A t(1;13)(p31.2;q22.1)-containing cell line was established by Epstein-Barr virus transformation of peripheral blood lymphocytes and cultured in RPMI 1640 medium supplemented with 10% fetal calf serum, 2 mM L-glutamine, and antibiotics at 37°C in a humidified atmosphere containing 5% CO2. Cells in log phase were treated for 16h with colcemid (0.05 mg/mL final concentration) to arrest cells in metaphase.

9.1.2.2 Flow karyotyping and sorting

Chromosomes were stained with chromomycin-A3 (CA3) and Hoechst 33258 (Ho) and analysed on a dual-laser beam flow cytometer (FACSVantage™ SE; Becton Dickinson, Franklin Lakes, NJ, USA). CA3 was excited with an argon ion laser tuned at 458 nm at 100 mW, and CA3 fluorescence was measured through a 550 nm longpass filter. Ho was excited with an argon ion laser tuned into the UV range (351 and 364 nm) at 125 mW laser power, and Ho fluorescence was measured through two KV 408 filters. The system was triggered on the CA3 fluorescence signal. In total, 6000 to 12000 chromosomes were sorted from each cluster. DNA obtained in this way was amplified by GenomePlex.

9.1.2.3 DNA amplification using GenomePlex

For the amplification of flow-sorted chromosomes we used the GenomePlex Whole Genome Amplification (WGA) Kit (Rubicon Genomics). GenomePlex WGA is based upon random chemical fragmentation and conversion of genomic DNA into a library of DNA molecules flanked by universal priming sites. DNA Fragments are amplified by standard PCR using universal oligonucleotide primers.

9.1.2.4 Array painting using 36K BAC array

Labeling and hybridization of DNA amplified by GenomePlex on 36K BAC array were performed as described in the Section 8.1.1.2.

9.1.2.5 Fluorescence in situ hybridisation

To confirm the breakpoint regions determined by array-CGH, we employed fluorescence in situ hybridization (FISH) experiments. A permanent lymphoblastoid cell line of patient 2 was established by EBV transformation according to standard protocols after informed consent. FISH was performed using three BAC clones at each breakpoint region. For the breakpoint on chromosome 1, BAC clones RP11-55O04, RP11-764P11 and RP11-746B05 were employed. BAC clones RP11-490G20, RP11-339I10 and RP11-702F16 were used for the breakpoint on chromosome 13. DNA samples were prepared according to standard protocols and were labeled by nick translation with either biotin-16-dUTP or digoxigenin-11-dUTP. Immunocytochemical detection of probes was performed as described elsewhere (Wirth et al., 1999). Chromosomes were counterstained with 4'-6-diamino-2-phenyl-indole (DAPI). Metaphases were analysed with a Zeiss epifluorescence microscope.

9.1.2.6 PCR fragment subarray

The genomic sequence of specific breakpoint spanning BAC clones for chromosome 1 (RP11-764P11) and chromosome 13 (RP11-339I10) were chosen as targets for the design of a PCR amplicon subarray.

A customized Perl script incorporated within CGHPRO was used to design the primers for the amplicons. The script first divided the BAC clones into evenly distributed intervals of 2 kb. Then, by using Primer3 (Rozen and Skaletsky, 2000), within each interval, it designed primers for generating PCR fragments ranging from 500 to 800 bp in size. To facilitate the subsequent amplifications using the same condition, primers were selected to have the same annealing temperature. Finally, to confirm that the amplicons are specific for the target region, the script searched the whole human genome for the presence of the amplicon sequences by using BLAST with the default parameter. Amplicons with more than one matches in the genome were excluded from PCR amplification and spotting. The primers designed for RP11-764P11 and RP11-339I10 are listed in the Table 6 and Table 7, respectively.

Table 6: List of primer pairs designed for RP11-764P11 (chromosome 1)

NO.	FORWARD PRIMER	REVERSE PRIMER
1	TCCAGCTTCATTCATAGGGC	CTCAAAGCGCTCTTACCCAC
2	ATTCTGGCTAGGTGTGGTGG	TTAAGCACCTGTGACCTCCC
3	TTGAGGAACTGGGGACATTC	CAGTCTCTGCTTTTGAGGGG
4	GCTGAGCAGAGAGGGATTTG	TGGCCTTAAAACTGGACCAC
5	CTCAGCTACAGGAACCCCAG	TGGGTAAAATGTCCCTCCTG
6	TAACTGGATCTTCCGCATCC	CCCACCTGACCAATATGGAG
7	CAGCACACTGATGGGTCTTG	ACGTGAAGAATGCAGAAGCC
8	GGAGCTGGTTTTTCAAAAGG	CACGTGCCTGTAAGCCTAGC
9	CTTCAGGCAAAAGGTTGAGC	TGAACCCAGGAGAAGTCCAG
10	GCCACAGAGTCTAACGAGGC	TTCTCTGCATTCCTCACACG
11	TGGCCTCGTGTCTGTAGTTG	CTGGATTCAGGCCCTAAGTG
12	TCTGTGGTGTTTGCTGCTTC	TACCTCTGATGATGGGGAGC
13	GAGCCAGGCGTTCTGTTTAG	CTGCAATTGACCCACAAATG
14	TAGCAGGTCACCCAGAGTCC	CCCTCGGAGCCTCTATTTTC
15	ATCACCAGTGAAAAGGACGG	GAAGAGTCTGGCCTCCAGTG
16	ATCTGGGACAACAGAGCTGG	CTGACAGAAGGCTCCAGACC
17	CTTTGGAAGACTGAGGCAGG	TAGTTTGGCTGTGTTTCCCC
18	GGAAGAGCTTTTCATGCCAG	TGCATAAGCTTTTGTGCCAG
19	CCAGTCCTCTCTTGCCTGAC	ACTCATGGCCTATGACCCAG
20	TCAGAAGAATGGCCCCATAG	TTATGTCCAGCCCCCAGTAG
21	CTCTAGCCTCATCACCCAGC	TTTTGCACTCTGTCACCCAG
22	GATGCCTGCTTTCTTCCAAC	TCACCTCACAGCGAAGTCAC
23	AAATCACATCAAGGAACCGC	TGCCAAGTGTAGTGTCTGCC
24	TGTGTCAGAGCCACAGAAGG	AACATCGTGCGTTTACCTCC
25	ATTCCCTTGGCTGTCAAATG	CTCAGCCCTTGGAGAAACAG
26	CACCTCCATGATCCCAATTC	ACTCATGGGAACAGGAAACG
27	AATTCCAGCACTCCGTGTTC	ACAGTGGACAGGTTTGAGGC
28	ATGGGCATGAAGATAGGCAC	TTCTACAGAGGGCACATCCC
29	TACCAGATGTGCAGAGCCAC	GGGCACAGTGGTATTATGGG
30	CTCAGCACACAGTAGGCCAG	GAGGCAGCCATCATTCTCTC
31	TAAATTCCCTGCCATTCTGG	GGTTGCTTGCTTGTAATCCC
32	ACAGCCATCTTTCAACCCTG	CACTAACAGGCCCTCTCTGC
33	CATACCTGGGTTGCTTCCAC	CTCATGGGCTTTAGCAGCTC
34	TTCTGCCCCTGTTTTCATTC	TTTTGGCTCTTTTTGGTTCC
35	TTGGAGACTCAGAAATGGGG	TTCCACATTTTCTTCCAGCC
36	CAGTGTCCCCAAAGAGGAAG	TTTAATCAGGGCTGGAGTGG
37	GTGAACTGGGACTAGCCAGC	ATGAGGATAAGGACCCCCAC
38	GAGCTCTGACTTCTGGGTGG	TCACCTCTTTCCTGGGATTG
39	GCCTGAAATGCTCTCTACCG	AATGCTCATCCAGCCAAATC
40	GATCTTCAGGGAAGGAAGGG	CTCCTCTGGATAAGGGGCTC
41	CTGCCTTGAGTGAAAGGAGG	GTTTCCTCTTCATGCCTTGC
42	GAAGGGTTGCTTCAGACTGG	ATGCAAGAAAGCACATGCAG
43	AGGCAGGTGGATCATTTGAG	AAAATGGCATTACTGGGCTG
44	TGAGATTTGGGGTGTGATTG	ATCCACCCTCTTTGGTCTCC
45	ATTAGGAGTGCAGTGGCACC	GCTAAGGCTGATGAAGTGGC

46	CACAATCTGCATGCTGTTCC	TTTCAAGCATAGGTCCTGGG
47	CTGCTCAGTCCTTCAGGGTC	GGCAAGATGAAGAGTCCTGC
48	TGACTTCCCCTCATGACTCC	CCCTTCCATCTTCTTCCCTC
49	TGACAGGGAAGGAAATGAGG	GCAGGAAATCTCTGAGGCTG
50	TGTGCAGTCCACTCAAAAGC	AAGATCCACACATTCCTGCC
51	ACGTCCACTCACCCTTTGTC	TCGTCTGGGTCTCAAATTCC
52	TTGGTGAGGAAGACCGAATC	GAGAGCGGAAATGGAAGTTG
53	TGCAGTTCAGGCAAATGAAG	TAAAAGAAAATTGGTGCGGG
54	ACATGGCAAGTCTCCCTCAC	TGGTCCTCATGTTCAAGCTG
55	ATGTGACCAAAGGATCTCCG	AGCCCAGTACCTGGATGTTG
56	CCAGGCAGTTCCAAGAAGAG	AGTCAATGGGGTGACTTTGC
57	TTGTCAAGGAGGGGAGAATG	AACACAAGAGTGGGCAAACC
58	GAAGGCAGGGAGATCACTTG	TCCCCAAACAGAGGACAATC
59	GAGCCTGTCTCTCAAGCACC	GAGCAAGTGAAGGGATCAGC
60	TCAAGGTTCTTACGGGCATC	AATGTTGATGAGGAGCACCC
61	CTTTAAATGTGCTCCTCGGC	TGGAAGGAGAGAACCACCAG
62	CTTTGTGGTTTTGGGTCCAC	TGATGTTTGAGGCCTTTTCC
63	TTTGATCTTAGACAGCCGCC	TGATCCAGCACATGCTTCTC
64	AGCATATCCTGGGCATGAAG	TGTGCTGGGTGCTCTGTTAG
65	CCCACAAATGTGACTGCAAG	CCCAAAGTGAGGTTGTTTCC
66	ACCATGGGAGCATGTTAAGC	GGGTCTGTAATGGCTTCCTG
67	AAGTGGTATTTGCTGGGTGC	TGGGACAAAGTCCAGGCTAC
68	AATTGGTTGGTTGGCTTCTG	TAGTTGCTTTTCCCCACCAC
69	ATGCTCCGTGCTAAGCTCTC	TGTCCATCCTTCCTTCCTTG
70	TGCCACATACTGAAAGCACC	CTGACTGGACTTCCTGGTCC
71	AAAATAAGGTCTGGTGGGGG	TGCATATTGGTCACAATGGG
72	CCTAAAATAGGGAGGACGGC	TTGCGGAAGCAACATACAAG
73	CAAAAGCAGAGGGAGACCAG	TCTCTGTCGGTGCGTATCAG
74	TACTTCCAGTCATGAGGCCC	AGCCTGAGTCATCATTTGGC
75	CATCCCTGAATACAAACGGG	CACCTTTCGGTCTTCTGCTC
76	GCACTCTCGGGACTTCTCAC	AGATCCACAATCAGGCAACC
77	ACAGAAACTGATGCCAAGGG	GCAGCTGCTAGCAATATCCC
78	TGGAACGAGCACAGTAGCAG	GTGGAAAGGAGGAAGCAATG
79	TGAAGGGCATTTACTCCAGG	CTAAGCCCCAGTCTGAGTCG
80	CTAGCTGGGAGGCTTGTTTG	CCCAATGAGCTCCTAAATGC
81	TGTACAATGACTTGGGGTGG	ACCCAGGCAGATTGTACAGC
82	CAATGGGAGAACACATGCAC	GGCAAGCCTAAAGCAGTCAC
83	AACTCACCCCTACCCAATCC	CCCTTTCCCTTCCTTCACTC
84	AGTCATGAAATCATGCCGTG	GCCCTTAGCCTTGTGTCATC
85	ACTCAGCATCATTGAAGCCC	AGGAACATGCCAGGAGTGAC
86	CTAGCACAGCCCAGATAGCC	TGCTTCCAGGATTTTTCCAG
87	AGGTGGACAGTAGGTGTGCC	GTAAGTGCTGGTCTGGGAGC
88	TAGGCAACCTCACCCTGAAC	TGAGCTGGCAAGTGTGAATC

Table 7: List of primer pairs designed for RP11-339I10 (chromosome 13)

NO.	FORWARD PRIMER	REVERSE PRIMER
1	GAATGTATGGGCTGTTTGCC	TTTTTAATGCCTTTGCCCTG
2	GGAAGCAGCAGGAGACAAAC	GCTCCCTCTCATCTCCACAC
3	ACCAGATTTCTCCCATGTGC	CTTGTGTGGGTGGTTCAGTG
4	GAGAAACGGTCATTTTTGCC	CTTTTCCTTCAAGGTGCTGC
5	GAGTCCACTCTTGCCTCCAG	TTTTGGGACTTGCATACCTTG
6	GCCGTCTGCTAGCTTTTG	AAGACCCATCAGTGTGCTG
7	TCATCGCGTAGTTCTTGTGC	CACCATTGCAGAGGATTGAG
8	TGTCCGTGTTTTTCAAATGG	CCGTGTTAACCAGGATGGTC
9	ATGGATCATGAAGGACCCTG	ACGTCTCACCCTAATGTGGC
10	TGCTGAATGCAGTTTTCCTG	CCTAACCCCCTTGTCTCTCC
11	AGCAGAGTTGTATGGGTGCC	TTGGCTTAGGTTTACTGGGG
12	TAATTCCCTTGTGGAGCTGG	AAATTGGGATCCGAGAAACC
13	GGGATTGGCTATTCCTCCTC	TGTGGTCCATTGTTTGTTGG
14	TCCAAAATGAGAAAATGCCC	CGCTGAGCCTTGGTTTCTAC
15	GGTGCCTCTCAAGGTACAGG	TCTCTGCCCAAGCTGACTTAG
16	TCAAAGCACAATTGAGGGTG	TCAACTTGTCCCTCAGAGCC
17	AATTCAACCTTTTGCCTCCC	CCCTAATGAATGGGATGTGG
18	CCTTCCTTGAGGGAGGAAAC	AAATCCAGTGATGGTAGGCG
19	ACAAGGCAAAGGGTCACTTG	GTGAGAGAGCCTGACATCCC
20	CTTGAGGTCAAGCCAGAAGG	TTCCCAAAGGCTTATGATGG
21	GGGCTAACCAATCAAAATGC	AATGCCCTTCTATGTGTGTGG
22	AACAACCGTGGATTCTCAGG	CTTGCAAACTCTCCTTTGCC
23	ATCACCTTTTCTGGCCACTG	GGGAACCAGAAGATGCAAAG
24	CTCCATTGTTCCATGGCTTC	TGGAAGGAAATTCCAAGCAC
25	AATTTTGGATCTTCCCCCAG	AGTTCACCTTGGACCCACTG
26	ATGGCCACAGTATGTCCTCC	GAGTTGTTTCTGGCCTCACC
27	TGAGTGAGCACGGACAAGAC	GGCTGGTGTCCTGAGACTTC
28	AGGCACTATGCTATGGCCTG	AGCCACCAGCTTTGTCTCTC
29	ACCTGCAAAGCAATTCCAAC	CTTTCACTCAGGCCAGGAAC
30	CACAATCATAGGCATGTCGC	CAGAGACAGAGCAATGCGAG
31	TGGCAATTGTCTCCACTTTG	GCAAAGCAGGAGTAAGGCAC
32	GCATAGGTAGGTGCCTCTCG	ACTGGGCAAGTCATCGAATC
33	AATTTTAAAAGTTGCCCCCTC	TGCCAGATGGAGAGATTGC
34	CCCCAGAAGTCCTCTGTTTC	AAATGTTGATATCCTGGCCG
35	CATGGGATTTTGTACAGGGC	AGTTACCAACCCCTGATCCC
36	CTCGGTTGGAACAAAAGAGG	AAACCATTGCCAAATCCAAG
37	CGCAAAAGCAGGAAGTATGG	GATATCGCTCCCATTTCTGG
38	ATGCCTTGAAAAAGAGGCAC	AGCAGGAAATGGTGATGAGG
39	TGTGTGAGTGCTTGGTAGCC	ACAAGGGTGTTCTGACCTGC
40	ACACAGGAAAAGCCCTTGTG	TCTTTCCCCATTCCTAACCC
41	AAAGACCGCATATGCCAAAC	CAAAGATGCTGTCCTTTCCC
42	AGGACTGCTCTGCCTACGAG	TAAGGTGGGAGGATTGCTTG
43	TGAGATAAAGCCAGGGATGG	GAAGGAACAAAGCAAGCAGG
44	AAATTCCTGGAGCATTGTGC	CCAGTGTTCCTCCTCTCTGC
45	AGGGGATGCCCAAAAATATC	GACTTTTGGCTGTTTCTCCG

46	TTGGATGGGGGAGTGTTATC	TACCCTCTAATGCGGCAATC
47	CTCAGAGGCAGATAGCCCAC	AGATGGGCAATGAGATCCAG
48	ATTTCCCTCCTTTGTGGGAC	TATTGCAGGGAGATTTTGGC
49	TGCTGAGGATTCATCCCTTC	GCTGTTTGCAACATTCATCC
50	TGGCAGATTTATGAGCTTTGC	TTCTTGGCTCCACCCTTATG
51	TTCCTCAGGCCTTCTTTCAG	CACCCCGTGTCCAAGTATTC
52	AGATGTTTGGAAGGCAATGG	TGAGATGAGGACTGGCTGTG
53	TGGCAATTTCCTGTGAATCC	GCCAGTGAGGAAGAGTCAGC
54	AATTTCTCCATCATCTGGGG	TCTCGTTTGGGAATCTCTCG
55	CGAATTTTCACCAGTCCCTG	TGGGAATGAGAAAGACAGGC
56	GACTTTGCCAACTTTGGAGC	CTTTGCTCTGTCTTTTCGCC
57	TTGAAGAGACTGCCCTTTCC	AAAAATGCCTTATGAAAGCCC
58	AACGCAGTTCAAGTCCAGTG	CTTCTTTCTGTGAGCTGCCC
59	ATATGGGCAGGATAAAGGGC	TCACACTTGGAAATGCCTTG
60	CTTGAGCCCAGAGACCACTC	AGGGCCCATTTTCATTTTTC
61	AAAGTGGGTGGAACACAAGG	CACATGACCACAAGGTGAGG
62	GGTTGAGCCAGCTCTCTTTG	AATTTCTGTGGTGTGCTCCC
63	GCAAGGGGTATAGCAGATGG	AACAACCACAAGAAGGCTGG
64	ACTGCTTCAAGCTCAGGCTC	GAATGATACGGTGGGGAATG
65	TAGACCCCAGATTGCTGTCC	TGCAGTTTAAACCTGGAGGC
66	ATCTTGGACTGTGTGGGCTC	TGCGTTTCTCCTGTGTATGC
67	ATGAAAAGTCCCTGTGGCTG	GCATTGTGACCAAGCATGAG
68	TTTGAGGAGTTGGGATCCAG	TGCAATGAAAGCCACAGAAG
69	TACCCCTGAAAATGACACGG	GGCAGATTGCTCTCGAACTC
70	GAGCTGGTGGACGAGTTAGG	ACCAAAACCAGCAAAGGTTG
71	GTCATGCCTTTGGAGGTTTC	TCCTGAGGAGCTGGGAGTAG
72	ATCCTCATTCATCCACTGCC	GGGTTGCAATGATTTTGAGC
73	AGGAAATGAAATCCCCCATC	GCTTTTCCTTGAATTGCAGC
74	TCTTACATTCACCCGCTTCC	TTGCAATTGCTTCCTGTGTC
75	GCCAGTAAGTGGGAAAGCTG	CTGGAGGAAAAAGCACAAGG
76	CAGCTCTGCCTTTGGAGAAC	AAAATTCCTTCTGTGCCTGC
77	AGAACTGGGAGTGTCCATGC	TCTTCAGCAACTCTGCCAAG
78	TTTGAGGCCCTGAGTTATGG	AGTAACTGTGCCTTGTGGGG
79	TGCCGTTTTAATCTGGTTCC	TCTTCTTTTGGACACCTGGC
80	AAACCTGGCAAGCACAGTTC	TTGCTTATCTAGGCATGGGC
81	TGGCTGACTGAATAAAGGGC	ATTGGACTGCTGGCCACTAC
82	GAGGGAGGGAGGGAAAGAC	CTCCCTCAGTTACCCTGCTG
83	TTGCAGTGCAGAGCAAAATC	CTCCGGTTACAGGTTTGAGC
84	TTTAGTGGTGGTTGCAGTGG	AGTGGTAATGTTCATGGGGC
85	AGGGCACGTGAGATACAGAC	GGCTTTTAAACACCCCTTGG
86	GGTGAAATTTCCACAATGGG	AGGGCCCTTATCTCTCTTGC
87	TAGGGACTGCCAAAAACTCC	TTTAGTTCCAGCATTTGGGC

HPLC-purified oligonucleotides were obtained from MWG Biotech AG (Ebersberg, Germany). Lyophilized primer were dissolved in H_2O to a final concentration of 100 µM and stored frozen at -20°C. All PCR products were amplified in a thermal cycler (TC9700, Perkin Elmer) under the conditions described below. Reaction mixtures of 50 µl contained 200 µM of each dNTP (Roche Molecular Biochemicals, Mannheim, Germany), 20 pmol of each primer, 100 ng of a genomic DNA preparation, 1× PCR buffer and 2U AmpliTaq (Perkin Elmer). Since the primers were selected to have the same annealing temperature, PCR products were amplified under the same conditions as following: initial denaturation (5 min at 94°C) followed by 35 cycles of denaturation (1 min at 94°C), annealing (40 sec at 59°C) and extension (2 min at 72°C), a final extension step was carried out for 5 min at 72°C. PCR products were ethanol precipitated, dissolved in 3xSSC/1.5M betaine and spotted on epoxy-coated slides (NUNC). Labeling and hybridisation of DNA from sorted chromosomes on the PCR amplicon subarray were performed as described in the Section 8.1.1.2.

9.1.2.7 Long-range PCR and sequencing

Long-range PCR amplification of breakpoint-spanning fragments was performed using specific primer pairs with one primer mapping to chromosome 1, and a corresponding primer mapping to chromosome 13 (Table 8). The ExpandTM Long Template PCR system (Roche Applied Science) was employed.

Table 8: Primer pairs designed to amplify the breakpoint-spanning fragments

NO.	FORWARD PRIMER	REVERSE PRIMER
1	ACAGCCATCTTTCAACCCTG	AGATGGGCAATGAGATCCAG
2	CACTAACAGGCCCTCTCTGC	CTCAGAGGCAGATAGCCCAC
3	ACAGCCATCTTTCAACCCTG	CTCAGAGGCAGATAGCCCAC
4	ACAGCCATCTTTCAACCCTG	AGATGGGCAATGAGATCCAG
5	CATACCTGGGTTGCTTCCAC	TATTGCAGGGAGATTTTGGC
6	CTCATGGGCTTTAGCAGCTC	ATTTCCCTCCTTTGTGGGAC
7	CATACCTGGGTTGCTTCCAC	ATTTCCCTCCTTTGTGGGAC
8	CTCATGGGCTTTAGCAGCTC	TATTGCAGGGAGATTTTGGC

Amplifications were performed in a thermal cycler (TC9700, Perkin Elmer) in 20µl reactions containing 50ng of DNA, 30 pmol of each primer, 10µl of FailSafeMix (0.4mM of each dNTP, 1×PCR buffer) and 3.5U of polymerase mix, with the following cycling parameters: after initial denaturation at 92°C for 2 min, 10 cycles of denaturation 30 sec at 92°C, annealing 30 sec at 58°C and extension 5 min at 68°C, followed by 20 cycles [30 sec at 92°C, 30 sec at 58°C and 5 min at 68°C (+20sec/cycle)] and a final extension step at 68°C for 7 min.

The three PCR products (No. 3, 4, 8) were used as templates for sequencing in both directions by use of BigDye Terminator chemistry (PE Biosystems). Separation and visualisation was performed on an Applied Biosystem 3730xl DNA Analyzer.

9.1.2.8 Agarose gel electrophoresis

DNA fragments were separated and visualized by agarose gel electrophoresis. Gels of 1% agarose (Invitrogen) in TBE buffer (0.1 M Tris, 0.1 M boric acid, 2 mM EDTA) were supplemented with 0.5 µg/ml ethidium bromide. At least 0.2 volumes of gel loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, and 30% glycerol) was added to the nucleic acid solutions before loading into the wells. DNA size markers HyperLadder I 100 Lanes (Bioline) were also loaded. Gels were run at 100 V for 30-45 min. Nucleic acids were visualized and pictures were taken using the E.A.S.Y Win32 gel documentation system (Herolab, Wiesloch, Germany).

9.2 Supplementary tables

Table S1: Segmental duplication content and DNA copy number polymorphisms in 25 patients with unbalanced aberrations (200kb)*.

case No.	LCR content** upper breakpoint	CNPs*** upper breakpoint	LCR content** lower breakpoint	CNPs*** lower breakpoint	size of homologous sequence(kb)	sequence identity
1	2.464	-	5.796	-	146	0.995
2	7.625	-	8.512	-	265	0.983
3	1.156	+	3.031	+	0	0
4	0	-	0	+	0	0
5	2.591	+	4.161	+	43	0.941
6	1.808	+	2.787	-	38	0.987
7	**4.992**	**+**	**5.082**	**+**	**190**	**0.982**
8	3.67	-	0	+	0	0
9	0.009	-	0.079	-	0	0
10	0	-	0	-	0	0
11	0	-	0	+	0	0
12	0.01	-	0.016	-	0	0
13	0	-	0	-	0	0
14	0	-	0	-	0	0
15	**2.755**	**+**	**2.182**	**+**	**6**	**0.906**
16	**1.14**	**+**	**2.973**	**-**	**0**	**0**
17	0	-	0	-	0	0
18	0.144	-	0.052	+	0	0
19	1.533	+	0	-	0	0
20	2.532	-	1.987	+	123	0.929

21	0	-	0	-	0	0
22	0	-	0	+	0	0
23	0	-	0	+	0	0
24	0	-	0	-	0	0
25	0	-	0	-	0	0

*Three patients with previously known genomic disorders are shown in bold. The 25 cases are sorted by aberration size. Calculation is based on a 200 kb interval centered around the breakpoint.

**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone

*** CNP: DNA Copy Number Polymorphism

Table S2: Segmental duplication content and DNA copy number polymorphisms in 41 mentally retarded patients with balanced translocation (200kb)*.

case No.	LCR content** breakpoint 1	CNPs*** breakpoint 1	LCR content** breakpoint 2	CNPs*** breakpoint 2	size of homologous sequence(kb)	sequence identity
1	0.006	-	0	-	0	0
2	0	-	0.021	-	0	0
3	0	-	0	-	0	0
4	0	-	0	+	0	0
5	0	-	0	-	0	0
6	0	-	0	-	0	0
7	0.639	-	0	-	0	0
8	0	-	0	-	0	0
9	0	-	0	-	0	0
10	0	-	0	-	0	0
11	0	-	0	-	0	0

12	0	+	0	-	0	0
13	0	-	0.497	-	0	0
14	0	-	0	-	0	0
15	0	-	0.02	-	0	0
16	0	-	0.272	-	0	0
17	0	-	0	-	0	0
18	0	-	0	-	0	0
19	0	-	0	-	0	0
20	0	-	0	-	0	0
21	3.007	-	2.676	+	0	0
22	0.057	-	1.307	+	0	0
23	0	-	0	-	0	0
24	0	-	0	-	0	0
25	0	-	0.008	-	0	0
26	0.016	-	0	-	0	0
27	0	+	0.023	+	0	0
28	0	-	0	-	0	0
29	0.032	+	0	-	0	0
30	0	-	0.175	-	0	0
31	11.858	-	0	-	0	0
32	0	-	0	-	0	0
33	0	-	0	-	0	0
34	0	-	0.007	-	0	0
35	0	+	0	-	0	0
36	0	-	0	+	0	0
37	0	-	0.008	-	0	0

38	0	-	0	-	0	0
39	0	-	0	-	0	0
40	0	-	0.517	-	0	0
41	0.208	+	0	-	0	0

*Calculation is based on a 200 kb interval centered around the breakpoint.
**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone
*** CNP: DNA Copy Number Polymorphism

Table S3: Segmental duplication content and DNA copy number polymorphisms in patients with unbalanced aberrations (deVries et al., 2005) (200kb interval)*.

patient	LCR content** upper breakpoint	CNPs*** upper breakpoint	LCR content** lower breakpoint	CNPs*** lower breakpoint	Size of homologous sequence(kb)	Sequence identity
1	0.016	-	0.01	-	0	0
2	0	-	0.02	-	0	0
3	0.027	-	2.032	+	0	0
4	0.075	-	0	-	0	0
5	0.023	-	0	-	0	0
6	0	-	0	-	0	0
7	0	+	0.008	-	0	0
8	0	-	0	-	0	0
9	0	-	0	-	0	0
9	0	+	0	+	0	0
9	0	+	1.14	+	0	0
9	1.795	-	0.116	-	0	0
10	5.32	+	3.321	+	250	0.989

11	0.834	-	6.779	-	0	0
12	0.021	-	0	+	0	0
13	0.108	-	5.826	+	0	0
14	2.083	+	0	-	0	0
15	1.8	+	2.962	-	52	0.934

*Calculation is based on a 200 kb interval centered around the breakpoint.

**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone

*** CNP: DNA Copy Number Polymorphism

Table S4: Segmental duplications at breakpoints of balanced and unbalanced aberrations in patients with mental retardation (200kb interval)*

	Unbalanced aberrations**	Balanced translocations
Total No. of aberrations	22	41
Aberrations with no LCRs***flanking the breakpoints	11 (50%)	23(56%)
Aberrations with LCRs*** flanking one breakpoint	2 (9%)	16 (39%)
Aberrations with LCRs*** flanking both breakpoints, but without homology	4 (18%)	2 (5%)
Aberrations with homologous LCRs*** flanking both breakpoints	5 (23%)	0 (0%)

*Data are based on a 200 kb interval centered around the breakpoint.

** Three cases with previous known genomic disorders (7, 15 and 16) have been excluded

**LCRs: Low Copy Repeats, same as segmental duplications

Table S5: DNA copy number polymorphisms at breakpoints of balanced and unbalanced aberrations with mental retardation (200kb interval)*

	Unbalanced aberrations**	Balanced translocation
Total No. of Aberrations	22	41
No. of Aberration without CNPs*** at the breakpoints	11 (50%)	32 (78%)
No. of Aberration with CNPs*** at one breakpoint	9(41%)	8 (20%)
No. of Aberration with CNPs*** at both breakpoints	2 (9%)	1 (2%)

*Data are based on a 200 kb interval centered around the breakpoint.
** Three cases with previously known genomic disorders (7, 15 and 16) have been excluded
**CNPs: copy number polymorphisms

Table S6: Segmental duplication content and DNA copy number polymorphisms in 25 patients with unbalanced aberrations (1Mb)*

Case No.	LCR content** upper breakpoint	CNPs*** upper breakpoint	LCR content** lower breakpoint	CNPs*** lower breakpoint	Size of homologous sequence (kb)	Sequence identity
1	2.298	-	1.185	-	218	0.988
2	3.051	+	1.873	+	549	0.979
3	1.319	+	0.925	+	138	0.983
4	0	-	0	+	0	0
5	1.332	+	1.38	+	433	0.975
6	0.695	+	0.712	-	70	0.987
7	**1.318**	**+**	**1.477**	**+**	**373**	**0.979**
8	0.737	-	0	+	0	0
9	0.002	-	0.017	-	0	0
10	0	-	0.003	-	0	0
11	0.008	-	0.016	+	0	0

12	0.055	-	0.002	-	0	0
13	0.001	-	0.01	+	0	0
14	0.022	-	0	-	0	0
15	**2.649**	**+**	**2.698**	**+**	**558**	**0.966**
16	**0.471**	**+**	**1.431**	**-**	**112**	**0.986**
17	0.003	-	0	-	0	0
18	0.03	-	2.291	+	0	0
19	0.981	+	0.002	-	0	0
20	2.035	-	1.856	+	362	0.951
21	0	-	0.065	+	0	0
22	0.008	-	0.715	+	0	0
23	0	-	0	+	0	0
24	0.004	+	0	-	0	0
25	0	+	0.004	-	0	0

*Three patients with previouly known genomic disorders are shown in bold. The 25 cases are sorted based on the aberration size. Calculation is based on a 1 Mb interval centered around the breakpoint.
**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone
*** CNP: DNA Copy Number Polymorphism

Table S7: Segmental duplication content and DNA copy number polymorphisms in 41 mentally retarded patients with balanced translocations (1Mb interval)*.

case No.	LCR content** breakpoint 1	CNPs*** breakpoint 1	LCR content** breakpoint 2	CNPs*** breakpoint 2	size of homologous sequence(kb)	sequence identity
1	0.006	+	0.003	-	0	0
2	0	+	0	-	0	0

3	0	+	0.005	+	0	0
4	0.061	+	0.007	+	0	0
5	0.003	-	0.03	-	0	0
6	0.009	-	0.035	+	0	0
7	0.421	-	0	-	0	0
8	0	-	0	-	0	0
9	0	-	0	-	0	0
10	0	-	0.001	+	0	0
11	0.003	-	0	-	0	0
12	0.001	+	0.009	+	0	0
13	0.003	-	1.82	-	0	0
14	0.126	-	0	-	0	0
15	0	-	0.006	-	0	0
16	0.001	-	0.326	-	0	0
17	0.001	-	0	-	0	0
18	0.006	-	0	-	0	0
19	0	-	0.003	-	0	0
20	0.019	+	0	+	0	0
21	0.999	-	1.858	+	0	0
22	0.053	+	0.87	+	0	0
23	0.031	-	0	-	0	0
24	0.001	-	0.004	-	0	0
25	0.007	+	0.002	-	0	0
26	0.005	+	0.01	-	0	0
27	0	+	0.005	+	0	0
28	0.002	+	0.018	-	0	0

29	0.006	+	0.019	-	0	0
30	0.019	-	0.035	-	0	0
31	2.828	+	0.002	-	0	0
32	0.003	-	0	+	0	0
33	0.905	+	0.006	+	0	0
34	0	-	0.013	-	0	0
35	0.007	+	0.019	-	0	0
36	0	-	0.008	+	0	0
37	0	-	0.004	-	0	0
38	0	-	0.003	-	0	0
39	0.005	+	0	-	0	0
40	0.003	-	1.819	-	0	0
41	0.968	+	0.919	+	0	0

*Calculation is based on a 1 Mb interval centered around the breakpoint.
**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone
*** CNP: DNA Copy Number Polymorphism

Table S8: Segmental duplication content and DNA copy number polymorphisms in pateints with unbalanced aberrations (deVries et al., 2005) (1Mb interval)*

patient	LCR content** upper breakpoint	CNPs*** upper breakpoint	LCR content** lower breakpoint	CNPs*** lower breakpoint	size of homologous sequence (kb)	sequence identity
1	0.033	-	0.013	+	0	0
2	0.019	-	0.008	+	0	0
3	0.008	+	0.418	+	0	0
4	0.03	-	0.004	-	0	0
5	0.005	-	0.002	-	0	0

6	0	-	0	-	0	0
7	0.002	+	0.002	+	0	0
8	0	-	0.002	-	0	0
9	0.019	+	0.0090	+	4	0.934
9	0.004	+	0	+	0	0
9	0	+	0.471	+	0	0
9	1.145	+	0.972	-	264	0.992
10	1.321	+	1.617	+	385	0.979
11	2.395	+	4.556	+	353	0.916
12	0.004	-	0	+	0	0
13	2.857	+	2.518	+	282	0.967
14	0.751	+	0.751	+	0	0
15	0.362	+	0.919	-	113	0.934

*Calculation is based on a 1 Mb interval centered around the breakpoint.

**LCR (Low Copy Repeats, same as segmental duplications) content is calculated using the formula (∑Length of Duplication * Copy Number)/ Length of Clone

*** CNP: DNA Copy Number Polymorphism

Table S9: Segmental duplications at breakpoints of balanced and unbalanced aberrations in patients with mental retardation (1Mb interval)*

	Unbalanced aberrations**	**Balanced translocations**
Total No. of Aberrations	22	41
No. of Aberration with no LCRs*** flanking the breakpoints	2 (9%)	3 (7%)
No. of Aberration with LCRs*** flanking one breakpoint	7 (32%)	18 (44%)
No. of Aberration with LCRs*** flanking both breakpoints, but without homology	7 (32%)	20 (49%)
No. of Aberration with homologous LCRs*** flanking both breakpoints	6 (27%)	0 (0%)

*Data are based on a 1 Mb interval centered around the breakpoint.
** Three cases with known genomic disorders (7, 15 and 16) have been excluded
**LCRs: Low Copy Repeats, same as segmental duplications

Table S10: DNA copy number polymorphisms at breakpoints of balanced and unbalanced aberrations in patients with mental retardation (1Mb interval)

	Unbalanced aberrations**	**Balanced translocation**
Total No. of Aberrations	22	41
No. of Aberration without CNPs*** at the breakpoints	6 (27%)	19 (46%)
No. of Aberration with CNPs*** at one breakpoint	13 (59%)	14 (34%)
No. of Aberration with CNPs*** at both breakpoints	3 (14%)	8 (20%)

*Data are based on a 1 Mb interval centered around the breakpoint.
** Three cases with known genomic disorders (7, 15 and 16) have been excluded
**CNPs: copy number polymorphisms

Curriculum Vitae

Name, First name	Chen, Wei
Date of birth	19/11/1972
Nationality	P. R. China
Marital status	Married
Address	Department of Human Molecular Genetics (Prof. Dr. Ropers) Max-Planck-Institute for Molecular Genetics Ihnestrasse 73, Berlin 14195, Germany
Phone	(49) (30) 8413 1244
E-mail	wei@molgen.mpg.de

EDUCATION

08/2002 - present	Ph.D. program, Max-Planck-Institute for Molecular Genetics (MPIMG), Berlin, Germany Supervisors: Prof. Hans Hilger Ropers
09/1999 - 07/2002	M.Sc. program in Medical Genetics, West China Hospital, Sichuan University, Chengdu Sichuan, China Supervisor: Prof. Zhang Sizhong
09/1989- 07/1993	Bachelor of Science (major in Biochemistry), Department of Biology, Xiamen University, Xiamen, Fujian, China

WORKING EXPERIENCES

08/2002 - present	Ph.D. student, Department Ropers, Max-Planck-Institute for Molecular Genetics, Berlin, Germany Supervisors: Prof. Hans Hilger Ropers
07/2002 - 09/1999	Research Assistant, West China Hospital, Sichuan University, Chengdu, Sichuan, China Supervisor: Prof. Zhang Sizhong
08/1993 - 08/1999	Research Assistant, School of Pharmacy, Sichuan University, Chengdu, Sichuan, China Supervisor: Prof. Zhang Hao

RESEARCH FIELDS

Human Molecular Genetics and Bioinformatics

- Role of non-coding RNA (especially microRNA) in the neuron development
- Array CGH data analysis
- Molecular mechanism of chromosome breakage

- Dissection of the genetic network accounting for locus heterogeneity and clinical heterogeneity
- Prediction of candidate disease-causing gene by bioinformatical means

PUBLICATIONS

Chen W, Erdogan F, Ropers HH, Lenzner S, Ullmann R:
CGHPRO – A comprehensive data analysis tool for array CGH. *BMC Bioinformatics 2005, 6:85*

Fikret Erdogan*, **Wei Chen***, Maria Kirchhoff, Vera M. Kalscheuer, Claus Hultschig, Ines Müller, Ralph Schulz, Corinna Menzel, Thue Bryndorf, Hans-Hilger Ropers, Reinhard Ullmann:
Impact of low copy repeats on the generation of balanced and unbalanced chromosomal aberrations in mental retardation. *Cytogenetic and genome research* (In press) *Shared first author

Budny B, **Chen W**, Omran H, Fliegauf M, Tzschach A, Wisniewska M, Jensen LR, Raynaud M, Shoichet SA, Badura M, Lenzner S, Latos-Bielenska A, Ropers HH:
A novel X-linked recessive syndrome characterized by mental retardation and primary ciliary dyskinesia is allelic to OFD1. *Human Genetics* [Epub ahead of print]

Wei Chen, Lars R. Jensen, Jozef Gecz, Jean-Pierre Fryns, Claude Moraine, Arjan de Brouwer, Jamel Chelly, Bettina Moser, H. Hilger Ropers and Andreas W. Kuss:
Mutation screening of brain-expressed miRNA genes in 464 patients with non-syndromic X-linked mental retardation. *European Journal of Human Genetics* (Submitted)

Fikret Erdogan, Reinhard Ullmann, **Wei Chen**, Marei Schubert, Sabine Adolph, Claus Hultschig, Hans-Hilger Ropers, Christiane Spaich, and Andreas Tzschach:
Characterization of a 5.3 Mb deletion in 15q14 by Comparative Genomic Hybridization using a whole genome ”tiling path” BAC array in a girl with heart defect, cleft palate and developmental delay. *American Journal of Medical Genetics* (Submitted)

Lars Riff Jensen, Steffen Lenzner, Bettina Moser, Kristine Freude, Andreas Tzschach, **Wei Chen,** Jean-Pierre Fryns, Jamel Chelly, Gillian Turner, Claude Moraine, Ben Hamel, Hans-Hilger Ropers§ and Andreas Walter Kuss:
X-linked mental retardation – a comprehensive molecular screen of 47 candidate genes from a 7.4 Mb interval in Xp11. *European Journal of Human Genetics* (Submitted)

H. Najmabadi, M. Motazacker, M. Garshasbi, K. Kahrizi, A. Tzschach, **W. Chen**, F. Behjati, V. Hadavi, S. Esmaili Nieh, S.S. Abedini, R.V. Vazifehmand, S. Ghasemi Firoozabadi, Payman Jamali, Seyed Mortaza Seifati, S. Lenzner, F. Ruschendorf, A. Kuss and H.H. Ropers:

Homozygosity mapping in consanguineous families reveals extreme heterogeneity of non-syndromic autosomal recessive mental retardation and identifies 8 novel gene loci. (Manuscript in preparation)

Fikret Erdogan, **Wei Chen**, Andreas Tzschach, Ger Arkesteijn, Vera Kalscheuer, Artur Muradyan, Krause-Plonka, Marei Schubert, Antje-Friederike Pelz, Claus Hultschig, H-Hilger Ropers, Reinhard Ullmann:
Two consecutive hybridizations on DNA arrays decipher the breakpoint sequences in a balanced translocation possibly predisposing to hematological malignancy. (Manuscript in preparation)

Zhang G, Zhang S, **Chen W**, Qiu W, Wu H, Wang J, Luo J, Gu X, Cotton RG:
Go!Poly: A gene-oriented polymorphism database. *Hum Mutat. 2001 Nov;18(5):382-7.*

Chen W, Zhang G, Zhang S:
Introduction to Go! Poly, a human genome polymorphism database. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2001 Dec;18(6):482-5.*

Chen W, Zhang G, Zhang S:
Discovery of Candidate SNP by Bioinformatic Methods. *Yi Chuan,2001,23(2):153-156.*

CONFERENCE CONTRIBUTIONS

Selected talk: German Society of Human Genetics, 03/06, Heidelberg

Selected talk: Fragilome – Chromosomal Instability, Fragile Sites, and Cancer, 02/2005, Heidelberg

Selected talk: German Society of Human Genetics Congress 2003, 10/2003, Marburg

Poster: Genetic Analysis: Model Organisms to Human Biology, 01/06, San Diego

Poster: 2nd Marie-Curie Conferences on arrayCGH and molecular cytogenetics, 10/2005, Bari

Poster: European Human Genetics Conference 2005, 05/2005, Prag

Poster: German Society of Human Genetics Congress 2005, 03/2005, Halle

Poster: European Human Genetics Conference 2004, 06/2004, Munich

Poster: Human Genome Meeting 2004, 04/2004, Berlin