University of Tartu

Faculty of Biology and Geography

Institute of Molecular and Cell Biology

Department of Bioinformatics

Priit Palta

Statistical methods for DNA copy-number detection

Master's thesis

Supervisor: Prof. Maido Remm, Ph.D.

Tartu 2007

Table of contents

Abbreviations and definitions

BAC	bacterial artificial chromosome
bp	base pair(s)
CGH	comparative genomic hybridization
ddNTP	dideoxynucleotides
DNA	deoxyribonucleic acid
FDR	false discovery rate
LOH	loss-of-heterozygosity
MAPH	multiplex amplifiable probe hybridization
PAC	P1 phage-derived artificial chromosome
PCR	polymerase chain reaction
ROC	receiver operating characteristic
SNP	single nucleotide polymorphism
TI	tolerance interval(s)
TIFF	tagged information file format
WGSA	whole-genome sampling assay

Introduction

Submicroscopic changes in DNA copy-number are most likely the cause of many genetic disorders (hereditary and *de novo*), play an important role in tumorigenesis and development of various diseases. The identifications of such malformation-associated altered regions in DNA give valuable information about the genes involved in the disease and would be one step towards understanding of the molecular mechanisms beneath.

Recent developments in microarray technology have enabled genome-wide investigations of copy-number changes by means of combining conventional cytogenetic and high throughput microarray methods.

To distinguish and confirm truly aberrated regions among normal variation and spurious findings, we need computational methods that can help the investigator to estimate the statistical significance of the results obtained from microarray experiments.

The purpose of this work was to give an overview of currently used microarray-based techniques for copy-number detection. In addition, we focus on the *in silico* statistical methods used to date to assist copy-number detection.

Chapter 1
Background review

1.1 Microarray-based copy-number detection

Several molecular cytogenetic methods such as comparative genomic hybridization (CGH) and multiplex amplifiable probe hybridization (MAPH) that have been evolved to microarray format allow specific and sensitive detection of genomic copy number alterations. Moreover, array-based genotyping platforms have been used to detect previously known and novel aberrations in the genomic DNA. Currently, the above-mentioned microarray techniques have been used for studying cancer genetics, constitutional diseases and human variation.

Copy-number detection by the means of microarray techniques is an indirect way to estimate genomic copy-number of the studied DNA by measuring the quantity of fluorescent signal intensity from the labeled DNA. The easiest way to find copy-number alterations in the studied DNA is to compare it with cytogenetically controlled, normal DNA. To rule out copy-number differences found due to the normal copy-number variation, several reference DNAs are often used, either as pooled reference set or as a set containing several individual genomic profiles.

1.1.1 Array-CGH

Microarray-based comparative genomic hybridization (abbreviated as array-CGH) is a technique for the genome wide detection of chromosomal imbalances. Array-CGH builds upon well-established comparative genomic hybridization procedure that was introduced by Kallioniemi and the others (Kallioniemi *et al.*, 1992). The principle of array-CGH is that the test (studied DNA) and reference DNA (normal control DNA) are stained using different fluorescent dyes and competitively cohybridized to DNA microarrays (Solinas-Toldo *et al.*, 1997). Genomic BAC clones, cDNAs, oligonucleotides or PCR-amplified sequences that are used as capture probes on the microarray are robotically spotted or *in situ* synthesized into distinct locations of the array matrix (Mantripragada *et al.*, 2004). To avoid non-specific hybridization on the microarray (e.g. to repetitive sequences), concurrent DNA (most frequently human repetitive sequences-specific Cot-1 DNA or salmon sperm DNA) is added into the hybridization solution. After the hybridization, microarray slides are washed to remove non-bound and weakly cross-hybridized DNA sequences whereby correctly bound sequences remain bound to their capture probes on the microarray. The quantity of correctly bound sequences in each spot is estimated by measuring the fluorescence intensities of fluorescence dyes by scanning the slides. This is done separately for both dyes and independent grayscale images (typically 16-bit TIFF images) are generated for both dyes. These images are then analyzed to identify the arrayed spots and to measure the relative fluorescence intensities for each array element. After microarray scanning, image analysis and correction of intensities for various intervening variables, gain or loss in the studied DNA can be indicated from the spots showing aberrant signal intensity ratios. Signal intensity ratio S_j for the j^{th} spot on the array is calculated as logarithmic ratio of both fluorescent signal intensities:

$$S_j = \log_2\left(\frac{R_j}{G_j}\right)$$

where (and also hereafter) j is an index running over all spotted probes (N probes) on the microarray, R_j is the signal intensity corresponding to the j^{th} probe in one (most frequently red dye-specific) channel and G_j in the other (green) channel.

Differences from the expected logarithmic ratio of zero can be interpreted as copy-number differences between the studied test and normal reference genomes. Since the location-related information for the capture probes in known, acquired copy-number reports can be directly mapped to distinct regions on chromosomes, giving the genomic profile for the studied DNA (Quackenbush, 2002; Solinas-Toldo *et al.*, 1997; Veltman *et al.*, 2002). General overview of the array-CGH process is depicted in the Figure 1.1.

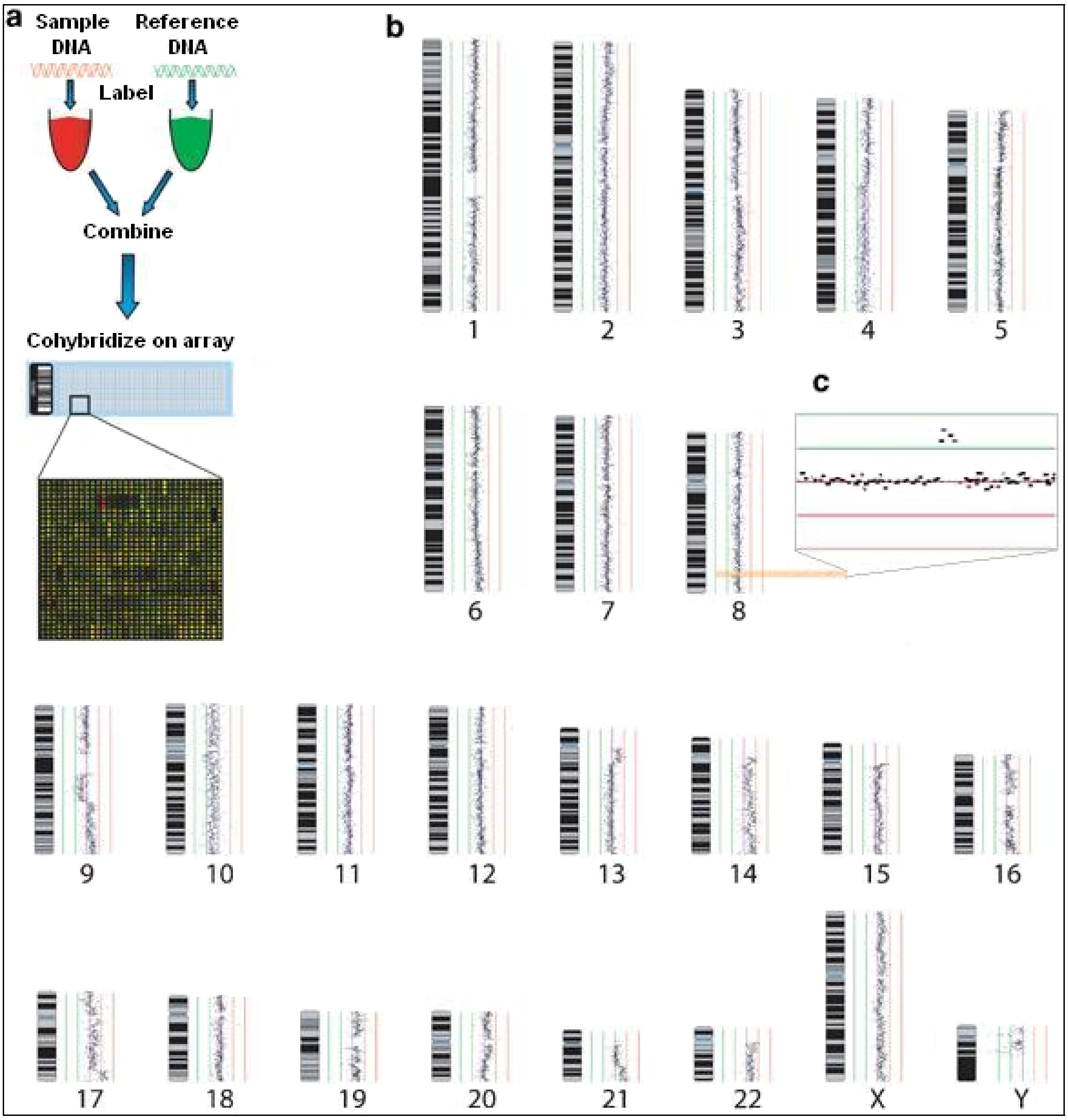

Figure 1.1. General principles of microarray-based comparative genomic hybridization. (a) Sample and reference DNA are differentially labeled with fluorescent dyes, which are typically cyanine-3 (green) and cyanine-5 (red). DNAs are mixed and cohybridized to a microarray containing spots of capture probes. The sample and reference bind specifically to those probes and the resulting fluorescence intensity ratios reflect their relative quantities. (b) Whole-genome idiogram of a small cell lung cancer cell line hybridized against a normal male reference on the submegabase resolution tiling array. Each black dot represents a single BAC clone spotted on the array. (c) Magnified view of an amplification at the *c-Myc* oncogene locus at 8q24.21 in the small cell lung cancer cell line. (Lockwood *et al.*, 2006).

1.1.2 Array-MAPH

Array-based multiplex amplifiable probe hybridization (abbreviated array-MAPH) is a molecular cytogenetic method which similarly to array-based comparative genomic hybridization is a progressive development of conventional, gel-based multiplex amplifiable probe hybridization. MAPH method enables accurate and reliable detection of changes in DNA copy-number with a theoretical resolution up to 100 bp (Armour *et al.*, 2000). Furthermore, since *in silico* designed and PCR-amplified capture probes can be made relatively flexibly, array-MAPH can be utilized for the determination of copy-number changes in virtually any targeted locus of the human genome.

The basic principle of the method is that the studied loci-specific probes can be quantitatively recovered and amplified after hybridization on a solid microarray matrix. In practice, the studied genomic DNA is denatured, immobilized on a membrane filter and hybridized to a mixture of *in silico* designed and PCR-amplified capture probes. Since in the hybridization solution the probes are in excess, every site that is recognized by the probes in the genomic DNA is occupied, so that the amount of bound probes depends on the number of available sites for each probe in the studied genomic DNA. After hybridization, filters are washed to remove non-specifically bound and unbound probes. Specifically bound probes are recovered from filters by denaturation and quantitative-phase PCR amplification. The amplified probe mixture is labeled with fluorescent dye and rehybridized to the microarray. Where, in distinct spots, there are sequences identical to probes which were initially hybridized to the studied DNA. After this, microarray slides are washed to remove non-specifically hybridized probes and scanned with microarray scanner. Raw signal intensities are extracted and capture probes and adequate signal intensities are resorted into their genomic order. Microarray signals are normalized (between-slide normalization) with respect to the control probe-specific signals on the microarrays. Raw signal intensities are transformed to logartihmic scale and the average μ_j, standard deviation σ_j and 90% tolerance interval (*TI 90%*) values are calculated for each j^{th} probe by using data from the control panel containing fluorescent signal intensities from several DNA samples of cytogenetically controlled and phenotypically normal individuals. This is done separately for both male and female reference DNAs. Signal intensity values of at least two adjacent probes deviating from the *TI 90%* values are considered indicative for the potential copy number change in the corresponding region (Patsalis *et al.*, 2007). Main steps of the array-MAPH method are described in Figure 1.2.

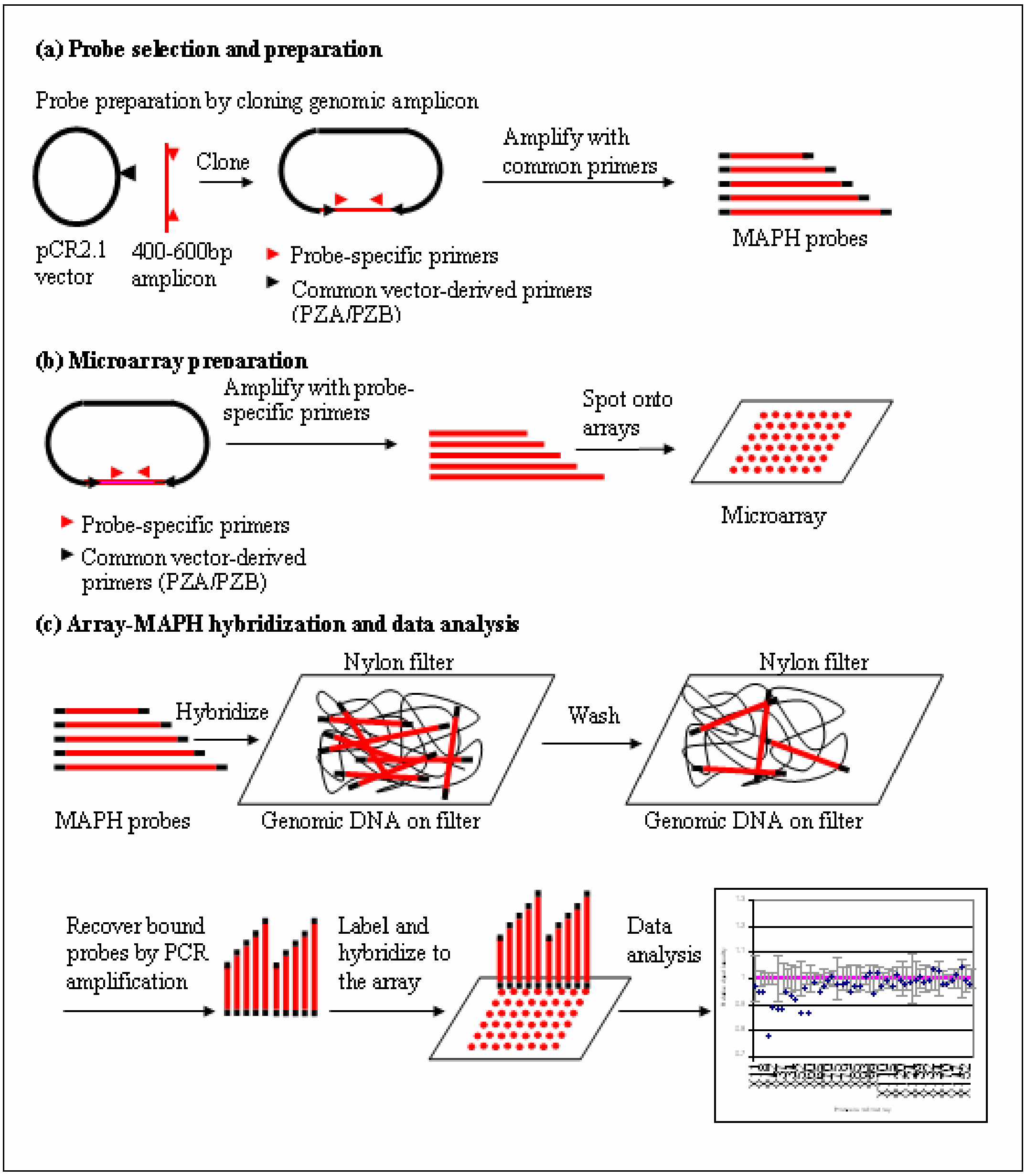
(a) Probe selection and preparation
Probe preparation by cloning genomic amplicon
Clone
Amplify with common primers
pCR2.1 vector
400-600bp amplicon
Probe-specific primers
Common vector-derived primers (PZA/PZB)
MAPH probes
(b) Microarray preparation
Amplify with probe-specific primers
Spot onto arrays
Probe-specific primers
Common vector-derived primers (PZA/PZB)
Microarray
(c) Array-MAPH hybridization and data analysis
Nylon filter
Hybridize
Wash
Nylon filter
MAPH probes
Genomic DNA on filter
Genomic DNA on filter
Recover bound probes by PCR amplification
Label and hybridize to the array
Data analysis

Figure 1.2. A flow diagram of array-MAPH methodology describing: (a) capture probe selection and preparation, (b) microarray preparation and (c) microarray hybridization and data analysis (Patsalis *et al.*, 2007).

The main drawback of the array-MAPH method is that compared to the microarray-based CGH the process is relatively laborious and time-consuming. Also, a potential drawback of single sample hybridization might be an increased influence of microarray-caused artefacts and variances. On the other hand, this approach has the advantage of being able to collect a pool of normal data for the subsequent analysis. In contrast to clone-based technologies, the method does not rely on clone availability from BAC, PAC or other libraries, as probes can be rapidly and almost unrestrictedly selected from any location in genome. Because specifically designed probes are relatively short, they can detect very small size genomic imbalances and allow developing extremely high-resolution array analyses (Patsalis *et al.*, 2007).

1.1.3 Copy-number detection with genotyping microarray platforms

In the past few years, researchers have also used genotyping platforms to detect copy-number changes in the human genome. Arrays of *in situ* synthesized short oligonucleotides originally designed for detecting single nucleotide polymorphisms (SNPs) have been used to assess DNA copy-number in the studied genome. Since this approach is regularly used in parallel with array-CGH, it is often referred to as SNP-CGH. The advantage of a combined SNP-CGH approach is the identification of allele specific gain and loss by SNP array and the robust copy number detection by array CGH (Kloth *et al.*, 2007; Peiffer *et al.*, 2006).

Genotyping platforms initially developed to qualitatively determine genotypes of the studied loci could be used for copy-number detection, since the fluorescent signal intensity from each feature (oligonucleotide probe) shows dosage-dependent response to variations in copy-number. Moreover, genotyping arrays enable to distinguish if the studied DNA has one copy from each parental chromosome or two copies of one parental chromosome, both of which will generate a signal characteristic to two copies.

The general principle of SNP-CGH is relatively simple. The test-DNA is studied in parallel on array-CGH and on genotyping platform. In the latter case, the DNA sample of interest is amplified and hybridized to microarrays carrying locus-specific capture probes. After hybridization, while amplified genomic sequences are specifically bound to the capture probes, array-based primer extension is carried out. During the primer extension step, the locus-specific oligonucleotide attached to the microarray is extended by one nucleotide (SNP) complementarily to the specifically bound (hybridized) genomic sequence. The more genomic sequences there are in the studied genomic material corresponding to locus-specific oligonucleotide probes on the microarray the more probes are elongated by one nucleotide during the extension step. The qualification (base calling) and quantification (fluorescent signal intensity level) of specific loci (the SNP nucleotide and the genomic sequence next to the SNP, respectively) is possible by employing different fluorophores that are directly (covalently linked) or indirectly (linked by immunohistochemistry) attached to the dideoxynucleotides (ddNTPs) appended in the elongation step. After the extension, arrays are washed and similarly to array-CGH and array-MAPH the proportion of correctly bound sequences and the added nucleotide for each spot is estimated by measuring the fluorescence signal intensities of fluorophores by scanning the microarray slides. In this case, this is carried out separately for all four fluorescent dyes (one responsive for each different nucleotide – A, T, C, G) and again, independent grayscale (TIFF) images are generated. Finally, by using appropriate computer software, based on fluorescent signal intensities calculated from the TIFF images, genomic profile is generated for each analyzed DNA. Genomic profile contains information about all interrogated loci (SNPs) on the microarray; their positions in the genome and their fluorescent signal intensity values. Since the studied genomic material is hybridized onto microarray alone, without controlled normal reference DNA, the copy-number for each studied locus cannot be directly estimated from its signal intensity profile as in the case of array-CGH. Rather, the DNA is analyzed in a similar manner with array-MAPH: to detect copy-number changes in the studied DNA its signal intensity profile is compared with cytogenetically controlled reference DNA-s (the control panel) (Bignell *et al.*, 2004; Peiffer *et al.*, 2006). By doing this, the ratio S_j is calculated for each interrogated loci (SNP) j:

$$S_j = \frac{x_j}{\mu_j}$$

where x_j is the studied DNA-specific signal intensity in the j^{th} locus and μ_j is equal to the average of j^{th} probe-specific signal intensities from the control panel. If the ratio is higher or lower than a predefined ratio threshold, the locus corresponding to the specific signal is considered as putatively altered.

One problem with oligonucleotide-based genotyping arrays is their higher variability that may have risen because of the intrinsic variability of the PCR-based approach used to amplify the studied genomic material (Bignell *et al.*, 2004; Lockwood *et al.*, 2006). Consequently, this raises the rate of false positives and false negatives i.e. the number of loci that appear to be aberrated (which actually are not) and the number of loci that appear to be normal, although their copy number is altered. Therefore, to diminish the rate of false positive findings, more than one-probe-specific signal is considered in the analysis. Bignell and his coworkers used three consecutive SNPs as an indication of a genomic aberration (Bignell *et al.*, 2004). Based on the analysis of ROC curves (a graphical plot of false positive rate vs. true positive rate while a sensitivity or threshold parameter is varied) Peiffer and others found that the number of false positives was minimal in case of 10-SNP "rule-of-thumb" – when they considered the average or median signal intensity of 10 consecutive SNPs (Peiffer *et al.*, 2006).

The major advantage of genotyping platforms compared to other copy-number detection methods such as array-CGH and array-MAPH is their capability of simultaneously profile the studied genome for both structural and genetic abnormalities. Moreover, simultaneous measurement of both signal intensity variations and changes in allelic composition makes it possible to detect both copy-number changes and copy-neutral events such as loss-of-heterozygosity (LOH). The studies using genotyping oligonucleotide arrays for copy-number detection demonstrate that combining genotype and copy-number analysis gives greater insight into the underlying genetic alterations, especially while studying cancer cells with identification of complex events including loss and amplification of loci.

1.2 Necessity for statistics and statistical methods used for copy-number detection

Microarray-based copy-number detection related computational and statistical requirements can be divided into three separate steps: data preprocessing (microarray data quality control and normalization), single-array methods and multi-array methods (Diskin *et al.*, 2006). Single-array methods are aimed at accurately identifying regions of gain and loss within one individual sample, including the optional characterization of breakpoints. Such methods range in complexity from simple thresholding-based decision to more sophisticated approaches that draw power from neighboring signals when making calls. Multi-array methods, which have received little attention to date, help to identify regions of consistent aberrations across multiple experiments. Such methods are required indeed to profile promiscuous for example tumor-related aberrations.

Copy-number detection even for single arrays is not a simple task and cannot be regarded as easy. The reason is that, even though underlying biological nature of DNA copy-number is always discrete (in one cell, DNA copy-number in a certain locus is always fixed to a firm level) the fluorescent signal intensity of the studied and reference DNA(s) from the microarray experiment are continuous. Therefore, there is a need for comprehensive and precise statistical methods dealing with copy-number estimation, which can help the investigator to estimate the statistical significance of the results obtained from microarray experiments (Peiffer *et al.*, 2006).

In statistics, a result is called *significant* if it is unlikely to have occurred by chance. Such random chance is measured by probability, usually referred to as *p-value*. The smaller is the probability (p-value) of random appearance of a certain result, the higher is the significance of the result. Still, it is important to note that however small the p-value, there is always a finite chance that the result is occurred by pure accident. If the p-value is smaller than predefined level of significance, we just have more statistical evidence that the event did not appear by chance. The significance level α is the probability that the null hypothesis will be rejected erroneously when it is actually true (a decision also known as a Type I error or *false positive*).

The seasures of statistical significance and statistical evidence are used in hypothesis testing. A null hypothesis (indicated as H_0) is set up to be nullified or refuted in order to support an alternative hypothesis (H_1). Nevertheless, the null hypothesis is presumed true until statistical evidence in the form of a statistical hypothesis test indicates otherwise. In case of copy-number detection, the null hypothesis usually declares that the DNA under investigation is *normal*. In case of human genomic DNA, *normal* means two copies for autosomal chromosomes or one and two copies for male and female sex chromosomes, respectively.

Since microarray data often consists of thousands of reporters (and thousands independent statistical tests are performed), one should also consider the overall rate of erroneous decisions and therefore multiple testing corrections. As described above, the significance level α is the probability that in one independent test, the null hypothesis will be rejected incorrectly when it is actually true. The correct decision will be made with the following probability

$$P_{\text{correct}} = 1 - \alpha$$

If we repeat our test for several times, e.g. ask whether all signal intensities from the current microarray represent loci with normal copy-number or not, we would also like to achieve the correct judgment for all signal intensities and appropriate genomic regions. Since all tests are considered autonomous from each other, the proper probability in the latter case would be

$$P_{\text{correct altogether}} = (1-\alpha) \times (1-\alpha) ... (1-\alpha) = (1-\alpha)^{N_t}$$

where N_t is the number of independent tests performed.

By using the opposite event, we can now calculate the probability of being wrong somewhere, i.e. that at least once, we will draw the incorrect conclusion:

$$P_{\text{wrong somewhere}} = 1 - P_{\text{correct altogether}}$$

Practically, this is the probability of false positives for the entire experiment, in the present case, for one studied microarray *m* and therefore can be re-written as follows:

$$\alpha_m = 1 - (1-\alpha)^{N_t}$$

where α is the probability of false positive result in a single test and N_t is the number of independent tests applied to one microarray-specific dataset (Draghici, 2003).

Accurate and straightforward statistical methodology is extremely important because in an actual analysis (e.g. in clinical diagnostics) with patient DNA, all aberrated regions should be found. In addition, both false positive and false negative regions should be minimal and their rate accurately predictable. It has been denoted, that copy-number studies of the human genome have false positive and false negative results and this has raised important concerns regarding the suitability of microarray-based copy-number detection for clinical diagnostic applications, since in a clinical diagnostic environment, reliable assays, providing clear and high quality results of measurable significance are required (Price *et al.*, 2005). Also, one must carefully weigh the detection of false positives (and the additional confirmation that would be needed) with the false negatives that may result in a missed diagnosis (Yu *et al.*, 2003).

1.2.1 Fixed threshold-based method

The easiest method used to identify putatively aberrated regions from a microarray-based copy number detection assay is the fixed threshold method. In this case, the deviation of probe specific signal is considered to be indicative for a putative aberration if it exceeds some fixed limit of deviation (Menten *et al.*, 2005). This method follows the ideology of the similar method used in case of expression microarray analysis – if gene *G* in condition *A* has, for example, two times higher expression than in condition *B*, it is considered to have different expression level between those conditions (Allison *et al.*, 2006; Quackenbush, 2002). Accordingly, in case of copy-number detection, regions with higher medians are considered to reflect gains and those with lower values are considered to reflect losses of the DNA material (Hupe *et al.*, 2004; Liva *et al.*, 2006).

Fixed threshold values are usually found by comparing one or several normal vs. normal experiments or by visual assessment of normal vs. normal tests (Lingjaerde *et al.*, 2005; Veltman *et al.*, 2002). The most basic possibility is to manually define the upper and the lower threshold (Kim *et al.*, 2005; Menten *et al.*, 2005), but it can be also calculated automatically (Lingjaerde *et al.*, 2005). There are also few modifications for this method – it is possible to smooth the data first and then set a threshold based on the smoothed results. Data smoothing is conducted by using moving average on the overall data (Menten *et al.*, 2005) or on the subsets of data points segmented as sets with equal copy-numbers (Chen *et al.*, 2005). The lack of fixed-threshold method is that it does not consider variance of the tests and offers no associated level of confidence (Allison *et al.*, 2006). Therefore, the use of simple ratio thresholds for calling gains and losses often leads to false negatives and can also lead to false positives (Diskin *et al.*, 2006).

1.2.2 Information-lending methods

The method described above has one crucial shortcoming – it does not consider the fact that different reporters (probes on the microarray) have different signal intensity and different variance of signal intensities. This problem is partly solved by statistically more sound methods that use overall signals from a given experiment (microarray analyzed) to estimate the normal variance of signal intensities (Allison *et al.*, 2006;

Quackenbush, 2002; Wang *et al.*, 2004). The simplest method of this kind involves calculating the mean and standard deviation of the distribution of signal intensity values and defining a global fold change difference and confidence, which is essentially equivalent to using a Z-score for the whole data set (Quackenbush, 2002). This method does not consider the fact that distinct reporters (i.e. probes) can have diverse variance in signal intensity. Rather, it just allows giving a better estimation of the possible variance, since all signals corresponding to different probes are considered in the calculation.

A more precise method proposed by Wang and his colleagues divides the studied data points into three different subsets. This is accomplished by using maximum likelihood method to fit a mixture of three Gaussian distributions (representing signals corresponding to amplification, deletion and normal copy-number) to a histogram of normalized signal intensity ratios. After dividing the data, average, standard deviation and relative proportion of the data points in each subset are found. Subset with the average signal intensity ratio closest to the zero is considered to represent the normal copy number. From that subset, 3σ upper and lower thresholds are determined. Data points from the two other subsets representing putative amplifications and deletions that fall outside the 3σ threshold are considered as aberrated (Wang *et al.*, 2004). Picard and his colleagues proposed another similar method; they assume that regions carrying different discrete copy number are seen as data sets with changes in the average and the variance of the signal intensities or changes in the average only (Picard *et al.*, 2005). The former model is also supported by other authors who have argued that despite the underlying cause, logarithmic microarray signals (that are assumed to have Gaussian distribution) having lower averages have higher variance and vice-a-versa, signals with higher averages have lower intrinsic variance (Quackenbush, 2002).

1.2.3 Permutation-based methods

If there is no data available to estimate the variance of signal intensities corresponding to each probe on the microarray, the permutation-based method should be considered to achieve more precise assessment of the variability of signal intensities, since it takes into account the peculiarity (variance, noise, etc) of the studied dataset (Yang & Churchill, 2007). Permutation-based method is somewhat mixture of fixed threshold method and information-lending method, since it uses both fixed threshold and assessment of the variability of all signal intensity values from a current microarray study. It is offering not only a nonparametric segmentation procedure, but also a nonparametric test of significance, i.e. once a single threshold parameter has been set, the method identifies the regions of copy-number change without assuming that the data follow a normal distribution and furthermore, tests the significance of these regions without making any other assumptions (Price *et al.*, 2005).

The basic principle of permutation-based method is that the actual data is randomly shuffled numerous (usually ~5000) times to count how often clearly defined genomic alteration would become to light just by chance (Diskin *et al.*, 2006; Myers *et al.*, 2005; Yang & Churchill, 2007). Genomic alteration is defined here as a region containing a certain number of consecutive signals having an average that deviates from a fixed threshold. By counting such randomly appearing genomic regions it is possible to estimate the statistical significance of such arrangements in the studied DNA. If no such regions emerge by data randomization, one can say that the actual finding was statistically significant. The statistical significance of an arrangement is estimated as the proportion of times that same or higher-scoring regions were found in numerous runs of data shuffling in which signal intensities where permuted between probes and the highest-scoring region in the permuted data was recorded in each run (Diskin *et al.*, 2006).

Figuratively, the null hypothesis of the permutation method is the studied genomic profile where originally in their genomic order positioned signal intensity values are randomly permutated (Price *et al.*, 2005). For example, if the studied genomic profile consists of N probes (studied loci) and there is putatively aberrated region, containing l probes, then there would be $N - l + 1$ permutations for this profile, each equally likely. Let us consider that X_A indicates an l_A signals long aberration starting from the j^{th} signal (probe) and that within this region, the average signal intensity $\overline{x}_A$ deviates from the predefined signal intensity threshold T. If we now run the permutation test for R_P times and additively count all randomly generated regions N_{X_P} where the length of those regions is $l_P = l_A$ and average over this region is $\overline{x}_P \geq \overline{x}_A$ (for copy-number gain) or $\overline{x}_P \leq \overline{x}_A$ (for copy-number loss), we can estimate the significance for the region X_A as the right-hand tail probability from the permutation distribution D:

$$p_{X0} = \frac{N_{XP}}{(N - l_A + 1) \times R_P}$$

It should be noticed, that the statistical significance found is strictly analyzed dataset-specific and cannot be directly compared to the probability evaluations from other analysis (Diskin *et al.*, 2006).

The main shortcoming of the permutation-based method is that the numerous data shuffling is time consuming. To overcome this problem, Myers and his colleagues proposed a simplified method where instead of calculating the distribution *D*, they approximate its probability density function with the normal distribution. If so, the statistical significance of predicted aberration is obtained by making 200 permutations of the actual data, estimating calculated parameters (average and standard deviation), and then integrating the tail of the underlying distribution beyond the observed value (Myers *et al.*, 2005).

1.2.4 Exact p-value method

As pointed out, the most basic approach to estimate copy-number is to use fixed signal or ratio thresholds to identify probes corresponding to putatively altered loci in the studied DNA. However, fixed threshold-based method or its modifications do not take into account systematic differences in each specific probe hybridization performance. In case there is additional data available (technical replicates, several reference experiments) to estimate the variance of signal intensities corresponding to each probe separately, it would be useful to employ this knowledge in more precise estimation of copy-number alterations. This would increase the predictive accuracy of copy-number determination by considering factors (e.g. repeat regions that can account greater variability in its hybridization and non-specific binding affinity) inherent in each different reporter (Margolin *et al.*, 2005).

To analyze copy-number changes in cancer cell lines with Affymetrix p501 genotyping platform, Bignell and his colleagues adopted the analysis methodology of Affymetrix expression arrays. To estimate the significance of the copy-number variation in the target cancer cell line, they compared it with a reference set consisting of 29 normal DNA samples (Bignell *et al.*, 2004). For any given SNP *j*, they assumed that its logarithmic, smoothed (mean over five consecutive SNPs) signal intensity values in the normal reference set S_j follow a Gaussian distribution: $S_j \sim N\left(\mu_j, \sigma_j^2\right)$ and which parameters where estimated by using the 29 normal reference samples as follows:

$$\hat{\mu}_j = \frac{1}{K}\sum_{k=1}^{K} S_{jk}$$

$$\hat{\sigma}_j^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(S_{jk} - \hat{\mu}_j\right)^2$$

where k=1,…K (K=29), represents the normal reference set. Assuming that the target cancer cell line has value S_j^C for the j^{th} SNP, the significance of the difference of S_j^C from the normal reference distribution S_j is measured by the p-value,

$$p_j = \min\left(1 - \Phi\left(\frac{S_j^C - \hat{\mu}_j}{\hat{\sigma}_j}\right), \Phi\left(\frac{S_j^C - \hat{\mu}_j}{\hat{\sigma}_j}\right)\right)$$

Calculated probability estimates, how likely is it for the normal population to have signal intensity values as the cancer cell line (Bignell *et al.*, 2004).

Objective of the work

The main goal of this thesis was to develop a statistical methodology for copy-number detection with the microarrays, specifically for technologies, which use control panel of normal references. The statistical method should help to identify putative copy-number changes in the studied DNA and determine the statistical significance of those findings. Additionally, our aim was to implement developed methods in command-line software programs.

Chapter 2
Material and Methods

2.1 Origin of the data

Microarray data used in this work was acquired from Dr. Ants Kurg's laboratory, Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu. The data consisted of 32 (64 subgrids) and 33 microarray slides (66 subgrids) from the human chromosome X-specific array, hybridized with cytogenetically normal male and female DNAs, respectively. Each microarray carried 558 chromosome X-specific probes, uniformly covering the whole chromosome X, and also 107 autosomal control-probes and 4 control probes from the human chromosome Y. All microarrays and studied DNAs were applied with the previously described array-MAPH methodology.

2.2 Data transformation, normalization and filtration

The first step in data manipulation was the transformation of raw signal intensities to logarithmic scale. This was done by taking logarithm with the base of two from raw signal intensity values. Logarithmic transformation helps to stabilize variation between different datasets and to adjust the raw data to normal distribution, which was required in the following data analysis. Therefore, if not denoted otherwise, all subsequent steps were carried out with logarithmic signal intensity values.

To apply between-slide normalization, signal intensities from different microarrays were rescaled with respect to the median of autosomal control probe-specific signals. This kind of 'rescaling' should make different microarrays more comparable, since although the overall signal intensity level is different on distinct array slides (due to several experimental dissimilarities, e.g. labeling and hybridization efficiency), independent use-proven control probes (as autosomal probes in case of chromosome X-specific microarray) should yield relatively invariable signal intensity values. For each microarray subgrid i, we calculated the median med_i of its autosomal control probe-specific signal intensities. Then, for each microarray, we calculated a correction coefficient c_i:

$$c_i = \frac{C}{med_i}$$

where C is previously defined baseline for the normalization, frequently used as C=2, as there are two copies of autosomal chromosomes in both male and female cells. Then, all raw signal intensities ($x_{R_{ij}}$) from i^{th} microarray were multiplied with the corresponding correction coefficient c_i:

$$x_j = x_{R_{ij}} \times c_i$$

After normalization, the median of autosomal probe-specific signal intensities should have been the same for all microarrays (and subgrids) and all other signals also more comparable.

To exclude signal intensities corresponding to deficient probes, which were not spotted on all microarrays, we filtrated the normalized microarray data. The capture probes that yielded signal intensities with a call rate less than 90% over male- of female-specific microarrays were discarded from the further analysis.

2.3 Statistical methods

2.3.1 Shapiro-Wilk test of normality

Shapiro-Wilk test of normality was used in order to validate the distribution of one-probe-specific signal intensities from the control panel consisting of normal reference DNAs. Shapiro-Wilk test considers the null hypothesis that sample $x_1,..., x_n$ came from normally distributed data set. Small values of the calculated test statistics W evidence divergence from normality and investigator may reject the null hypothesis. Based on theoretical distribution of the W statistics, we can estimate the significance of the obtained W value by p-value (Shapiro S. S., 1965). In practical evaluation of the data, we used different significance level α to evaluate if one-probe-specific data follows a normal distribution. If the test significance was smaller than α, the null hypothesis (that the data set is normally distributed) was rejected.

2.3.2 Corrections for multiple testing

To address the multiple testing problem in the copy-number detection, we used and implemented two multiple testing correction methods that help to adjust the significance level for numerous independent statistical tests. The easiest multiple testing correction is the Bonferroni method, in case of which the corrected significance level α_m for microarray m would simply be

$$\alpha_{\mathrm{m}} = \frac{\alpha}{N_{\mathrm{t}}}$$

where α is the probability of false positive result in a single test and N_t is the number of independent tests inquired from one microarray-specific data. The null hypothesis is rejected for tests, which yield p-value smaller than the Bonferroni-corrected value of α_m. Bonferroni method is relatively strict, since as the number of tests increases, the significance level quickly decreases (Draghici, 2003).

To allow less conservative adjustments of the significance level, we also used method called the false discovery rate (FDR). False discovery rate assumes that the null hypothesis is correct for all tests, i.e. there are actually no altered regions in studied DNA, and tries to control the expected proportion of wrongly rejected null hypothesis (false positives). According to the FDR method, the proved tests are ordered progressively by the significance values obtained from individual independent tests. Then, each j^{th} p-value (significance value) is compared with its specific threshold t_j, which is calculated from the formula

$$t_j = \frac{j}{N_t} \times \alpha$$

where α is the probability of false positive result in a single test and N_t is the number of independent tests inquired from one microarray-specific data. The null hypothesis is rejected for the tests, where $p_j < t_j$ (Draghici, 2003).

2.4 Programs used

Data transformation and normalization was carried out with the spreadsheet program Microsoft Office Excel 2003 (Microsoft Corporation, Redmond, WA, USA).

Shapiro-Wilk test of normality was carried out by the means of statistical software R for Windows, version 2.4.0. Within the R program, the *shapiro.test()* function was called with the *shapiro_test_for_signals.r* script.

Chapter 3
Results

As a practical outcome of this thesis, we developed two statistical methods to assist the microarray-based copy-number detection by assigning the statistical significance for putatively aberrated loci.

1. Parametric method using the distribution-based probabilities assumes that one-probe-specific data X is normally distributed ($X \sim N\left(\mu, \sigma^2\right)$) and that different probes can have different mean and variance.
2. Nonparametric method using signal intensity ranking does not require any prior knowledge about the distribution of data.

Since in the case of a usual copy-number estimation experiment there is no preliminary knowledge about the changes in the studied DNA, both methods share the same null hypothesis – the investigated DNA has normal (typically two) copy number. Therefore, we cannot directly estimate the probability for a region to be altered; we rather calculate the probability for a region to be consistent with the null hypothesis. In other words, we calculate the probability that the locus corresponding to the studied signal intensity has normal copy number. If such probability is smaller than the predefined significance level, we can consider it as an indicator for a putative copy number change.

3.1 Probe-by-probe copy-number detection method

Probe-by-probe copy-number detection method is different from the thresholding, information-lending and permutation-based methods, since in the former case all calculations are made separately for each interrogated reporter (capture probe) on the microarray. For each probe, we estimate all relevant parameters of corresponding signal intensities (e.g. average, variance and rank) by exclusively using only this one-probe-specific information. Probe-by-probe method assumes that each studied probes are independent from each other; i.e. if the signal x_{10} has a very small probability of being normal (and might correspond to a copy-number alteration), we do not make any presumptions about the condition of x_9 or x_{11} nor any other signal and appropriate locus.

3.1.1 The parametric method

The first, parametric method is somewhat similar to the method proposed by Bignall and his colleagues (both developed independently). The method is based on the assumption that one-probe specific normalized signal intensities from the different microarray experiments carried out with controlled normal individuals have similar signal intensity value. Due to the experimental variability, those readings yield in slightly different signal intensity values. Since the additional experimental noise is expected to be random and to have the normal distribution, the final signal intensities are also expected to be normally distributed (considering the fact that the family of normal distributions is invariable under linear transformations). The former is also supported by the fact that different experimental factors that add up to the final variability can have any kind of distributions but the sum of those factors still approximates the normal distribution (central limit theorem). Therefore, we can assume, that j^{th} probe-specific signal intensities in the control panel are approximately normally distributed stochastic variables with the mean $\bar{x}_j$ and with the standard deviation s_j, which we can estimate as follows:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}$$

and

$$s_j = \sqrt{\frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(x_{ji} - \bar{x}_j\right)^2}$$

In the both equations, x_i is the i^{th} signal of the j^{th} probe-specific signal intensities from the control panel and n_j is the total number of j^{th} probe-specific signal intensities in the control panel.

By using the estimated parameters and the probabilities corresponding to the standard normal distribution-specific cumulative probability function $\Phi(x)$, we can calculate the probability p_j of the j^{th} probe-specific signal intensity x_j of the studied DNA to belong to the same dataset with cytogenetically normal references:

$$p_j = \min\left(1 - \Phi\left(\frac{x_j - \bar{x}_j}{s_j}\right), \Phi\left(\frac{x_j - \bar{x}_j}{s_j}\right)\right)$$

where x_j is the studied DNA-specific signal intensity value, $\bar{x}_j$ is the average and s_j is the standard deviation of j^{th} probe-specific signal intensities from the control panel (i.e. normal references). The principal of the parametric method is depicted in Figure 3.1.

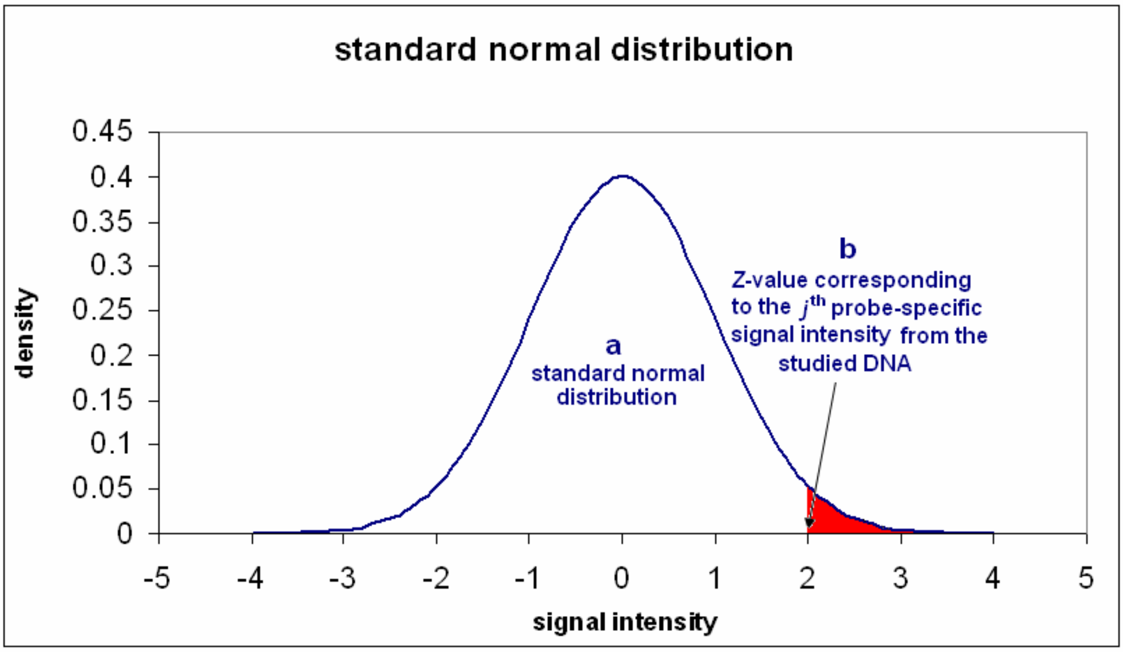

Figure 3.1. The principle of the parametric method. From normal references, the average and the standard deviation of signal intensities for the j^{th} probe are calculated. (a) The studied DNA-specific signal intensity is compared with the standard normal distribution $N(0,1^2)$. (b) For the studied DNA-specific signal intensity, the Z-value is calculated (in the latter formula, Z-value corresponds to the argument of the Φ). By using cumulative probability function $\Phi(x)$, we can calculate the probability p_j (red-colored fraction under the distribution curve) that the studied signal intensity belongs to the same dataset with cytogenetically normal references.

In biological terms, by applying the latter formula, we can evaluate the likelihood that the studied DNA-specific locus corresponding to the j^{th} probe on the microarray is from the same group with the control panel signals, i.e. it is normal. If the studied signal intensity significantly deviates from microarray signals seen for that probe before, we can mark the corresponding probe and genomic locus as putatively altered. By considering the direction of the trending signal intensity, we can deduce the underlying biological event. Signal intensity deviating upwards or below from the certain threshold can indicate putative gain or loss in the specific locus of the studied DNA, respectively.

The easiest way to use the parametric method is to define a fixed p-value threshold for the studied signal intensities. This is equivalent to the use of tolerance intervals (*TI*), in case of which the scientist defines the quantile (the lower threshold value) and the complementary quantile (the upper threshold value) for the control panel-specific dataset. Then, one can mark all probes that deviate from their specific *TI* values as altered. If the aberrated region spans more than one probe, consecutive probes should trend to the same side to be indicative for putative copy-number change.

3.1.2 The nonparametric method

If the microarray data is not normally distributed, the parametric methods will yield in incorrect results. In a real-life situation, this occurs if the data contains many outliers. If this is the case, it is helpful to use nonparametric methods that do not use exact signal intensities for copy-number detection but rather signal intensity ranking.

Our parametric method assumes that the j^{th} signal intensity of the studied DNA is normal, i.e. it shares the same copy-number with control panel-specific probes. Among the j^{th} probe-specific signal intensities from the control panel, the signal intensity corresponding to the studied DNA has a rank, which is a random discrete variable that has a uniform distribution.

If the studied DNA-specific locus is duplicated or amplified, the corresponding signal intensity ranks relatively higher and therefore most of the control panel-specific signal intensities are expected to be smaller. Inversely, if the studied DNA-specific locus is deleted, the corresponding signal intensity is lower than most of the control panel-specific signal intensities. By the discrete uniform probability function, we can then calculate the likelihood that the studied DNA-specific signal intensity x_j is ranked as X_j in the joint set of control panel-specific signals and the studied signal intensity. This probability can be calculated as follows:

$$P(X_j) = \frac{1}{n_j + 1}$$

where n_j is the total number of references corresponding to the studied probe.

Now, by additively considering the number of reference signals that are bigger or smaller than the signal intensity corresponding to the studied probe, we can estimate the probability, that studied signal intensity was normal. This probability would be

$$p_j = 2 \times \min\left((X_j - 1) \times \left(\frac{1}{n_j + 1} \right), (n_j + 1 - X_j) \times \left(\frac{1}{n_j + 1} \right) \right)$$

where X_j is the rank of the studied DNA-specific signal intensity and n_j is the total number of references corresponding to the studied probe.

The calculated probability is considered significant, if it is smaller or equal to predefined significance level α. It has to be noted, that to use strict significance levels, the reference dataset has to be relatively big.

If we now consider more than one-probe-specific rank obtained with probe-by-probe nonparametric method, we can find consecutive signals, which continually rank very high or low. For that, we can use the sliding window method.

3.2 The sliding window method

If the microarray data is noisy, simple probe-by-probe method would give the investigator a high number of false positive (if studied experiment were noisy) and false negative results (if reference experiments are noisy). Therefore, we suggest a simple sliding window method, which is more robust in case of low-quality experiments and that can be used with both parametric and nonparametric probe-by-probe copy-number detection methods. Sliding window method assumes that consecutive signal intensities are not reliant to each other and can be viewed as a series of independent data points. Still, one has to consider that there is no justified principal for applying the sliding window on signal intensities corresponding to loci that are physically located apart, i.e. on different chromosomes.

A sliding window of fixed length w is used to go through successively organized probabilities for signal intensities. A simple calculation shows, that for a genomic profile of length N there would be $N_w = N - w + 1$ of such windows.

If utilized with parametric method, the sliding window method evaluates the probability that signal intensities in the current window are normal. For that, the sequential probabilities are multiplied in each step. This is justified, because in case the signal intensities in the windows were spurious artifacts of the experiment, they would be expected to appear randomly and independently from each other. For the j^{th} window, the probability p_{wj} estimates the likelihood of a normal reference panel to contain a region of length w with such signal intensity values just by chance:

$$p_{wj} = \prod_{i=j}^{j+w} p_i$$

The product gives us the probability evaluation that the region in the j^{th} window is normal. To decide over its merit, we can compare this value with the theoretical probability for this region to be altered. The latter is calculated with the presumption that probability for one probe to deviate significantly just by chance is α – the theoretical rate of false positives. If the real-data based probability is smaller than the theoretical probability, we can mark the region in the studied window as putatively altered. In the next step, the window is shifted by one probe (and corresponding signal intensity) and new probability is calculated. The processes of window shifting and probability calculations are illustrated in Figure 3.2.

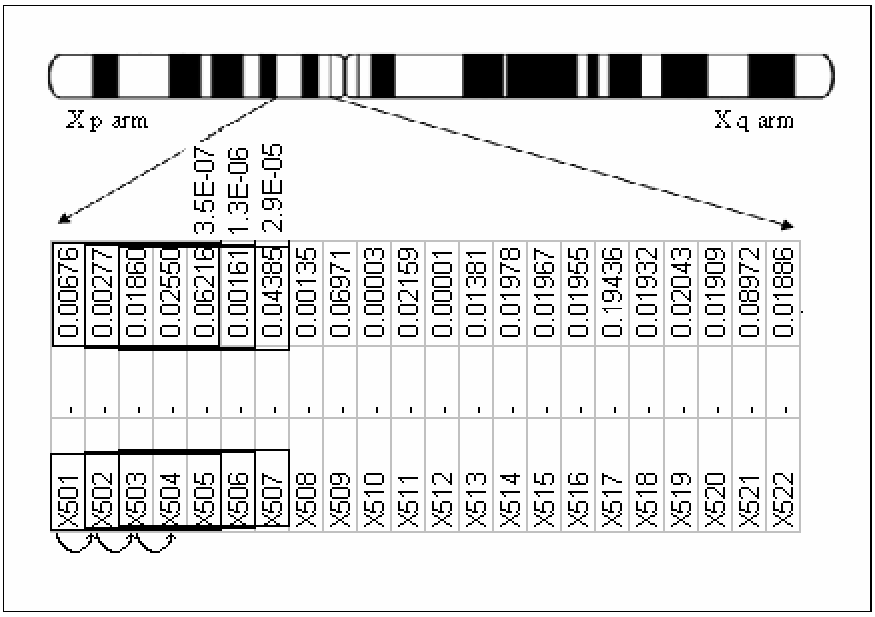

Figure 3.2. The sliding window method utilized with the parametric method. Probes and corresponding p-values obtained with the parametric or nonparametric method are ordered into their genomic order. Fixed length window (in the current example, window

length *w*=5) is shifted through the data and for each window (bordered boxes), p-values corresponding to consecutive probes are multiplied and the product is stored. Calculated products estimate the probability for the probes (and adequate genomic region) in the window to be normal.

If used in conjunction with the nonparametric method, the sliding window method evaluates the probability that signal intensities in the current window are normal by summing-up ranks corresponding to consecutive signal intensities. For the j^{th} window in length *w*, we can calculate the sum S_{wj} of the ranks:

$$S_{wj} = \sum_{i=j}^{j+w} X_i$$

where X_i is the rank of the studied DNA-specific signal intensity.

If the sum of consecutive signals is very small or big, it is considered indicative for putative copy-number changes. In addition, if the window length is longer than four consecutive probes and corresponding signal intensities, the distribution of such sums can be fairly approximated with the normal distribution. By using the latter fact, we can then estimate if the sum of the ranks in the current window is significantly high or low, which is indicative for putative copy-number loss or gain, respectively.

As noted, the sliding window method is more robust and insensitive to spurious microarray signals, allowing detection of larger aberrated regions even if a small number of probes in the studied area were acting as false negatives – appearing normal even though they were actually altered.

3.3 Normality test of experimental array-MAPH-specific microarray data

To decide, which *in silico* copy-number detection method should be more appropriate with the array-MAPH method, we analyzed 64 and 66 datasets from experiments carried out with normal reference male and female DNAs, respectively. The microarray data was transformed, normalized and refined as described in Material and Methods. For each j^{th} probe, the average $\bar{x}_j$ and the standard deviation s_j were estimated. By using one-probe-specific signal intensity parameters from normal male (one copy of chromosome X) and female (two copies of chromosome X), we estimated male and female-specific distributions of signal intensities. Example of actual and simulated (with calculated parameters) signals intensity distributions for one probe (probe X117) from the human chromosome X-specific microarray are depicted in Figure 3.3.

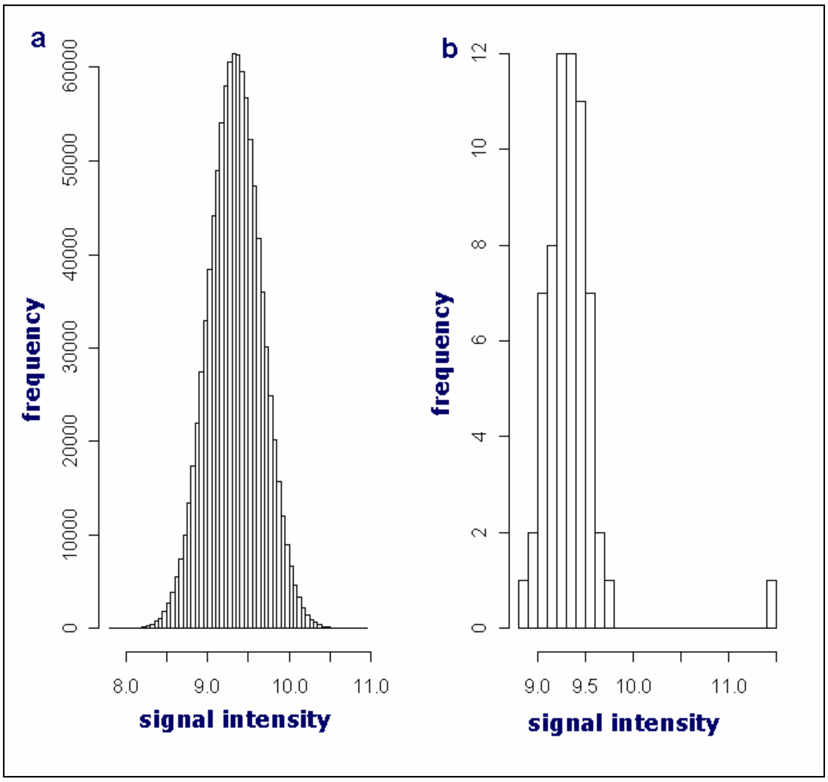

Figure 3.3 Comparison of (a) simulated and (b) actual signal intensity distributions.

Furthermore, we used the Shapiro-Wilk test to estimate if one-probe-specific microarray data follows a normal distribution, which is frequently assumed in the pertaining literature and was presumed by our parametric method for probe-by-probe copy-number detection. The results of this test are presented in Table 1.

Table 1. Results of the Shapiro-Wilk normality test conducted separately for male and female-specific data. The male-specific dataset contained signal intensities for 491 probes from 64 subgrids (32 microarray slides) and the female-specific dataset contained signal intensities for 490 probes from 66 subgrids (33 microarray slides).

Test significance level *α*	Number and percetage of rejected null hypothesis	
	Male data	Female data
0.05	213 (43.38%)	267 (54.89%)
0.01	121 (24.64%)	176 (35.91%)
0.005	88 (17.92%)	139 (28.36%)
0.001	24 (4.89%)	74 (15.1%)
0.0005	11 (2.24%)	53 (10.81%)
0.0001	0 (0%)	37 (7.55%)

As one can see from the results of the Shapiro-Wilk normality test, one-probe-specific microarray data tends to follow the normal distribution, although the number of probes, for which the null hypothesis of the normality was rejected at less strict significance level, is relatively high. This suggests that the 'normality' of the array-MAPH data is arguable and by default, the nonparametric method should be the method of choice.

3.4 Implemented programs

All developed methods for copy-number detection were implemented in the PERL scripting language. Both parametric and nonparametric methods were implemented in the program called *calc_p.pl* and the sliding window method was implemented in the *windows.pl* program. Additional and shared statistical functions were implemented in the *statistics.pm* module.

As input data, the *calc_p.pl* uses the output file from a program called MAPHStat, developed to resort and normalize array-MAPH-specific data. Input data contains probe-by-probe organized signal intensities corresponding to the studied DNA and reference DNAs. The *calc_p.pl* program also takes the number of references and used method as input parameters. From the UNIX command line, the program is executed as follows:

```
$> calc_p.pl <MAPHStat output file> <number of normal references> <method, 'p' for
parametric /'r' for ranking>
```

In principal, for each studied DNA-specific signal intensity, we calculate its probability to have the normal copy number. This is done by using the parametric on nonparametric method. The output file *p_values.txt* contains for each capture probe the probe ID, trend of the deviation (gain or loss) and the probability for the corresponding locus of having normal copy number.

The input file for the *windows.pl* program is the *p_values.txt* obtained by running the *calc_p.pl* program. It also takes the length of the window, the significance level and the multiple comparison correction method as input parameters. From the UNIX command line, the program can be run by:

```
$> windows.pl <input file containing IDs and probabilities> <number of consecutive
probes studied> <significance level> <multiple comparison correction;
'N':one/'B':onferroni/'FDR'>
```

The sliding window of length *w* is used to go through probabilities corresponding to consecutive probes. In each step, those probabilities in the current window are multiplied and the product is compared with the theoretical probability for a region of *w* probes long to be altered just by user-specified chance. This, if chosen, has been corrected for the multiple testing. The output of the program is probe IDs for significantly deviating signals, the statistical significance and the theoretical significance of findings and the trend of the deviation.

Chapter 4
Discussion

In recent years, malignant copy-number changes that have an effect on the development of cancer, congenital and *de novo* raised disorders have been studied more and more frequently. The same is true for normal variation of the genomes – in order to declare something malignant and defective, we have to know what is *normal* in the first place. While the means of microarray technology have offered high throughput techniques for copy-number studies, also straightforward statistical methods providing clear and high quality results of measurable significance are required.

Statistical methods currently used for copy-number detection lack in control over false positive and false negative results. Some methods are too liberal and some methods are too conservative. To fill this gap, we propose two statistical methods for probe-by-probe copy-number detection and a sliding window method that uses information acquired from probe-by-probe method but is more robust and insensitive to single false positive and false negative signals.

Firstly, unlike the thresholding, information-lending methods or permutation-based methods, parametric probe-by-probe copy-number detection enables to assess each probes inherent signal intensity average and variance. Therefore, there is no need to estimate the variance over all signal intensities corresponding to different capture probes and to use it as the fixed estimation of signal intensity variance for all probes. Moreover, not only does the parametric method allow estimating the statistical significance of putative copy number changes probe-by-probe, but it also helps us to predict the rate of false positive results and, if a suitable data is available, the rate of false negative results. For example, the comparison of male and female-specific chromosome X microarray data allows estimating the proportion of signals corresponding to the copy-number gain (one additional copy) and loss (heterozygous deletion) that would be missed just by chance. This is possible, since if we consider male-specific data from the chromosome X-specific microarray as normal (copy number is one), female-specific data should appear as a gain (copy-number is two) and vice versa, by reckoning the data corresponding to normal females as normal, male-specific data should look like a heterozygous deletion. If we assume, that j^{th} probe-specific data is normally distributed, the proportion b_j of false negatives in the former case can be roughly estimated as

$$\beta_j = \Phi\left(\frac{\overline{x}_{jM} + s_{jM} \times q_{\frac{\alpha}{2}} - \overline{x}_{jN}}{s_{jN}} \right)$$

where $\overline{x}_{jM}$ and s_{jM} are the male dataset-specific average and standard deviation, $\overline{x}_{jN}$ and s_{jN} are the female dataset-specific average and standard deviation. $q_{\frac{\alpha}{2}}$ is the standard normal distribution-specific $\frac{\alpha}{2}$ quantile. It is important to note, that false negatives calculated with such one-sided formula is only a fraction of all false negatives. Similar approach should be considered to calculate the false negatives from the other tail of the distribution. In the present case, it would mean taking into account the parameters of the signal intensity values that correspond to the copy-number of zero. Moreover, higher order amplifications (in our case more than one additional copy) should be consulted in a similar manner.

The main disadvantage of the parametric probe-by-probe copy-number detection method is that it assumes the normal distribution of the data. If this is not the case, the method can mediate miscalculated significances and therefore incorrect decisions. Accordingly, the distribution of the analyzed data should be appraised in advance using the parametric method.

If the data does not follow a normal distribution, our nonparametric statistical test should be the method of choice. The nonparametric method does not require any prior knowledge about the data since it is insensitive to the distribution of the data. Rather, it uses robust assumption that just by chance, half of the signals in one-probe-specific reference set should be higher and the other half lower than the signal intensity corresponding to the studied DNA. The main drawback of the nonparametric method is that in order to assign significant p-values, it needs relatively large reference dataset.

The overall size of the reference set is not directly fixed, but intelligibly, larger set will yield in a better estimation of probe-specific parameters in case of the parametric method and more precise significance level in case of the nonparametric method. Both will result in more accurate evaluation of the copy-number in the interrogated locus. Moreover, if the appropriate reference data is available for cases of copy number gain and loss, it should be involved in the analysis to facilitate the estimation of false negative results.

In practical work (scientific study or clinical screening), one should also consider the trade-off between allowed false positive and resulting false negative results. Use of less strict significance level to call putative alterations reduces the rate of false negatives but consequently, increases in the rate of false positive regions where the copy-number is assigned incorrectly. Inversely, stricter significance level will decrease the proportion of false positives and inevitably, increase the rate of false negative results, i.e. the number of missed copy-number alterations. Therefore, it is important to use optimal significance level so that in an actual analysis with patient DNA both false positive and false negative regions would be minimal and their rate accurately predictable.

If the microarray data is noisy, it is reasonable to use several consecutive signal intensities for copy-number detection. For that purpose, we present the sliding window method. Sliding window method is more robust and less sensitive to spurious microarray signals. However, since this method draws power from several neighboring probes, it also decreases the resolution of the copy-number detection experiment. Therefore, one should carefully select an appropriate length of the window. If the window is too short, obtained results will be enriched with false positives and vice versa; if the window is too long, method can miss true copy-number alterations, especially short ones.

Both parametric and nonparametric methods and the sliding window conception can be used for copy-number detection with other techniques than array-MAPH. The main requirement of our methods is that in conjunction with the studied DNA-specific genomic profile there are several normal references available. By default, this demand is fulfilled in case of copy-number detection with genotyping microarrays and can be realizable with the array-CGH method, since in clinical practice, all experiments are usually carried out with technical replicates. Furthermore, as the signal intensity data is available for normal references as well (fluorescence intensities corresponding to test and reference DNA-specific dyes are read and stored separately), it would be considerably easy to collect a reference set of normal genomic profiles from array-CGH assays. And although the statistical methods and implemented programs presented in the current work were developed primarily for the array-MAPH method, they need only small modifications to be usable with other copy-number detection methods.

Further improvements of our copy-number detection methods foresee a better incorporation of statistical significances (acquired from probe-by-probe methods) and genomic data available. For example, instead of just multiplying the probabilities representing consecutively studied genomic loci, it would be wise to weight the product with respect to the distance between simultaneously interrogated regions. It is intuitive that if the distance between those markers is small, they should have a higher probability for being both altered at the same time. Contrarily, if the distance between two virtually successive capture probes is relatively long, it is less likely that they both occurred in the same molecular event (deletion, duplication or amplification). Rather, at least on of them is a false positive result.

If the investigators have scoped out to find malignant changes in the studied DNA, it would be useful to take into account the normal variation of the genomes. For that purpose, it would be easy to check each putatively aberrated locus against public domain databases comprising information about normal copy-number variations. If the region is presented in such database, one can discard it from the further analysis.

Lastly, another improvement would be the modification of the sliding window method. Since alterations in the studied DNA are barely in the same size, it would be smart to use dynamic length for the window *w*, which would assist the investigator with more exact breakpoints for the altered region.

Summary

Copy-number alterations in human genomic DNA are most likely the cause of many human health problems, constitutional disorders and play an important role in tumorigenesis. To detect such copy-number changes, the methods most currently used compare the studied DNAs with normal reference DNAs.

In the current thesis, we described microarray-based techniques used for copy-number detection. We talked about the main concepts of copy-number detection with array-CGH and array-MAPH methodologies and genotyping platforms. We also described and discussed statistical methods that are used to assess copy-number detection with above-mentioned techniques.

As a practical work of the current thesis, we developed two statistical methods for copy-number detection with the array-MAPH method. The parametric method can be utilized if microarray data is normally distributed. If the data is noisy and includes many outliers, the nonparametric procedure should be the method of choice, since it does not assume any distribution for the data and is insensitive for the outliers. Both methods help the investigator to find putatively altered regions in the studied DNA and to assign a statistical significance for those findings.

The novelty of our methods is that the signal intensity data is analyzed separately for each capture probe presented on the microarray. The latter should improve the accuracy of the results, since it enables to consider each probes inherent signal intensity average and variance. Moreover, the sliding window method enables to find exact statistical significances over longer studied regions.

Together with small modifications to our methods and proper data manipulation, our methods can be successfully utilized with array-CGH and genotyping platforms also. The main requirement in that case is that there would be several normal references available for the data analysis.

Kokkuvõte (Summary in Estonian)

Magistritöö

Statistilised meetodid koopiaarvu määramiseks

Priit Palta

Koopiaarvu muutused inimese genoomses DNA-s mõjutavad tõenäoliselt eelsoodumust paljudele haigustele, on konstitutsiooniliste hälvete põhjuseks ja osalevad otseselt ka kasvajarakkude ja -kudede väljakujunemisel. Et sellised pahaloomulised muutused DNA koopiaarvus üles leida, võrdlevad koopiaarvu määramise metoodikad enamasti uuritavat DNA-d teiste, tsütogeneetiliselt kontrollitud normaalsete DNA-dega.

Käesoleva töö kirjanduse ülevaates kirjeldati kolme erinevat mikrokiibipõhist koopiaarvu määramise metoodikat; mikrokiibi võrdlevat genoomset hübridisatsiooni, mikrokiibi multipleks amplifitseeritavate proovide hübridisatsiooni ja koopiaarvu määramist genotüpiseerimiskiipidega. Kirjeldati ja arvustati ka ststistilisi meetodeid, mida on kasutatud koopiaarvu määramiseks eelmainitud koopiaarvu määramise meetoditega.

Teeside praktilise tulemusena töötati välja kaks statistilist meetodit koopiaarvu määramiseks mikrokiibipõhise multipleks amplifitseeritavate proovide hübridisatsiooni metoodikaga. Esimest, parametrilist meetodit saab edukalt kasutada, kui mikrokiibi andmed ei sisalda palju erindeid ja on normaaljaotusega. Kui aga mikrokiibi andmed on mürarikkad ja sisaldavad palju erindeid, tuleks kasutada mittepatameetrilist meetodit, mis ei eelda andmetele mingit eelnevalt teadaolevat jaotust ja on tundetu erinditele. Mõlemad meetodid aitavad eksperimenteerijal leida võimalikke koopiaarvu muutusega lookusi ja anda nendele leidudele statistilise olulisuse hinnang.

Meie meetodite uudsuseks on see, et signaalide analüüs viiakse läbi iga lookus-spetsiiflise proovi puhul eraldi, sõltumatult teiste proovide signaalidest. See võimaldab tõsta tulemuste täpsust, kuna iga proovi keskmiste signaalitugevuste ja signaalide intensiivsuste varieeruvuse kaudu võetakse arvesse iga proovi spetsiifilisi omadusi. Libiseva akna meetodi kasutamine võimaldab aga anda täpse statistilise olulisuse hinnangu pikematele uuritavatele regioonidele

Lisaks, meie poolt välja töötatud statistilisi meetodeid saab kasutada ka teiste koopiaarvu määramise metodikate puhul, peaasi, et lisaks uuritava DNA kiibi andmetele on kasutada ka mitmete normaalsete kontroll-DNA-de samasisuline info.

Acknowledgements

I would like to thank my supervisor Prof. Maido Remm who throughout my years in bioinformatics has guided me with insightful thoughts and proper critics. Maido, from you I have learned my affection to systematic thinking and work. In my glimpse to statistics, Tõnu Möls has been my guide. Thank you! I would like to acknowledge Ants Kurg, who has been my mentor in academic performance and writing. For teaching molecular cytogenetics and always reminding me to be careful with my words, I would like to thank Katrin Männik. I cannot forget the gratitude to my colleagues; Lauris, Reidar, Reedik, Hedi, Priit and all others, you have been the greatest one can ever wish to work with.

I would also like to thank my friends and my family for their support. Most importantly, I would like to express my gratefulness to my sweetest Mirell. Thank you for everything!

References

Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55-65.

Armour, J. A., Sismani, C., Patsalis, P. C. & Cross, G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* **28**, 605-609.

Bignell, G. R., Huang, J., Greshock, J. & other authors (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* **14**, 287-295.

Chen, W., Erdogan, F., Ropers, H. H., Lenzner, S. & Ullmann, R. (2005). CGHPRO -- a comprehensive data analysis tool for array CGH. *BMC Bioinformatics* **6**, 85.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, C. J., Jr., Weber, B. L., Maris, J. M. & Grant, G. R. (2006). STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* **16**, 1149-1158.

Draghici, S. (2003). Multiple comparisons. In *Data Analysis Tools for DNA Microarrays*, pp. 225-227.

Hupe, P., Stransky, N., Thiery, J. P., Radvanyi, F. & Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413-3422.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-821.

Kim, S. Y., Nam, S. W., Lee, S. H., Park, W. S., Yoo, N. J., Lee, J. Y. & Chung, Y. J. (2005). ArrayCyGHt: a web application for analysis and visualization of array-CGH data. *Bioinformatics* **21**, 2554-2555.

Kloth, J. N., Oosting, J., van Wezel, T., Szuhai, K., Knijnenburg, J., Gorter, A., Kenter, G. G., Fleuren, G. J. & Jordanova, E. S. (2007). Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex genetic alterations in cervical cancer. *BMC Genomics* **8**, 53.

Lingjaerde, O. C., Baumbusch, L. O., Liestol, K., Glad, I. K. & Borresen-Dale, A. L. (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 821-822.

Liva, S., Hupe, P., Neuvial, P., Brito, I., Viara, E., La Rosa, P. & Barillot, E. (2006). CAPweb: a bioinformatics CGH array Analysis Platform. *Nucleic Acids Res* **34**, W477-481.

Lockwood, W. W., Chari, R., Chi, B. & Lam, W. L. (2006). Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur J Hum Genet* **14**, 139-148.

Mantripragada, K. K., Buckley, P. G., de Stahl, T. D. & Dumanski, J. P. (2004). Genomic microarrays in the spotlight. *Trends Genet* **20**, 87-94.

Margolin, A. A., Greshock, J., Naylor, T. L., Mosse, Y., Maris, J. M., Bignell, G., Saeed, A. I., Quackenbush, J. & Weber, B. L. (2005). CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics* **21**, 3308-3311.

Menten, B., Pattyn, F., De Preter, K. & other authors (2005). arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics* **6**, 124.

Myers, C. L., Chen, X. & Troyanskaya, O. G. (2005). Visualization-based discovery and analysis of genomic aberrations in microarray data. *BMC Bioinformatics* **6**, 146.

Patsalis, P. C., Kousoulidou, L., Mannik, K. & other authors (2007). Detection of small genomic imbalances using microarray-based multiplex amplifiable probe hybridization. *Eur J Hum Genet* **15**, 162-172.

Peiffer, D. A., Le, J. M., Steemers, F. J. & other authors (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16**, 1136-1148.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. & Daudin, J. J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**, 27.

Price, T. S., Regan, R., Mott, R. & other authors (2005). SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **33**, 3455-3464.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet* **32 Suppl**, 496-501.

Shapiro S. S., W. M. B. (1965). An analysis of variance test for normality (complete samples). In *Biometrika*, pp. 591-611.

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

Wang, J., Meza-Zepeda, L. A., Kresse, S. H. & Myklebost, O. (2004). M-CGH: analysing microarray-based CGH experiments. *BMC Bioinformatics* **5**, 74.

Veltman, J. A., Schoenmakers, E. F., Eussen, B. H. & other authors (2002). High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet* **70**, 1269-1276.

Yang, H. & Churchill, G. (2007). Estimating p-values in small microarray experiments. *Bioinformatics* **23**, 38-43.

Yu, W., Ballif, B. C., Kashork, C. D. & other authors (2003). Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. *Hum Mol Genet* **12**, 2145-2152.