Normalization and analysis of high-dimensional genomics data

Mattias Landfors

Department of Mathematics and Mathematical Statistics
Umeå 2012

ISBN: 978-91-7459-402-7
Cover: Print & media, Scanned image of a microarray.
Electronic verion available at http://umu.diva-portal.org/
Printed by: Print & Media
Umeå, Sweden 2012

To my mother

Abstract

In the middle of the 1990's the microarray technology was introduced. The technology allowed for genome wide analysis of gene expression in one experiment. Since its introduction similar high through-put methods have been developed in other fields of molecular biology. These high through-put methods provide measurements for hundred up to millions of variables in a single experiment and a rigorous data analysis is necessary in order to answer the underlying biological questions.

Further complications arise in data analysis as technological variation is introduced in the data, due to the complexity of the experimental procedures in these experiments. This technological variation needs to be removed in order to draw relevant biological conclusions from the data. The process of removing the technical variation is referred to as normalization or pre-processing. During the last decade a large number of normalization and data analysis methods have been proposed.

In this thesis, data from two types of high through-put methods are used to evaluate the effect pre-processing methods have on further analyzes. In areas where problems in current methods are identified, novel normalization methods are proposed. The evaluations of known and novel methods are performed on simulated data, real data and data from an in-house produced spike-in experiment.

Keywords: normalization, pre-processing, microarray, downstream analysis, evaluation, sensitivity, bias, genomics data, gene expression, spike-in data, ChIP-chip.

List of papers

This thesis is based on the following papers:

I. Rydén P, Andersson H, Landfors M, Näslund L, Hartmanova B, Noppa L, Sjöstedt A: Evaluation of microarray data normalization procedures using spike-in experiments. *BMC Bioinformatics* 2006, **7**:300.

II. Landfors M, Fahlén J, Rydén P: MC-normalization: a novel method for dye-normalization of two-channel microarray data. *Statistical applications in genetics and molecular biology* 2009, **8**(1):Article 42.

III. Freyhult E, Landfors M, Önskog J, Hvidsten TR, Rydén P: Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC Bioinformatics* 2010, **11**:503.

IV. Önskog J, Freyhult E, Landfors M, Ryden P, Hvidsten TR: Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinformatics* 2011, **12**:390.

V. Landfors M, Philip P, Rydén P, Stenberg P: Normalization of high dimensional genomics data where the distribution of the altered variables is skewed. *PloS one* 2011, **6**(11):e27942.

Table of Contents

Abbreviations

bp	base pairs
cDNA	complementary deoxyribonucleic acid
ChIP	chromatin immunopresipitation
DE	differentially expressed
DSE-test	detection of skewed experiments test
DNA	deoxyribonucleic acid
DT	decision tree
FDR	false discovery rate
FN	false negative
FP	false positive
FPR	false positive rate
HMM	hidden markov model
IC-curve	intensity-concentration curve
i.i.d.	independent identically distributed
mRNA	messenger ribonucleic acid
NDE	non-differentially expressed
PMT	photomultiplier tube
RNA	ribonucleic acid
ROC-curve	receiver operator characteristics curve
SNP	single nucleotide polymorphisms
TF	transcription factor
TN	true negative
TP	true positive
TPR	true positive rate

Introduction to the high through-put methodology

In the last decades a lot of work has been put into mapping the genome of many species in order to understand the biological processes on a molecular level. The introduction of the microarray technology in the mid 1990's allowed for genome wide analyzes in a single experiment (Schena et al. 1995). Since then, high through-put methods such as microarrays have become quite common tools in the study of many biological processes.

The study of genome wide gene expression is a common application, where one the objectives is to identify and characterize genes involved in different processes. A medical orientated application is to identify genes that change expression during infection of a pathogen. Either to identify potential vaccine candidates or biomarkers that can be used to quickly identify the disease and prescribe proper medication. A research oriented application is to characterize the function of genes in order to construct a model for a biological process.

Gene expression can be studied either by measuring the amount of RNA (ribonucleic acid) or the amount of proteins. For both approaches high through-put methods are available, i.e transcriptomics and proteomics. Similar analyzes are also possible for studies of metabolites (metabolomics) and studies more focused on the interaction between the proteins and DNA (Deoxyribonucleic acid), e.g. protein binding pattern studies (ChIP-chip experiments). Other areas where high through-put methodologies are used are studies to identify mutations (SNPs, Single Nucleotide Polymorphisms) in the genome which may cause diseases or increase the risk of getting a disease. In recent years a methodology called deep sequencing (or next generation sequencing) has also been introduced. This is a high through-put methodology which allows for the determination the DNA- or RNA-sequence (sequencing) and the RNA expression on a genome level bringing the amount of information obtained in each experiment even higher.

Although the high through-put methods described above are highly different from a biological perspective, they are similar on the nature of the data produced. These methods provide analyzes over the entire set of all possible variables in an individual, e.g. genome wide analyzes. Therefore the number of variables studied is typically very large, ranging from hundreds up to millions depending on the processes studied. The data are also heavily affected by variation, due to the complicated experimental procedures. This

variation needs to be removed in order to draw biologically relevant conclusions.

The data analysis is often divided into two steps, pre-processing which aims to remove the technical variation and downstream analysis which is all further analyzes performed to answer the biological question posed. As such, the downstream analyzes are more dependent on the design of the experiment than the technology used. Therefore the downstream analysis mostly consists of classical statistical analysis methods, while pre-processing methods are novel methods developed for these types of experiments.

A common goal is to identify variables that differ between two treatments, e.g. which genes, proteins or metabolites are different in infected and uninfected tissue, patients and healthy individuals or virulent and non-virulent strains of bacteria. Analysis of such experiments often involves hypothesis testing of all variables, often with a small number of observations in each test. Due to the large number of tests and the small sample size, a large amount of the variables are likely to be falsely identified. The list of candidates obtained is often reduced through biological knowledge of the variables relevance to the topic studied. It is also recommended and often required for publication that the final results are verified with other methods, e.g. quantitative polymer chain reaction or other follow-up studies.

Often the underlying biological question is more complex and requires the analysis of several different treatments. To determine functions of genes, identify potential virulence factors or vaccine targets, studies of gene expression over time can be necessary, e.g. how does gene expression in a host or a pathogen change during infection. Other studies aim to identify subgroups of a disease and the genes that differ between the subgroups. In these studies the samples and variables are often analyzed using clustering methods. A more clinical application can be to predict diseases or strains of pathogens based on gene expression using classification methods. Either expressions from all genes or a sample of genes are used for this purpose. Meta analysis is not uncommon on data from high through-put methods. In these analyzes data from several different experiments are combined in order to model biological processes, e.g. determine the function of the genes or to model gene interactions.

In this thesis, data from high through-put methods for studying gene expression and protein binding patterns are used. Pre-processing methods are evaluated with respect to their performance in downstream analysis and novel pre-processing methods are developed.

Introduction to microarray technology

Introduction to gene expression

In the field of molecular biology three concepts are of greater importance, DNA, RNA and proteins. Here, a brief description of the three concepts and the interaction between them is presented as a background to the microarray technology.

With exception of viruses all living organisms are composed of cells. Each cell contains an exact copy of the DNA. The DNA contains all the genetic information of an organism and determines its characteristics, e.g. species, gender and eye color. The RNA acts as an intermediary between the DNA and the proteins, where the composition of proteins determines and regulates the states and functions of the cells.

The process in which DNA is converted into proteins is described in the central dogma of molecular biology. The dogma was first stated by Francis Crick in 1958 (Crick 1958). The DNA is transcribed into messenger RNA (*mRNA*), which in turn is translated into protein, as illustrated in Figure 1. While all proteins are translated from the DNA, all DNA is not translated into protein. In fact only a small percentage of the DNA is coding for proteins. Still non-coding DNA also has function, e.g. regulation of DNA transcription or RNA translation, but it will not be described in this thesis. Genes can be thought of as the protein coding regions in the DNA and when (m)RNA is transcribed from that region the gene is said to be expressed.

The molecular components of both DNA and RNA are called nucleotides. In DNA there are four types of nucleotides, which bind uniquely together and form four matched pairs (*base pairs*). Thus, DNA is a double stranded sequence of nucleotides, i.e. two sequences (*strands*) of nucleotides bound together by hydrogen bonds. The strand transcribed into RNA is often referred to as the coding strand and the other the template strand. The coding strand is said to be complementary to the template strand and consequently also to the RNA, as illustrated in Figure 1. When a gene is expressed, the protein coding region of the coding strand is separated from the template strand and transcribed into RNA (*RNA transcript*). Thus, RNA is a shorter single stranded sequence of nucleotides. In the translation process, triplets of nucleotides (*codons*) in the RNA transcript are converted into amino acids. The protein is the sequence of amino acids formed through this process.

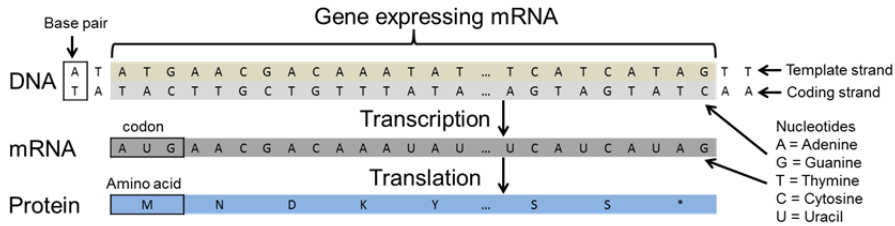

Figure 1. The DNA consists of four matched pairs of nucleotides (base pairs). When a gene is expressed, the sequence of the coding strand is separated from the template strand. The coding strand is then transcribed into a single sequence of nucleotides (mRNA), where one of the nucleotides is replaced. In the translation process, amino acids are formed from triplets of nucleotides (codons) in the mRNA. The protein is the amino acid sequence translated from the mRNA.

Introduction to gene expression microarrays

While the goals of microarray experiments can be quite diverse, a common use of microarray experiments is to compare the gene expression between different samples. Samples generally refer to biological replicates of cells, e.g. cells extracted from cell lines or different individuals. Henceforth we consider comparisons between treatment samples and reference samples. Treatment samples are samples treated in one or more ways, e.g. infected, cancer cells or different strains of bacteria. The reference samples can be considered to be the 'natural state' of the sample, e.g. uninfected, healthy cells or wild type strain.

The comparison of gene expression between the samples is done by comparing the amount of mRNA transcribed by genes in both samples (on RNA level). Genes transcribing different amount of RNA, i.e. different gene expression, in the samples are called differentially expressed.

The microarray

There are many types of microarrays, e.g. Affymetrix, Agilent, spotted cDNA (*complementary DNA*) arrays and spotted oligo arrays. Although they are constructed in different ways, the general idea behind them is the same. The array is composed of thousands of probes. A probe consists of many copies of single stranded DNA. The single stranded DNA in each probe is complementary to a specific RNA transcript. Depending on the method used to construct the arrays the DNA in the probes will differ in length, Affymetrix has a length of 25 base pairs (bp), oligo arrays around 70 bp and cDNA arrays around 500 bp.

For spotted cDNA arrays the single stranded DNAs are dissolved in a solution and dropped (spotted) on to the physical array by small pins. The physical array is usually a glass slide about 2x5 cm in size, but other types of arrays also exist. It should be noted that for spotted arrays the probes on the array are grouped in several small areas (sub-grids), where each area is spotted with the same pin.

Since all RNA-transcripts are not known in all species, some probes represent hypothetical genes, i.e. a DNA region that is likely to code for a protein. These hypothecial genes may or may not be correct. However, in the subject of gene expression microarrays, genes will henceforth refer to probes even though all probes do not necessarily correspond to an actual gene.

Experimental procedure

In order to describe the problems faced in microarray data analysis, a short description of the experimental procedure is provided. An overview of the procedure is visualized in Figure 2.

Extraction, labeling and hybridization

In the first step of the microarray experiment total RNA is extracted from the samples. Note that total RNA represents the (m)RNA transcribed by all genes currently expressed. The (m)RNA from the samples are labeled with a fluorescent dye and reverse transcribed into single stranded DNA. The labeled extract is deposited on the microarray. In a process, called hybridization, the cDNA from the sample binds to the single stranded DNA of the corresponding gene. After the hybridization the arrays are washed to remove any unbound labeled extract from the array.

Scanning

After hybridization and washing the arrays are scanned using a confocal laser microscope, which produces an image of the array. The laser sends a beam of light at focused part (5-10 micrometer) of the array. The wavelength of the light corresponds to the wavelength of the fluorescent dye. When hit by the light the dye molecules emit fluorescent light. The intensity of the light emitted corresponds to the amount of dye molecules and consequently the amount of mRNA present. The array is traversed and the intensities are stored as pixels in a 16-bit image of the array, as illustrated in Figure 3. Note that light from the laser can be reflected from the surface of the array and any unbound DNA will also emit light which will be recorded. This light is generally referred to as background noise.

Scanners often use photo-multiplier tubes (*PMT*), which multiplies the amount of photons. This increases the intensity and makes it possible to identify even small amount of light.

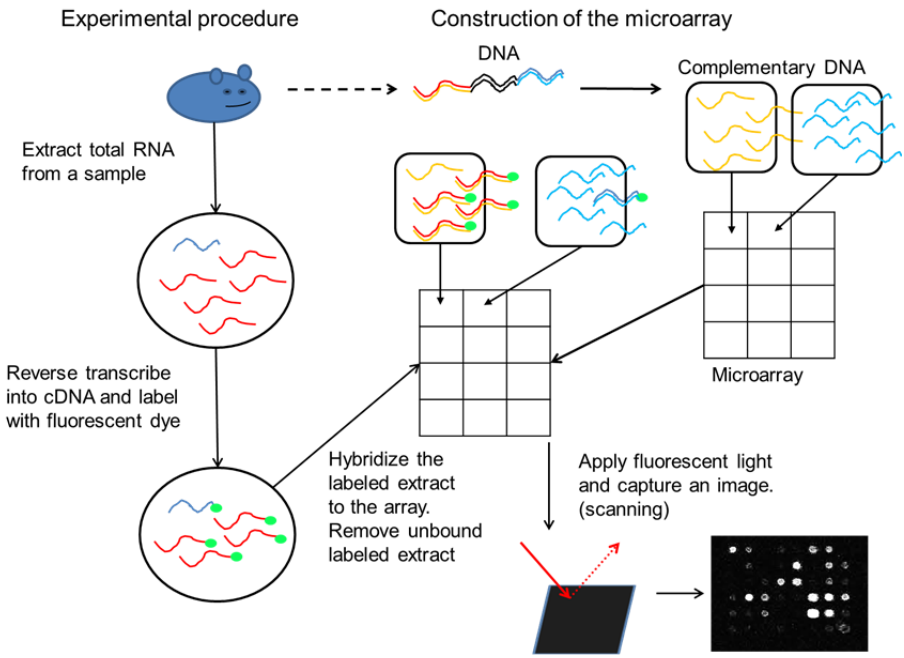

Figure 2. Experimental overview of a gene expression microarray experiment. The microarray is constructed from single stranded DNA complementary to the RNA transcripts of interests, e.g. all genes. The RNA-transcripts in a sample are extracted, reverse transcribed into single stranded DNA and labeled with a fluorescent dye molecule. Here, the blue and the red lines represent two RNA-transcripts. The labeled extract is poured onto the array and in a process, called hybridization, the DNA in the sample binds to the complementary DNA on the array. The arrays are then washed to remove unbound labeled extract. After hybridization and washing the arrays are scanned. In scanning, light from a laser is focused on the array and the fluorescent light emitted from the array is recorded in an image.

Image analysis

In order to acquire data which can be analyzed, the images are processed in an image analysis program. The image analysis consists of two parts, segmentation and data extraction. Usually a grid containing the information about location and identity of the probes are provided by the array manufacturer. For spotted arrays the locations are not always exact and must first be identified. Image segmentation usually refers to a process which identifies sub-areas that differ from its background. For microarrays this refers to the process of identifying the pixels which correspond to a probe, e.g. finding the edge of the probes. There are many segmentation methods

employed in different image analysis programs. For a description and comparison of the methods implemented in commonly used image analysis programs, see Yang *et al.* (2002).

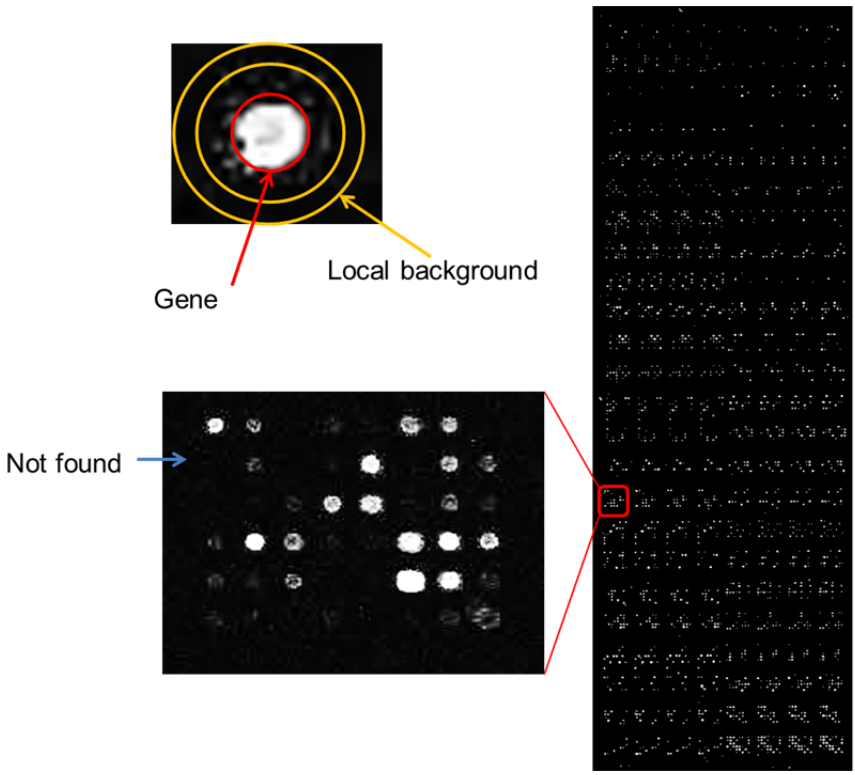

Figure 3. An image of a scanned microarray and visualization of a gene marked as not found. An example of image analysis is to define the genes as circles of variable sizes. The intensity of the pixels inside the red circle corresponds to the gene. The local background can be estimated as an area outside of the gene. Here, the pixels between the two orange circles are used to estimate the local background.

Once the probes have been identified the grid is aligned. This process is automatic in most image analysis programs. However, manual verification and correction of the grid placement is often needed. The main reason for the necessity of the manual verification is due to probes not emitting any light beyond background level. For these probes, segmentation will not work and they can cause displacements in the grid alignment. These probes are marked as *not found* by the image analysis program. To identify the pixels of these probes, a circle of size and location specified in the grid is often used. Note that this may be a different approach than for probes identified through segmentation. To help with manual verification and correction a number of probes emitting a lot of light are often spotted on the arrays. The manual

correction can also be complemented with identification of probes where the spotting or hybridization has failed.

After the segmentation and alignment of the grid, the mean and median intensity of the pixels in each area containing a probe are calculated. In addition to the intensities of each probe, an estimate of the background noise is also provided. For spotted arrays this usually consists of calculating the mean and median of an area outside of the probe, as illustrated in Figure 3. The image analysis programs also provide different quality control measures including a *flag* indicating if the probe was identified (found), not found or marked as bad by either user or program.

Technical variation

Variation is one of the primary problems dealt with in the field of statistics and it can be either random or systematic. While random variation occurs in any experimental data, systematic variation is less common. Systematic variation refers to biases or patterns in the measurements caused by identifiable sources, e.g. un-calibrated equipment reporting too low or too high values. For biological experiments the variation can be divided into biological and technical variation. Biological variation represents difference between biological sources, e.g. different individuals will not have exactly the same amount of expression in the genes. Technical variation refers to differences between repeated measurements which are caused by the experimental procedures and not being true differences in the biological material. In a good experimental design the biological variation is only considered to be random.

Data obtained from a microarray experiment are highly affected by technical variation. This is due to the complicated experimental procedures involved and the nature of the methods used. This variation is both random and systematic. The systematic variation will likely distort the biological interpretation of the data if it is not removed.

Systematic variation of microarray experiments

In general it is assumed that the observed increase of signal intensity is proportional to the increase of the amount of mRNA, i.e. increase in *concentration*. However, the systematic variations present in microarray data will both distort the linearity and cause difference in the constant of proportionality between experiments. Microarray experiments can be categorized as one or two channel experiments. In one channel experiments each array corresponds to one sample. In two channel experiments a

reference and a treatment sample labeled with different dyes are hybridized to the same array. The systematic variations of microarray experiments are usually divided in three separate categories, dye and array error, background noise and saturation.

The *dye error* is considered to originate from the difference in emission strength (quantum yield) of the two fluorescent dyes used in the two channel systems. This will cause one channel to systematically report higher intensities than the other (i.e. larger constant of proportionality). In theory, the expected difference originating from the difference in dye emission should be constant. Although the largest effect in the dye error is due to the difference in the dyes, what is estimated as dye error is often a combination of errors originating from several factors (many of which the effects are incalculable or inseparable).

In one channel experiments, the same dye is used for all samples. The *array error* is considered to originate from the difference in hybridization effect, washing and amount of material loaded on to the arrays. This error causes a difference in the proportionality between arrays, thus distorting the ratio between treatment and reference. All of these errors are also present in the two channel experiment. However, the effects between arrays are considered to be a minor issue as both the reference and the treatment are hybridized to the same array. The effects are therefore either considered to affect both on an equal amount or incorporated in the dye error.

The second most studied systematic variation is the effect of background noise. *Background noise* is considered to originate from several sources, such as unwashed (labeled) mRNA, stray light, dust, light reflections from the surface of the array and dark noise (caused by small electric currents passing through the PMT). This causes signals with low intensity to be censored from below, i.e. no genes have lower values than the background level.

The last of the systematic variations is called *saturation*. It originates from limitations in the resolution of the scanners. The scanners of today are limited to 16-bit images, thus pixels with an intensity higher than 2^16-1 will be truncated, i.e. all signals above this value will be set to the value. When the concentration increases for probes with high concentration more pixels become saturated. This causes the increase in the mean intensity (i.e. gene intensity) to decrease, until all pixels are saturated and no increase is observed. It is recommended to use the median to calculate the gene intensity as it less affected by the decrease, at least until 50% of the pixels are saturated.

Introduction to ChIP-Chip experiments

Mechanisms of gene regulation

As previously mentioned, genes encode proteins and proteins determine the function of a gene. The largest class of proteins is called *transcription factors* (TFs), which are proteins that bind to the DNA close to a gene, i.e. promoter region. When bound, a TF serves as an activator or repressor of mRNA transcription for the gene, thus regulating gene expression.

Another mechanism of gene regulation is chromatin modification. In eukaryotic cells, i.e. cells with a nucleus, the sequence of DNA molecules is located within the nucleus. In human cells the approximately 2 meters long sequence of DNA-molecules is compacted down to a 10 μm space. In the first level of compaction the DNA is at regular intervals wrapped around an assembly of specific proteins (*histones*) forming a beads on a string like structure. The DNA is further compacted, although the complete mechanism of compaction is at this date unknown. The complex of DNA, histones and structural proteins that makes the structure of the DNA within the nucleus is called chromatin. One of the chromatin modifications that affect gene regulation are changes to the structure of the histones through post-translational modifications, making the DNA more or less accessible for transcription factors.

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) is an experimental technique for extracting and characterizing DNA bound by chromatin complexes. Such characterizations can be identification of binding sites of TFs or histone modifications. Here, a brief description of experimental procedure of (cross-linked) ChIP is described.

The protein bound to the DNA in the cell is stabilized (cross-linked), i.e. the bonds between the DNA and the proteins, including those within the chromatin, are strengthened. The DNA is then extracted and sheared usually through sonication. A protein specific antibody coupled to a magnetic bead is used to extract the chromatin bound to the protein. The bond between the DNA and proteins are then destroyed (reverse cross-linked) and the DNA fragments can be analyzed. One such analysis is to identify the binding sites of a protein, e.g. promoter regions, through ChIP-chip experiments.

ChIP-chip experiments

In ChIP-chip experiments the chromatin immunoprecipitated DNA (treatment) is compared with either input DNA, i.e. DNA immunopresipitation without antibody, or DNA immunopresipitated with a mock antibody (reference). The objective of the ChIP-chip experiment is to identify regions in the data where the protein is bound rather than to identify differential gene expression.

The microarrays used in ChIP-chip experiments are designed in a different manner than the gene expression arrays. For gene expression experiments the probes on the array each correspond to a gene, for ChIP-chip experiments the probes represent small parts of the entire genome. These parts are chosen sequentially along the genome with either a small overlap or with a small distance in between, depending on the size of the array and the size of the genome. Due to this design the number of probes is a lot larger than the gene expression arrays, ranging in millions compared to tens of thousands.

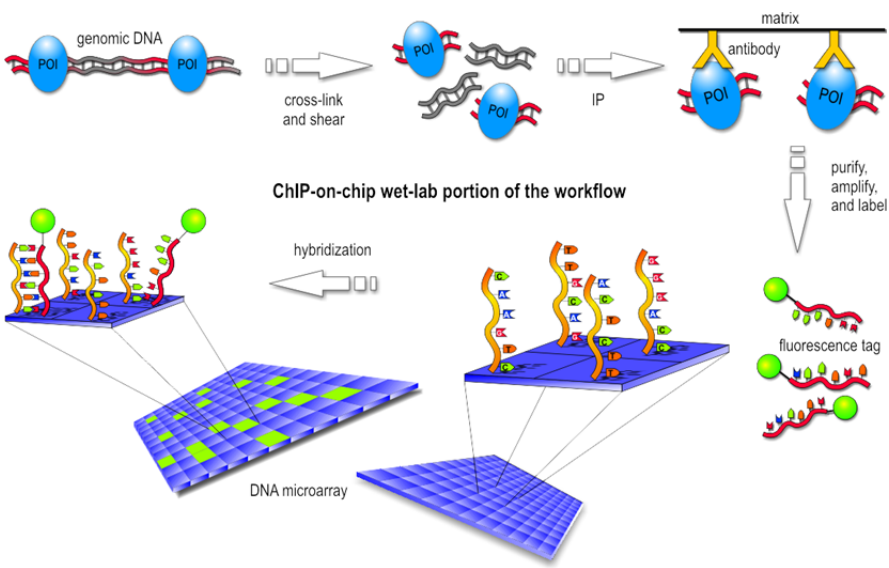

Figure 4. An overview of ChIP-chip experiments, the protein of interest (POI) is bound to the DNA in vivo, the DNA is extracted and sheared. An antibody specific to the protein of interest is used to immunoprecipitated the DNA. The Immunoprecipitated DNA is then labeled with a fluorescent dye molecule and hybridized to a microarray, (Courtesy of author Thomas Hentrich, wikipedia)

The experimental procedures of labeling, hybridization and image analysis and consequently the technical variation for ChIP-chip experiment are essentially the same as for gene expression array. However, the protein

bound regions will span several probes on the array and ChIP-Chip experiments will therefore have a dependency structure for probes that are close to each other in the genome. Although this does not affect pre-processing it does complicate the downstream analysis.

Pre-processing

Pre-processing aims to remove the technical variation in the data and is a vital step in order to infer biologically relevant conclusions. Pre-processing usually consists of several steps; filtration, background correction, censoring, saturation correction and dye or array normalization. Filtration describes how to process probes flagged in the image analysis and censoring is a complement to background correction.

Data model

Two general assumptions are that the observed signal is proportional to the mRNA concentration and that the background noise is additive. Thus, a simple model of the expected value of the observed signal (Y) for a gene would be the linear relationship,

$$E(Y) = \alpha + v\gamma.$$

where the intercept α describes expected value of the background noise, the proportionality constant v describes the expected value of the array error and γ the expected mRNA concentration. Incorporating random variation in the model gives,

$$Y = \alpha + v\gamma + \varepsilon + \delta.$$

Here, ε denotes the zero mean technical variation and δ the zero mean biological variation. Since the primary object is to model the technical variation, δ is assumed non-existent throughout this thesis. A similar model was introduced by Rocke and Durbin (2001). Their model is,

$$Y = \alpha + e^{\eta}\gamma + \varepsilon,$$

where the background noise is assumed to be constant, the proportionality constant is $v = e^{\eta}$, where $\eta \sim N(0, \sigma_{\eta})$ and the technical variation is $\varepsilon \sim N(0, \sigma_{\varepsilon})$.

In general the normalization methods make assumptions about the background noise and array errors. A more general model would assume that the expectations of both background noise and array error are non-constant across all n arrays and all N genes. This yields the model,

$$Y_{ij} = \alpha_{ij} + \nu_{ij}\gamma_i + \varepsilon_{ij},$$

where the subscripts i and j denote genes and arrays respectively.

Background correction

In general, the background corrected signals (X) are obtained by subtracting an estimate of the background noise,

$$X_{ij} = Y_{ij} - \hat{\alpha}(i,j).$$

Considering non-expressed genes, i.e. genes with $\gamma_i = 0$, the observed signal are observations of the background noise ($Y_{ij} = \alpha_{ij} + \varepsilon_{ij}$). Under the assumption that the background noise is constant across genes, a natural estimator is the average of the non-expressed genes, if any of these are known.

In the general model the background noise is assumed to be non-constant across genes and arrays. Under this assumption it would not be possible to estimate the background noise of the individual genes without further measurements. For oligo and cDNA type arrays the image analysis programs provide an estimate of the signal for pixels in a local area outside of each gene. This signal can be used to estimate the background noise in the observed signal of the gene. Although this is generally labeled as local background correction, its performance may depend on the different estimation methods used by the image analysis programs. See Yin *et al.* (2005) for a description of the local background estimation methods in the most commonly used image analysis programs.

For the local background correction we can make the assumption that the background noise present in the observed signal and the estimated local background are independent identically distributed (*i.i.d.*) random variables. The background corrected signals would then be reduced to,

$$X_{ij} = \nu_{ij}\gamma_i + \varepsilon_{ij} + \varepsilon_b,$$

where ε_b is a symmetrical random variable with mean zero.

Under the i.i.d. assumption, the background corrected signals of non-expressed genes are reduced to $X_{ij} = \varepsilon_b$. Consequently half of the signals are expected to have negative values. It is often considered that approximately 60% of all genes are non-expressed. Hence, assuming a correct background correction, approximately one third of the background corrected signals are expected to be negative. Other than being non-intuitive (negative signals are non-existent) a negative signal in itself is not a problem. The problem of negative values arises when genes are negative in one treatment and positive in the other. The log transformed ratio of observed signals between the two treatments (*log-ratio*) is often considered. This cannot be calculated for negative values and these genes become problematic to normalize further. A possible solution is to exclude these observations. This may cause the observations to be too few for further analysis of a particular gene. Consequently this gene is excluded from further analysis. It is likely that most of these genes are non–expressed in all treatments. However, it is also possible that some of these genes have changed between being non-expressed and being expressed, due to the treatment. These highly interesting genes will likely be missed if genes with negative values are excluded. An alternative solution, is censoring the data so that all genes have at least a certain value. Censoring genes is a more sensible approach than excluding them. However, the problem of properly choosing a censoring value is to my knowledge not solved.

The major problem with background correction is an increase in the variance of the log-ratios when it is applied, as illustrated in Figure 5. This increase in variance has a negative impact on the performance of certain downstream analyzes (Qin and Kerr 2004) and is most problematic for genes expressed at a low level. Due to this negative impact, whether or not to apply background correction has been a topic for discussion (Scharpf et al. 2007).

The background estimates and background correction methods for Affymetrix type of arrays are quite different from cDNA and oligo arrays. However, the two major problems described above also exist in this type of arrays. Consequently, many of the background correction methods proposed for either type, addresses the problems of the increase in variance and the problem of negative values. For a description and comparison of background correction methods for two channel microarrays, see e.g. Ritchie *et al.* (2007). For a comparison of background correction and normalization methods for Affymetrix type of arrays, see e.g. Zhu *et al.* (2010).

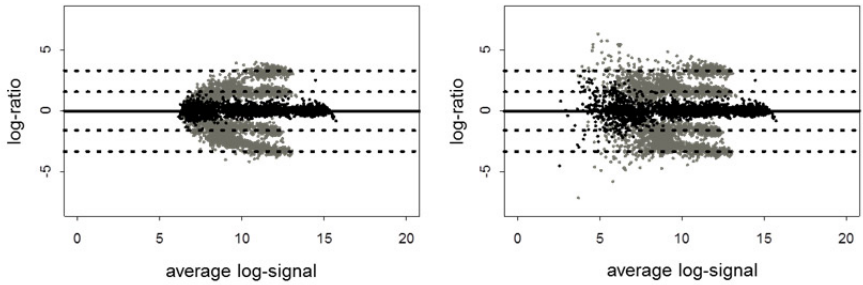

Figure 5. Log-ratio vs. the average log-signal of dye normalized data without background correction (A) and with background correction (B). The solid lines represent the true log-ratio of the NDE-genes (black dots) and the dotted lines the true log-ratios of the DE-genes (grey dots). The picture illustrates the increase of the variance for genes with low concentration as background correction is applied.

Saturation correction

In microarray terminology saturation is a (right) truncation of the pixels of a gene and the estimated signal is usually calculated as mean or median of several pixels. The most common approach for saturation correction is to set the laser power/PMT settings on the scanner to a level so that an acceptable low number of genes are severely affected by saturation. Although this approach is possibly the best approach for saturation correction, lowering the scanner level will increase the amount of signals at, or below, the background level. Thus, it will cause further problems due to background noise. If we consider the genes calculated as the median of the pixels, then the effect of saturation on gene level is essentially a right truncation too. Thus the signals can be modeled as,

$$Y_{ij} = \min\left(\alpha_{ij} + v_{ij}\gamma_i + \varepsilon_{ij}, 2^{16} - 1\right).$$

A few approaches have been proposed which commonly require the arrays to be scanned at several different scanner settings and combined (Dudley et al. 2002; Bengtsson et al. 2004; Lyng et al. 2004).

The Restricted Linear Scaling in Paper I is a slight modification of the method proposed by Dudley *et al.* (2002). Consider an array scanned at two different scanner settings, then the model can be written as,

$$Y_{ijk} = \min\left(\alpha_{ijk} + v_{ijk}\gamma_i + \varepsilon_{ijk}, 2^{16} - 1\right),$$

where $k = 1{,}2$ denotes the scan.

As the signal observed is an increasing function of the laser power/PMT setting, then we can denote the scans so that $\nu_{ij2} > \nu_{ij1}$ for all arrays and genes. Let I denote the interval in where $Y_{ijk} \approx \nu_{ijk}\gamma_i$ for both scans, i.e. where the signals are not heavily affected by either background noise or saturation. Regression is used to estimate Y_{ij2} as a linear function of Y_{ij1} within the interval I and used to predict signals saturated in Y_{ij2}. The saturation corrected signals (for median intensities) are thus obtained through,

$$\hat{Y}_{ij2} = \begin{cases} Y_{ij2} & Y_{ij2} < 2^{16} - 1 \\ a + b * Y_{ij1} & Y_{ij2} = 2^{16} - 1 \end{cases}.$$

Here, a and b are the regression coefficients estimated in the interval I. Note that genes saturated in the lower scan are not possible to correct in the higher scan. To correct these, the signals in the lower scan can first be corrected with a third (even lower) scan. The procedure can thus be extended until the lowest scan does not include any saturated genes.

Array and dye normalization

In this section we consider a model of the expected values of background corrected signals without saturation, i.e.

$$E(X_{ij}) = \nu_{ij}\gamma_i,$$

and focus on the models and normalization methods of the array error ν. For simplicity we will consider only a model with two samples, where the one channel model is equivalent to the two channel model.

$$E(X_{it}) = \nu_{it}\gamma_{it} = 2^{C_{it}}2^{\mu_{it}} = 2^{C_{it}+\mu_{it}}$$

$$E(X_{ir}) = \nu_{ir}\gamma_{ir} = 2^{C_{ir}}2^{\mu_{ir}} = 2^{C_{ir}+\mu_{ir}},$$

where C denotes the log-transformed dye (or array) error with base 2 and μ the log-transformed concentration with base 2.

Considering the model it should be noted that that given the observed values, the value of C or μ cannot be quantified. However, in microarrays the relative expression between two samples is studied. Therefore, the log transformation with base 2 of the ratio between the treatment and the reference samples is used. Henceforth we will refer to this as the *M-value* (or log-ratio),

$$M_i = log_2\left(\frac{X_{it}}{X_{ir}}\right) = log_2(X_{it}) - log_2(X_{ir}) =$$
$$C_{it} - C_{ir} + \mu_{it} - \mu_{ir} + \varepsilon_{mi} = \Delta_i + \theta_i + \varepsilon_{mi}.$$

Here, Δ denotes the difference in logarithm of dye errors and θ the difference in logarithm of mRNA concentrations between the samples (*true log-ratios*) and ε the zero mean random error. If we can obtain an unbiased estimator of the dye errors, the normalized M-values can be obtained,

$$E(M'_i) = E\left(M_i - \widehat{\Delta}_i\right) = \theta_i.$$

Normalizing the M-values corresponds to normalizing the signals so that the dye errors are equal for all samples (but not removed). If we consider *non-differentially expressed genes* (NDE-genes), where the concentration is equal for both samples, then $E(M_i) = \Delta_i$. Assume that the dye error is constant across the array ($\Delta_i \equiv \Delta$), then an unbiased estimator is the average M-value across the NDE-genes. Obviously, which genes are NDE-genes is not known, however it is believed that under most conditions only a small fraction of the genes are differentially expressed. Thus an average M-value of all genes is an estimator of the dye error with a small bias.

$$E\left(\widehat{\Delta}^{global}\right) = E(\bar{M}) = \Delta + \frac{\sum_{i=1}^{N}\theta_i}{N} = \Delta + \frac{\sum_{DE}\theta_i}{N}$$

Note that if the true log-ratios of the differentially expressed genes (DE-genes) are symmetrically distributed around zero, the estimator is unbiased. Under this assumption the average M-value should also be centered at zero and the above normalization method assures this. Consequently, the average of the log-signals of both samples will be equal. This normalization method is referred to as the global dye normalization. The one channel extension is called global scaling and sets the average of the log-transformed signals from all arrays to the same value.

In the general model previously described, it is not possible to estimate the dye errors for completely individual values. Therefore, some assumptions about the dye error are needed. Several normalization methods with

different assumptions of the dye errors have been proposed. The two most commonly used normalization methods are the MA-loess (Dudoit et al. 2002) and the quantile normalization (Bolstad et al. 2003).

The *MA-loess* assumes that the dye error is dependent on the intensity so that genes with similar expression have similar values. It uses the A-value as an estimate of the average intensity,

$$A_i = log_2\left(\sqrt{X_{it}X_{ir}}\right) = \frac{log_2(X_{it})+log_2(X_{ir})}{2},$$

and models the dye error as a function of the A-values. In the MA-loess normalization method the dye error is estimated by modeling the M-values as a function of the A-values through local regression, i.e. loess (Cleveland 1979),

$$\hat{\Delta}^{MA}(A_i) = f(M, A, h) = B(A_i)A_i.$$

Here, $B(A_i)$ are the regression coefficients estimated by weighted least squares in a neighborhood of A_i and h is a smoothing parameter for determining the size of the neighborhood.

The transformation from log-signals to M and A is equivalent to a 45 degree rotation of the axes. Thus, MA-loess corresponds to local orthogonal regression of the log-signals. Given the estimated dye error, normalized log-signals can be obtained through back-transformation, by

$$\log_2(X'_{it}) = \frac{2A_i+M'_i}{2} = \frac{2A_i+\left(M_i-\hat{\Delta}^{MA}(A_i)\right)}{2} = \log_2(X_{it}) - \frac{\hat{\Delta}^{MA}(A_i)}{2}$$

$$\log_2(X'_{ir}) = \frac{2A_i-M'_i}{2} = \frac{2A_i-\left(M_i-\hat{\Delta}^{MA}(A_i)\right)}{2} = \log_2(X_{ir}) + \frac{\hat{\Delta}^{MA}(A_i)}{2}.$$

Consider the following primary assumptions:

(1) Only a small fraction of the genes are differentially expressed.

(2) The true log-ratios are symmetrically distributed around zero.

Given the similarities of the MA-loess and the global method, it can be argued that if the assumptions (1) and (2) are locally true, i.e. true in all local neighborhoods, then the MA-loess method will provide an unbiased estimator of the dye error.

Yang *et al.* (2002) extended the MA-loess to account for spatial errors by estimating the dye errors within sub-grids, i.e. probes spotted (printed) by the same tip, of the array (print-tip MA-loess).

Further extensions to account for spatial effects have been proposed by Wilson *et al.* (2003). A lot of methods with essentially the same idea, by using other methods such as splines, have also been proposed. A one channel version of the MA-loess was suggested by Bolstad *et al.* (2003), which iteratively normalizes pairs of arrays until the array errors are sufficiently small.

In Paper II we proposed a model where the dye error is dependent on the intensity of each individual channel, rather than the average. We introduced the MC-normalization, which is based on the model proposed. In the MC-normalization the normalized signals are obtained through,

$$\log_2(X'_{it}) = \log_2(X_{it}) - \alpha\widehat{\Delta}^{MC}(\log_2(X_{it}))$$

$$\log_2(X'_{ir}) = \log_2(X_{ir}) + (1-\alpha)\widehat{\Delta}^{MC}(\log_2(X_{ir})),$$

where α is a parameter that is used to normalize the data towards the channel with the smallest error.

In the *quantile normalization*, the distributions of the observed signals for the genes on each array are considered. If we let the function $g(\gamma_j)$, denote the deterministic distribution of the mRNA concentrations of all genes on an array j, the distribution should be equal for all samples with the same treatment. Under assumption (1) the distribution should be approximately equal for all samples, i.e.

$$g(\gamma_j) \approx g(\gamma_i),\ \forall i,j.$$

Under the additional assumption that the array errors change the distribution of the mRNA levels in the same way for each sample, then the distribution of the expected values should be approximately equal and consequently also the distribution function of the observed signals. The objective of the quantile distribution is to normalize the data so that the distributions of each array are equal. For this purpose, the empirical cumulative distribution function is estimated for each sample and an average distribution function is estimated, i.e.

$$u_{ij} = F_{nj}(X_{ij}), \quad F_{nm}^{-1}(u_i^m) = \frac{\sum_{j=1}^{n} F_{nj}^{-1}(X_{ij})}{n}.$$

The normalized signals are then calculated from the inverse of the new distribution function,

$$X_{ij}^{quantile} = F_{nm}^{-1}(u_{ij}).$$

A two channel version of the quantile normalization can easily be constructed by performing the normalization for each paired sample, however it has been argued that the methodology is too aggressive for two channel experiments (Ballman et al. 2004).

Invariant methods

The two primary assumptions (1) and (2) in the previous section are not sufficient conditions to guarantee success of either the MA-loess or the quantile normalization. However, if both assumptions are invalid both methods will fail in normalizing the data, i.e. the dye/array errors will not be set to an equal level in the samples.

Assumptions (1) and (2) are often true in many gene expression studies. However, in several areas the assumptions are likely not valid, e.g. ChIP-chip experiments, gene expression studies of apoptosis (cell death) or SNP studies of copy number variations. In ChIp-chip experiments it can often be expected that a larger proportion of probes are different between the treatment and reference samples compared to most gene expression studied. Furthermore, in many ChIP-chip studies, all true differences can be expected to have higher amount of DNA in the treatment sample (only enriched regions). Thus, both assumptions may be invalid for ChIP-chip experiments.

A few methods have been developed to normalize data when the two assumptions are not valid. One approach is to use an *invariant method*, which identifies a subset of the data where the assumptions are valid. This subset is then used to estimate the normalization function. The normalized data are then obtained by applying the normalization function to all data. Several methods to identifying the subset have been proposed. They are often based on outlier detection, either through model assumption or through analysis of the differences in the within array ranks (Oshlack et al. 2007; Ni et al. 2008; Pelz et al. 2008; Knott et al. 2009). In Paper V, we introduced an invariant method based on Hidden Markov Models (HMM), which uses the dependency structure inherent in ChIP-chip experiments to identify an invariant subset.

Introduction to Hidden Markov Models with discrete time

Hidden Markov Models were introduced by Baum and Petrie (1966) and have several pattern recognition applications, e.g. speech recognition, handwriting recognition and gene finding.

A Hidden Markov Model is a process whose observable outcome is determined by an underlying, unobservable Markov process. Here only finite state Markov processes, i.e. Markov chains, with discrete time are considered. A Markov chain is a discrete random process $Z(t)$ which satisfies the Markov property, i.e. the preceding state of the process depends on the present state and is independent of the past (and the future),

$$P(Z(i+1) = z_{i+1}|Z(i) = z_i, Z(i-1) = z_{i-1}, \dots, Z(1) = z_1) = P(Z(i+1) = z_{i+1}|Z(i) = z_i) \ \forall i.$$

Hidden Markov Models has an underlying Markov chain with *s* states. The observable outcome Y_t is a random variable whose distribution is a function of the states. In Markov model terminology the transition matrix (T) describes the probabilities of changing between states given the present state of the Markov model,

$$T = \begin{bmatrix} p_{11} & \cdots & p_{s1} \\ \vdots & \ddots & \vdots \\ p_{1s} & \cdots & p_{ss} \end{bmatrix},$$

where $p_{ij} = P(Z(t+1) = j|Z(t) = i)$. The emission probabilities are the collection of distributions for the outcome of the process,

$$\{f_1(y), f_2(y), .., f_s(y)\}.$$

Most often the states of the Markov chain represent parameters for a common distribution of the outcomes, $f_i(y) = f(y|\theta_i)$. The initial state probabilities (π_0) describe the probabilities for which state the Markov chain is in at the first time point. Hidden Markov models are often represented by diagrams as illustrated in Figure 6.

Baum *et al.* (1970) introduced a maximum likelihood method which is referred to as a special case of the EM-algorithm (Dempster et al. 1977) to estimate the parameters of a Hidden Markov Model given the observed data. To predict the states of the Hidden Markov Model, given the (estimated) parameters the Viterbi algorithm (Viterbi 1967) can be used.

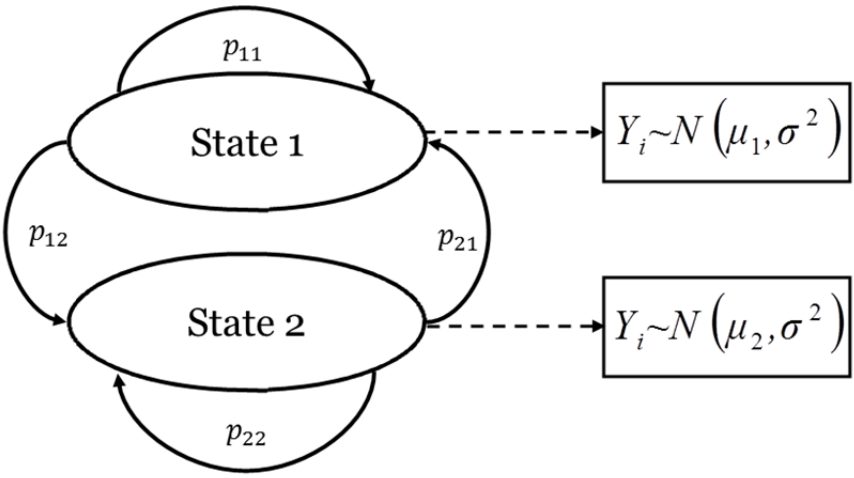

Figure 6. Diagram of a Hidden Markov Model with two states and the emission probabilities are normal distributions with the same variance and means corresponding to each state.

Downstream analysis

Note on terminology

Machine learning is a term in computer science on the subject artificial intelligence. It describes the science of learning a computer to make decisions based on pattern recognition in empirical data. Machine learning can be unsupervised or supervised, where the unsupervised learning corresponds to clustering in statistics terminology and supervised learning to classification. Although the computer science terminology is used in Paper IV, the statistics terminology will be used in this section.

Identification of DE-genes

In this section we consider gene expression microarray experiment with two treatments (treatment and reference), where we have log-transformed signals (Y) from an experiment with N genes and $2n$ samples,

$$Y_{ijt} = log_2(X_{ijt}) = C_i + \mu_{it} + \varepsilon_{ijt},\ i = 1 \dots N, j = 1 \dots n,$$

$$Y_{ijr} = log_2(X_{ijr}) = C_i + \mu_{ir} + \varepsilon_{ijt},\ i = 1 \dots N, j = 1 \dots n.$$

Here, C denotes the normalized array error, μ the true log concentration and ε the random noise. Subindex t and r denote treatment and reference samples.

The aim of the experiment is to identify the DE-genes, i.e. we want to find which of the N genes that have different (true) concentrations in the two treatments, $\{i: \mu_{it} \neq \mu_{ir}\}$. For a particular gene, a statistical test such as the T-test can be performed to test the hypothesis $\mu_{it} \neq \mu_{ir}$ against null hypothesis $\mu_{it} = \mu_{ir}$. Based on the test we can either reject or not reject the null hypothesis at a specified significance level. If the null hypothesis is rejected the gene is called differentially expressed.

If the true concentrations are known the genes can be categorized based on the truth and the result of the tests, as tabulated in Table 1. A gene can either be a true positive (a gene correctly called DE), a false positive (a gene incorrectly called DE), a true negative (a gene correctly called NDE) or a false negative (a gene incorrectly called NDE.

Table 1 cross-tabulation of the truth and the call made for a gene.

		Truth	
		Non-differentially expressed	Differentially expressed
Call	Non-differentially expressed	True negative	False negative
	Differentially expressed	False positive	True positive

The significance level controls the risk that the gene is classified as false positive. When we are testing a large number of genes, we can expect that, a percentage equal to the significance level will be false positives (mass significance). This problem has been known in statistics for a long time and several propositions have been made to correct for mass significance. One classical way is to correct the significance level so that the probability of one false positive is controlled (Family Wise Error Rate). However, due to the low amount of observations in a typical microarray experiment these are too conservative and it is unlikely that any DE-genes are identified (proportion of true positives is close to zero). Other methods to control the False Discovery Rate have been proposed. The *False Discovery Rate* (FDR) is defined as the expected proportion of false positives among the genes called differentially expressed, i.e.

$$FDR = E\left(\frac{FP}{FP+TP}\right),$$

where FP and TP are the number of false positives and true positives respectively.

In the microarray setting, the genes are often sorted according to the value of a test statistic (or p-value), in order of how likely the gene is differentially expressed. A cut off (τ) is then decided and a set of genes (Ω) with test-statistic values (or p-values) more extreme are identified. The genes in this set are the genes declared differentially expressed. Here, a larger statistic is considered more extreme,

$$T_{[1]} > T_{[2]} > \cdots > T_{[k]} > T_{[l+1]} > \cdots > T_{[N]}$$

$$\tau = T_{[k]},\ \Omega = \{i: T_i > \tau\}.$$

The cut off is often decided so that an estimate of FDR is acceptably low. A few methods have proposed to estimate the FDR. These methods can be based on on resampling techniques (Tusher et al. 2001) or adjustments to the p-values (Benjamini and Hochberg 1995). Adjusting p-values is equivalent to adjusting the significance level.

It has previously been shown that the classical T-test is not a well performing test for microarray experiments (Qin and Kerr 2004). One reason for this, is that the sample sizes in microarray experiements are typically very low (n=2-8). Due to the low sample size, the power, i.e. the probability of a true positive, of any classical tests is likely low. In order to improve the results of microarray experiments, new tests or adaptations of classical tests have been proposed. The idea behind some of these tests, is is adjust for the low amount of information in one gene, by borrowing information from other genes. In one such approach, the variation of the test statistics is reduced for individual genes, by adjusting the estimated variance. Tusher *et al.* (2001) added a small value to all variances to avoid extreme values of the T-statistic, thus reducing its variation. Baldi and Long (2001) suggested to model the variances as a function of the intensities by using loess. The individual variances are then estimated as a weighted average of the loess function and the sample variances,

$$\hat{\sigma}_i^2 = \frac{(n-1)s_i^2 + K\sigma_g^2}{n + K - 2} \quad i = 1, .., N$$

where σ_g^2 is the global variance estimated as a function of the mean intensity, s_i^2 the sample variance and K a tunable parameter to define the influence of the global variance.

While both of these methods are based on the T-statistic other statistics have also been proposed. Lönnstedt *et al.* (2002) suggested the B-statistic, which is a log-odds ratio from the Bayesian approach. The a priori distributions of

the mean and variance are assumed to be normal and inverse gamma distribution respectively. The B statistic is defined as,

$$B_i = log\frac{p}{1-p}\frac{1}{\sqrt{1+nc}}\left[\frac{a+\frac{(n-1)}{n}s_i^2+M_i^2}{a+\frac{(n-1)}{n}s_i^2+\frac{M_i^2}{1+nc}}\right]^{v+\frac{n}{2}}$$

where M_i is the mean log-ratio, s_i^2 is the sample variance for gene i, a and v are hyperparameters of the inverse gamma prior, c is a hyperparameter for the normal prior with non-zero means and p is the proportion of genes with non-zero means.

Identification of enriched regions

Hypothesis testing assumes that all variables are independent. Although this assumption may not be completely true in gene expression studies, it is definitely invalid in certain high through-put methods. In ChIP-chip experiments probes from closely related parts of the genome are dependent. Furthermore, the number of probes in ChIP-chip experiments ranges in the millions, while the number of arrays is often smaller than in gene expression studies. This would make the proportion of false positive using traditional analyzes a severe problem. However, the aim of data analyzes for ChIP-chip experiment is not to identify individual probes but rather regions where all probes are affected. Therefore, downstream analysis takes a different approach from that of gene expression studies.

A simple analysis consists of smoothing the intensities along the genome (e.g. local median of intensities from several probes) and regions above a cut off are considered affected (enriched). More elaborate analyzes, uses the dependency structure through Hidden Markov Models in order to identify the altered regions (Li et al. 2005; Munch et al. 2006; Qin et al. 2010).

Clustering

In Paper III, the effect pre-processing has on clustering of microarray data is studied in a large scale evaluation of clustering methods and pre-processing methods. Here, a short introduction to clustering its use in microarrays data analysis is provided.

Clustering is a statistical method for grouping variables based on their similarity in the observed data. In analysis of microarray data, clustering is often performed in order to identify subgroups in the samples and the genes

that separate these groups. Therefore, clustering is performed on genes and samples individually and then simultaneously visualized in a heat map, as illustrated in Figure 7. Throughout this section we consider clustering of samples on normalized log transformed signals (Y) from n samples from data with N genes.

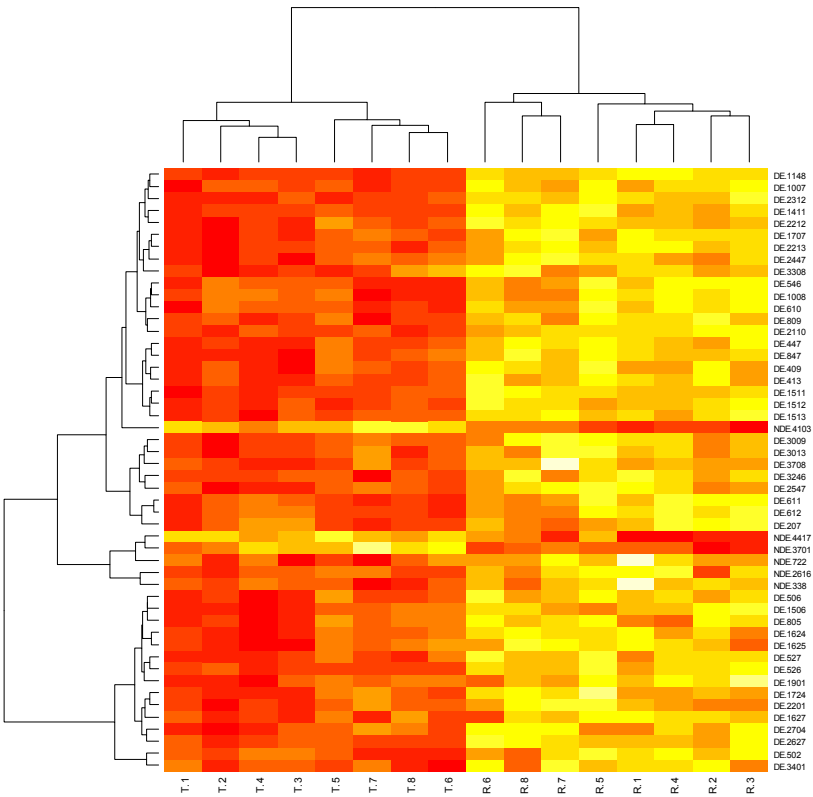

Figure 7. Heat map for the 50 genes with the highest values of test-statistic in a sample of the Lucidea data (Paper I). Genes (rows) and samples (columns) are clustered individually using the average linkage method of hierarchical clustering with a Euclidian distance matrix. The colors represent the values of the genes in each sample. It goes from low values (red) to high values (yellow). The sample names denote reference (R) or treatment (T) and an array number. The gene names denote non-differentially expressed (NDE) and differentially expressed (DE) genes and a gene number. The sample contains 5% DE-genes, all with higher concentration in the reference sample.

One of the classical clustering methods is (agglomerative) hierarchical clustering. It starts with every variable as a group. The two groups with the smallest distances between them are joined together into a new group. The joining procedure is continued until only one group remains.

In this method the (N-dimensional) pairwise distances are a measure of (dis)similarity between variables and several ways to define the distances exists, e.g. Euclidian distances,

$$d(Y_{.j}, Y_{.k}) = \sqrt{\sum_{i=1}^{N} (Y_{ij} - Y_{ik})^2},$$

or distances based on correlation,

$$d(Y_{.j}, Y_{.k}) = 1 - cor(Y_{.j}, Y_{.k}).$$

Essentially the hierarchical clustering builds a tree, where each node is a cluster. Given a specific number of clusters the tree is cut at the level where the number of nodes is equal to the number of cluster.

Other classical clustering methods, e.g. K-means, start with a given number of clusters and the (N-dimensional) centers of the clusters. Which cluster each variable belongs to are then chosen so that the distances between the variables and the center of the clusters are minimized.

Classification

In Paper IV, the effect pre-processing has on classification of microarray data was studied in a large scale evaluation of combinations of classification and pre-processing methods. Here, a short introduction to classification and its use in microarrays data analysis is provided.

Classification is in some ways similar to clustering. In clustering, the aim is to identify groups in the data, while in classification the true groups (*classes*) of the data are known and the aim is to build a model (*classifier*) from the data that can separate future data in to the classes (*classify*). In the microarray settings, classification is often performed in order to identify variables which can be used to predict different subtypes of cancer, strains of bacteria or diseases.

The data are often divided into a training set and a test set, where the training set is used to estimate the classifier and the test set is used to evaluate the classifier. This procedure is used to avoid over-fitting, i.e. when the classifier is too well adapted to the observed data but not a good model for general data. To get a better estimate of the classifiers performance, cross validation can be performed. In cross validation, the data are iteratively divided into new test and training sets of fixed sizes and the average performance of the classifier is estimated.

Consider the training set to be data in the form of $\{\boldsymbol{x}_i, y_i\}, i = 1,2,\ldots,n$ where y_i is the true class for the i:th sample and $\boldsymbol{x}_i$ is a vector of N variables. As a simplified and general description for classification, it can be seen as two functions,

$$\hat{y}_i = \theta_{H\to Y}(f_{X\to H}(\boldsymbol{x}_i, \boldsymbol{\alpha}), \boldsymbol{\beta}),$$

where the function $f_{X\to H}(\boldsymbol{x}_i, \alpha)$ maps the variables into a new space (H) that best separates the samples and the decision function $\theta_{H\to Y}(h_i, \beta)$ maps the new space into the classes (Y). Given the training set, the unknown set of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated, so that the separation into the true classes are optimized. This is often done by minimizing the error (or a function of the error), $\epsilon = g(y_i, \hat{y}_i)$, where g is a cost function. Several classification methods have been proposed, which essentially differ in the functions used. When the number of classes is two, the decision function is often a hard limiter function, e.g.

$$\theta_{H\to Y}(h_i, \beta) = \begin{cases} class\ 1 & h_i > \beta \\ class\ 2 & h_i < \beta \end{cases}.$$

A classical statistical classification method is linear discriminant analysis (LDA). The idea in LDA is to transform all variables to a single variable through a linear combination, (i.e. project the data from N-dimensions down to one). The coefficients in the linear combination are chosen to give the best separation of the true classes in the new variable.

One of the approaches, not following the general model described, is decision trees (DT). In DT a tree is built, where each node in the tree splits the data into groups based on the values of most separating variable in the group. The most separating value can be identified using different measures, e.g. information gain and Gini index (Raileanu and Stoffel 2004). An example of a decision tree with three variables follows,

$$\hat{y}_i = \begin{cases} class\ 1 & x_{i[1]} > \alpha_1 \\ class\ 2 & x_{i[1]} < \alpha_1\ and\ x_{i[2]} < \alpha_2\ and\ x_{i[3]} < \alpha_3 \\ class\ 1 & x_{i[1]} < \alpha_1\ and\ x_{i[2]} < \alpha_2\ and\ x_{i[3]} > \alpha_3 \\ class\ 1 & x_{i[1]} < \alpha_1\ and\ x_{i[2]} > \alpha_2\ and\ x_{i[3]} < \alpha_3 \\ class\ 2 & x_{i[1]} < \alpha_1\ and\ x_{i[2]} > \alpha_2\ and\ x_{i[3]} > \alpha_3 \end{cases}$$

where $x_{i[\cdot]}$ denotes the variables ordered after a separation measure. The example DT is also illustrated in Figure 8.

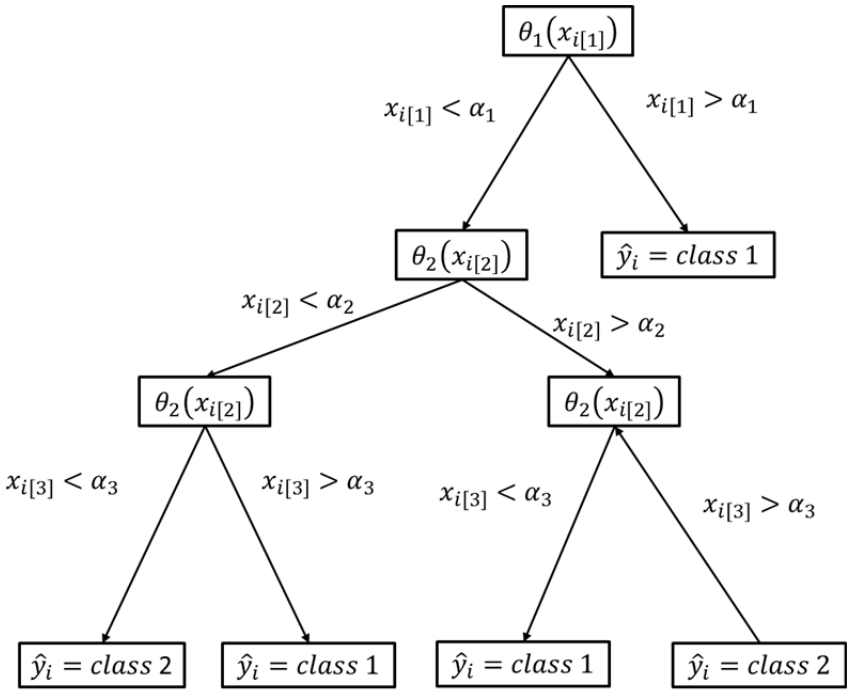

Figure 8. An example of a decision tree for two classes and three variables. $x_{[\cdot]}$ denotes the variables ordered after a separation measure.

Gene selection

As most genes are not differentially expressed and contain no information for clustering and classification, a step to reduce the data is often performed. For some of the classification methods the variables are mapped to a space of higher dimensionality than the variable space. Therefore, it may even be necessary to reduce the data in order for the analysis to be computationally acceptable. For classification problems, the gene selection is often the same as the analysis to identify differentially expressed genes. For clustering the true classes are unknown and one approach is to identify the genes with the highest variation across all samples. Another approach is to use data reduction techniques such as principal component analysis.

Evaluation of pre-processing methods

Novel pre-processing methods are commonly evaluated on simulated or real data sets. Although simulated data sets are good complements in evaluation, simulating realistic data is not a trivial task. Data from a real experiment has the advantage that, it is what the method is supposed to be applied to. The major disadvantage with real data is that the truth is not known. Therefore, a large problem with real data is the construction of reliable and relevant estimators for evaluation of analysis methods. One approach of evaluation using real data is performed by evaluating the variance, i.e. MSE of the experiments. The variance is admittedly tied to the ability to identify differentially expressed genes, however due to the small fraction of DE-genes only a small number of NDE-genes will influence the results. These NDE-genes can be considered to be outliers and will not have a great effect on the estimate of variance. Other evaluations using real data have used classification error rates in order to evaluate the performance of pre-processing methods (Wu et al. 2005).

Another possibility for evaluating analysis methods is plasmodes, i.e. real data sets whose structure is known (Mehta et al. 2004). Spike-in experiments are microarray experiment where the probes are artificially constructed, so that the true concentrations are known. These have the advantage of going through all the experimental steps, while the truth is known. The disadvantage of spike-in experiments is that they do neither reflect biological variation nor the biological properties of a sample (e.g. distribution of expression). Many spike-in experiments only have a few spike-in genes, while all other genes are real genes. A drawback of this is that the spike-in genes do not nescessarily share the same properties in hybridization as real genes.

In evaluating microarray data we argue that two properties are of importance, the ability to identify differentially expressed genes and the ability to provide accurate estimates of the true amplitude of differential expression (*fold change*). By using simulated data or data from spike-in experiments these properties can be estimated and used in evaluation of pre-processing methods.

Considering the categorization described in Table 1, the proportion of DE-genes correctly called differentially expressed is a measurement of the ability to identify differentially expressed genes. This measurement is often referred to as *true positive rate* (TPR) or *sensitivity*,

$$sensitivity\ = TPR = \frac{TP}{N_{DE}} = \frac{TP}{N*p_{DE}}.$$

Here, p_{DE} denotes the proportion of DE-genes. The sensitivity can be seen as a function of the cut off used in the statistical analysis. One approach of determining the cut off is to estimate and control the false discovery rate. However, the false discovery rate is a function of the expected sensitivity. Therefore, we use the false positive rate to determine the cut off. The *false positive rate* is the proportion of NDE-genes incorrectly called differentially expressed, i.e.

$$FPR = \frac{FP}{N_{NDE}} = \frac{FP}{N*p_{NDE}}.$$

Here, p_{NDE} denotes the proportion of NDE-genes. The FPR is directly tied to the *specificity*,

$$specificity = 1 - FPR = \frac{FN}{N_{NDE}} = \frac{FN}{N*p_{NDE}}.$$

The sensitivity is often visualized as a function of the FPR in a *Receiver Operator Characteristics-curve* (ROC-curves), as illustrated in Figure 9.

Considering the false discovery rate, this can be written as a function of the expected TPR and the FPR,

$$FDR = E\left(\frac{FP}{FP+TP}\right) = E\left(\frac{FPR*p_{NDE}}{FPR*p_{NDE}+TPR*p_{DE}}\right).$$

Assuming that the expected sensitivity, i.e. power of the tests, of the experiment is 80% and 10% is an acceptable level of the FDR. Then the expected FPR for experiments with 1% and 5% differentially expressed gene is approximately 0.5 and 0.1 percent respectively. Therefore we argue that the sensitivity should be evaluated at a false positive rate much less than 5%. Note that, with 1% DE-genes and a total of 40 000 genes, the NDE-genes which are decisive for the sensitivity, is the 36 NDE-genes with the most extreme values of the test-statistics (or p-values).

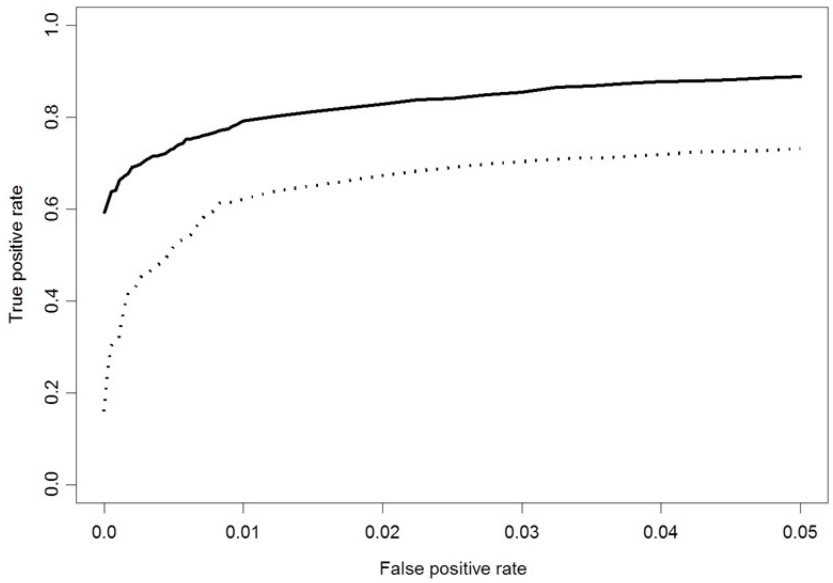

Figure 9. A receiver operator characteristics curve for normalizations with background correction (dotted) and without background correction (solid).

Arguable the ability to identify differentially expressed genes is the most important property of a microarray experiment. For further downstream analysis such as clustering, classification and meta analyzes the accuracy of the estimated fold change may be of importance.

Consider a model where the the array error is a function of the concentration. After normalization the array error is equal in both samples,

$$Y_{ijt} = log_2(X_{ijt}) = C_i(\mu_{it}) + \mu_{it} + \varepsilon_{ijt}, \qquad i = 1 \dots N, j = 1 \dots n$$

$$Y_{ijr} = log_2(X_{ijr}) = C_i(\mu_{ir}) + \mu_{ir} + \varepsilon_{ijt}, \qquad i = 1 \dots N, j = 1 \dots n.$$

The log-ratio then becomes,

$$M_{ij} = C_i(\mu_{it}) - C_i(\mu_{ir}) + \mu_{it} - \mu_{ir} + \varepsilon_{ijm}, \qquad i = 1 \dots N, j = 1 \dots n.$$

$$E(M_{ij}) = \text{bias}_i + \theta_i$$

For NDE-genes the expected value of the difference between the array errors is zero. For DE-genes difference depends on the remaining array error. Thus the estimated fold change of DE-genes will be biased.

If the true log-ratio is known we can calculate the the bias. The bias is thus a measure the (in)ability to provide accurate estimates of the fold change. In order to estimate the average bias over several genes with different true log-ratio, the reflected bias was constructed,

$$reflected\ bias\ = \frac{\sum_{i=1}^{N_{DE}} sign(\theta_i)(\bar{M}_i - \theta_i)}{N_{DE}}.$$

The reflected bias is based on the assumption that the log-ratios are distributed around zero and constructed so that negative and positive bias always mean underestimation and overestimation of the true log-ratio, respectively.

How pre-processing affects downstream analysis

Since the introduction of the microarray technology a score of methods have been developed for pre-processing and downstream analysis. A large part of the work presented in this thesis is evaluation of pre-processing methods with respect to downstream analysis. Here a brief summary of some of the findings is presented as a concluding section.

It has been known that microarrays generally underestimate the true log-ratios of the differentially expressed genes. In evaluating the effects of the pre-processing, an often recurring trend is a tradeoff between the bias and the sensitivity. The clearest and most noted tradeoff is in the effect of background correction. Applying background corrections yield a significant decrease in the bias, but also a decrease in the sensitivity. One of the results in Paper I shows that censoring, as a complement to background correction, can reduce this tradeoff. However, how to properly choose the censoring value is unknown. In Paper II we introduced a novel dye normalization method which is designed to minimize the bias, under certain model assumptions. The evaluation of the method shows that the bias in microarray experiments can be reduced without a loss of sensitivity.

One interesting result of Paper I is that the bias is dependent on the underlying mRNA concentration and the size of the true log-ratio. This suggested one of our working hypotheses, that bias may have a deteriorating

effect on further downstream analysis such as clustering and classification. The results in Paper III and Paper IV show a rather complex picture. Although normalization generally has a positive effect, there are no general differences between normalization methods. Whether a normalization method have a large bias (i.e. no background correction) or a small bias (background correction) has no general influence. In these studies several methods performed better on data without background correction, others on background corrected data and some equally well on both. Although this suggests that some methods are more sensitive to bias, the best normalization for a particular analysis may not be the one with the smaller bias.

One of the problems in microarrays is the amount of variation or differences between experiments, as seen in Paper III and IV. The variation between data sets has a large impact on the performance of downstream analysis. Consequently, the best normalization methods are dependent on the data and its intended use. This is likely a reason why no standardized procedure for pre-processing and data analysis has been accepted.

The effects of saturation correction are not always clear, often very few of the differentially expressed genes are saturated and little information can be gained by the correction. However, in Paper I, we note that the constrained model, which combines information from several scans, is one of the best performing methods with respect to sensitivity.

In microarray analysis dye or array normalization is considered a necessary step. Although this was not always clear with respect to clustering and classification, it is definitely true for identification of altered variable, e.g. differentially expressed genes. The performances of the different dye or array normalization methods such as quantile, MA-loess and MC-loess are often quite similar with respect to the sensitivity. However, in Paper II we note that methods accounting for spatial errors had significantly higher sensitivity than global methods and in Paper V it was clear that the quantile normalization was a better performing method on one channel experiments than the one channel version of MA-loess.

In Paper V, we also show that when assumptions (1) and (2) are not valid the methods will fail to normalize the data. Invariant methods are less sensitive but can have a negative effect when the assumptions are valid.

Summary of papers

Paper I

In this paper we evaluate known pre-processing methods and different combinations of background correction, filtration, censoring and dye normalization for two channel experiments. The evaluations are performed using an in-house produced spike-in experiment (Lucidea experiment). The pre-processing methods are evaluated by measuring two properties, the ability to identify differentially expressed genes (sensitivity at fixed specificity) and the ability to provide accurate estimates of the true log-ratios (bias).

We show that there is often a tradeoff between bias and sensitivity. Furthermore, we show that the bias depends on the underlying mRNA concentration, and DE-genes with low concentrations have larger bias. We introduced a novel filtration method, called partial filtration, which along with the constrained model (Bengtsson et al. 2004) is one of the best performing methods with respect to sensitivity.

We show, in accordance to previous published results, that analyzes using background correction have considerable lower bias and sensitivity. Furthermore, we show that censoring with an optimal value results in high sensitivity and relatively low bias, although choosing the optimal value is far from trivial.

Paper II

We introduce a novel dye-normalization method, called MC-normalization. The method is evaluated along with the MA-loess normalization. The methods are evaluated using the Lucidea experiment and a publicly available spike-in data. The spike-in data is from Agilent type of arrays and is a part of the MAQC project (Consortium et al. 2006). The evaluation is based on the sensitivity and the bias.

We propose a model, where the dye error is dependent on the intensity of each individual channel. The IC-curve is introduced as a visualization of the dye errors from each channel for spike-in experiments. A general method for identifying the channels with the lowest bias is also proposed. The MC-normalization is a method designed to normalize the data towards the channel with the lowest error.

We give theoretical proof, that under the model proposed, the MA-normalization introduce a bias among the DE-genes and the MC-normalization has lower bias. The evaluation corroborates the theoretical proof and shows that the methods have similar sensitivities. The evaluation also shows that methods applied to sub-grids, i.e. print-tip normalizations, have considerably higher sensitivity than those applied to the entire arrays, i.e. global methods.

Paper III

In this paper, publicly available two-channel data sets, where the true classes are known, is used to evaluate the individual performances of normalization, gene selection and cluster analysis methods and to discover possible synergistic effects in combinations thereof. In total 2780 analyzes are performed, each analysis consists of a unique combination of normalization method, missing values imputation, standardization, gene selection and cluster analysis method.

We show that the choice of analysis has a significant impact on the performance. However, there is a large variation in performance between the data sets and no general recommendation can be made. Among the methods, gene selection followed by cluster analysis method, have the largest impact on the performance. In general, normalization has a positive effect on the performance, but we are unable to determine any relative difference between normalizations methods.

Paper IV

Here, we provide a large scale evaluation of combinations of classification, gene selection and normalization methods in order to identify individually well performing methods and synergies between method choices. The data sets used in the evaluation are the same data sets used in paper III.

The evaluation show that support vector machines with radial, linear and polynomial basis perform consistently well across all data set and is the best performing classification method, which has also been shown in previous studies of classification methods. Synergistic effects are identified between the best performing methods, gene selection based on the T-test and a relatively high number of genes selected. Similarly to Paper III there is a large variation in performance between the data set and the general trend is that normalization has a positive effect on classification, although no general trend between the normalization methods can be seen.

We also show that the performance of the best individual clustering methods can be severely reduced by unfortunate choices of the other methods, including normalization.

Paper V

In pre-processing of microarray experiments two assumptions are often made, (1) only a small fraction of the genes are differentially expressed and (2) the true log-ratios of the DE-genes are symmetrically distributed around zero. In studies of gene expression related to apoptosis, protein binding studies and a number of other studies, these assumptions are likely not valid. The performances of standard and invariant normalization methods are evaluated using subsets of the Lucidea experiment. The subsets are constructed so that the two assumptions are violated to different degrees. We show that violation of the two assumptions have a deteriorating effect on the performance of the normalization methods. We also show that invariant methods based on ranks have a better performance on such data, but a worse performance when assumptions are valid.

ChIP-chip experiments where the two assumptions are known to be incorrect and have a dependency structure between probes are considered. We introduce a novel invariant normalization based on HMMs.

Although it can be known from the type of experiment or experimental design whether the experiment is skew (i.e. where the second assumption is invalid), it is not always the case. Therefore we introduce the DSE-test (Detection of Skewed Experiments) for identifying experiments where the assumptions are likely to be invalid.

Both the DSE-test and the novel normalization were evaluated on simulated and data from three ChIP-Chip experiments. We show that the DSE-test is able to identify skew experiments and have high power. The evaluation also show, that the HMM-normalization have a higher sensitivity at similar specificity than the invariant method based on ranks and the normalization based on all data.

Acknowledgements

First I would like to thank my supervisor *Patrik Rydén* for his support, guidance and masterful skill of finding interesting projects. This has been a much appreciated experience for me and I thank you for giving me the opportunity.

I would like to thank my co-supervisor and co-author *Anders Sjöstedt* who was instrumental in giving me this opportunity. I would also like to thank my former and current supervisors and co-supervisors, *Lennart Bondesson*, *Sara de Luna* and *Leif Nilsson*. Thank you for all the support you have given me.

During the work on my thesis I had the pleasure of working with a lot of very talented people. My first co-workers and the co-authors of the first paper, who introduced me to the field and the biology behind the experiments, *Henrik Andersson, Blanka Andersson* and *Laila Noppa*. I immensely enjoyed working with you. Thank you for everything.

Eva Freyhult, *Jenny Önskog* and *Torgeir Hvidsten*. Thank you for all the support and the time we spent together during our work on the third and fourth paper. I hope you do well in all your future endeavors.

I would like to thank *Philge Philip* for all the interesting discussions during our work together on the fifth paper. And thanks to the second co-author, *Per Stenberg,* a positive and excellent scientist, who can make his work sound so interesting that you begin to question your own choice of career. Thank you for your invaluable efforts to further my knowledge of the underlying biological processes.

Jessica Fahlén, teacher, student, co-worker and co-author. You are incredibly talented and the time we spent working on the second paper and the book chapter was one of the best times during my PhD-studies. Thank you.

I would also like to thank two persons that have been there to support me when I needed it the most.

Linda Junfors, you are one of the strongest and kindest persons I know. A simple thank you is not enough to truly express my gratitude for everything you have done. Thank you for being a shoulder I could lean on when I needed it. And thank you for all the support and help you gave me during the time we worked together.

Marie Honn, you are one of the kindest and most wonderful persons I know. Thank you for being a pillar of support when I needed it, it has meant more to me than I could say. Thank you for making my world a better place.

I would also like to take the opportunity to thank some of the people who made the time I spent on this thesis a much more pleasant time.

Helena Lindgren, I have always looked up to you, thank you for all the support, discussions and for being a most excellent scientist. I will miss working with you.

Marie Lindgren, thank you for being there to answer all my questions, for all the discussions and for making life more interesting.

Malin Vonkavaara, thank you your support, it has been fun working with you, I will miss it.

I would also like to thank *Lina Schelin*, for her support in all the teaching related parts of my studies.

Looking back over the years I realized just how many people that have contributed to making a hospitable workplace and I would like to thank:

The current the members of Patrik's group, *Rafael Björk* and *Therese Gabrielsson-Kellgren*.

All my 'biological friends' at the department of clinical bacteriology both present and past.

Past and present co-workers at the department of mathematics and mathematical statistics.

Past and present members of the computational life science cluster.

All the people I have had the pleasure of working with in the various projects not included in the thesis.

Last but not least I would like to thank my friends and family both distant and close. I would especially like to thank:

My mother *Katrin Landfors* for her continuous support, for taking an interest in my work and for helping me with everything.

My brother *Mikael Häggblad* and my sister *Maria Larsson* for their support, understanding and for always being there when I need it. My sister's husband Bert-Olov Larsson, who has always been there to help me with all the little things outside of work.

Slutligen vill jag även tacka mina underbara systersöner *Hampus Larsson* och *Pontus Larsson*.

References

Baldi, P. and A. D. Long (2001). "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes." Bioinformatics **17**(6): 509-519.

Ballman, K. V., D. E. Grill, A. L. Oberg and T. M. Therneau (2004). "Faster cyclic loess: normalizing RNA arrays via linear models." Bioinformatics **20**(16): 2778-2786.

Baum, L. E. and T. Petrie (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." Annals of Mathematical Statistics **37**(6): 1554-&.

Baum, L. E., T. Petrie, G. Soules and N. Weiss (1970). "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." The Annals of Mathematical Statistics **41**(1): 164-171.

Bengtsson, H., G. Jonsson and J. Vallon-Christersson (2004). "Calibration and assessment of channel-specific biases in microarray data with extended dynamical range." BMC Bioinformatics **5**(1): 177.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Cleveland, W. S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association **74**(368): 829-836.

Consortium, M., L. Shi, et al. (2006). "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nature Biotechnology **24**(9): 1151-1161.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B (Methodological) **39**(1): 1-38.

Dudley, A. M., J. Aach, M. A. Steffen and G. M. Church (2002). "Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range." Proc Natl Acad Sci U S A **99**(11): 7554-7559.

Dudoit, S., Y. H. Yang, M. J. Callow and T. P. Speed (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." Statistica Sinica **12**(1): 111-139.

Knott, S. R., C. J. Viggiani, O. M. Aparicio and S. Tavare (2009). "Strategies for analyzing highly enriched IP-chip datasets." BMC Bioinformatics **10**: 305.

Li, W., C. A. Meyer and X. S. Liu (2005). "A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences." Bioinformatics **21 Suppl 1**: i274-282.

Lyng, H., A. Badiee, D. H. Svendsrud, E. Hovig, O. Myklebost and T. Stokke (2004). "Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction." BMC Genomics **5**(1): 10.

Lönnstedt, I. and T. P. Speed (2002). "Replicated microarray data." Statistica Sinica **12**(1): 31-46.

Mehta, T., M. Tanik and D. B. Allison (2004). "Towards sound epistemological foundations of statistical methods for high-dimensional biology." Nat Genet **36**(9): 943-947.

Munch, K., P. P. Gardner, P. Arctander and A. Krogh (2006). "A hidden Markov model approach for determining expression from genomic tiling micro arrays." BMC Bioinformatics **7**: 239.

Ni, T. T., W. J. Lemon, Y. Shyr and T. P. Zhong (2008). "Use of normalization methods for analysis of microarrays containing a high degree of gene effects." BMC Bioinformatics **9**: 505.

Oshlack, A., D. Emslie, L. M. Corcoran and G. K. Smyth (2007). "Normalization of boutique two-color microarrays with a high

proportion of differentially expressed probes." Genome Biol **8**(1): R2.

Pelz, C. R., M. Kulesz-Martin, G. Bagby and R. C. Sears (2008). "Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data." BMC Bioinformatics **9**(1): 520.

Qin, L. X. and K. F. Kerr (2004). "Empirical evaluation of data transformations and ranking statistics for microarray analysis." Nucleic Acids Res **32**(18): 5471-5479.

Qin, Z. S., J. Yu, J. Shen, C. A. Maher, M. Hu, S. Kalyana-Sundaram, J. Yu and A. M. Chinnaiyan (2010). "HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data." BMC Bioinformatics **11**: 369.

Raileanu, L. and K. Stoffel (2004). "Theoretical Comparison between the Gini Index and Information Gain Criteria." Annals of Mathematics and Artificial Intelligence **41**(1): 77-93.

Ritchie, M. E., J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway and G. K. Smyth (2007). "A comparison of background correction methods for two-colour microarrays." Bioinformatics **23**(20): 2700-2707.

Rocke, D. M. and B. Durbin (2001). "A model for measurement error for gene expression arrays." J Comput Biol **8**(6): 557-569.

Scharpf, R. B., C. A. Iacobuzio-Donahue, J. B. Sneddon and G. Parmigiani (2007). "When should one subtract background fluorescence in 2-color microarrays?" Biostatistics **8**(4): 695-707.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Tusher, V. G., R. Tibshirani and G. Chu (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-5121.

Wilson, D. L., M. J. Buckley, C. A. Helliwell and I. W. Wilson (2003). "New normalization methods for cDNA microarray data." Bioinformatics **19**(11): 1325-1332.

Viterbi, A. J. (1967). "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." Ieee Transactions on Information Theory **13**(2): 260-+.

Wu, W., E. P. Xing, C. Myers, I. S. Mian and M. J. Bissell (2005). "Evaluation of normalization methods for cDNA microarray data by k-NN classification." BMC Bioinformatics **6**: 191.

Yang, Y. H., M. J. Buckley, S. Dudoit and T. P. Speed (2002). "Comparison of methods for image analysis on cDNA microarray data." Journal of computational and graphical statistics **11**(1): 108-136.

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**(4): e15.

Yin, W., T. Chen, S. X. Zhou and A. Chakraborty (2005). "Background correction for cDNA microarray images using the TV+L1 model." Bioinformatics **21**(10): 2410-2416.

Zhu, Q., J. Miecznikowski and M. Halfon (2010). "Preferred analysis methods for Affymetrix GeneChips. II. An expanded, balanced, wholly-defined spike-in dataset." BMC Bioinformatics **11**(1): 285.