Hybridization biases of microarray expression data

A model-based analysis of RNA quality and sequence effects

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALUM
(Dr. rer. nat)

im Fachgebiet

Informatik

vorgelegt

von Diplom Bioinformatiker Mario Fasold

geboren am 9. Juli 1981 in Dresden

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Peter F. Stadler, Universität Leipzig
2. Dr. Andrew Harrison, University of Essex, United Kingdom

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 11. Juni 2013 mit dem Gesamtprädikat magna cum laude.

Abstract

Modern high-throughput technologies like DNA microarrays are powerful tools that are widely used in biomedical research. They target a variety of genomics applications ranging from gene expression profiling over DNA genotyping to gene regulation studies. However, the recent discovery of false positives among prominent research findings indicates a lack of awareness or understanding of the non-biological factors negatively affecting the accuracy of data produced using these technologies. The aim of this thesis is to study the origins, effects and potential correction methods for selected methodical biases in microarray data.

The two-species Langmuir model serves as the basal physicochemical model of microarray hybridization describing the fluorescence signal response of oligonucleotide probes. The so-called hook method allows to estimate essential model parameters and to compute summary parameters characterizing a particular microarray sample. We show that this method can be applied successfully to various types of microarrays which share the same basic mechanism of multiplexed nucleic acid hybridization.

Using appropriate modifications of the model we study RNA quality and sequence effects using publicly available data from Affymetrix GeneChip expression arrays. Varying amounts of hybridized RNA result in systematic changes of raw intensity signals and appropriate indicator variables computed from these. Varying RNA quality strongly affects intensity signals of probes which are located at the 3' end of transcripts. We develop new methods that help assessing the RNA quality of a particular microarray sample. A new metric for determining RNA quality, the degradation index, is proposed which improves previous RNA quality metrics. Furthermore, we present a method for the correction of the 3' intensity bias. These functionalities have been implemented in the freely available program package *AffyRNADegradation*.

We show that microarray probe signals are affected by sequence effects which are studied systematically using positional-dependent nearest-neighbor models. Analysis of the resulting sensitivity profiles reveals that specific sequence patterns such as runs of guanines at the solution end of the probes have a strong impact on the probe signals. The sequence effects differ for different chip- and target-types, probe types and hybridization modes. Theoretical and practical solutions for the correction of the introduced sequence bias are provided.

Assessment of RNA quality and sequence biases in a representative ensemble of over 8000 available microarray samples reveals that RNA quality issues are prevalent: about 10% of

the samples have critically low RNA quality. Sequence effects exhibit considerable variation within the investigated samples but have limited impact on the most common patterns in the expression space. Variations in RNA quality and quantity in contrast have a significant impact on the obtained expression measurements.

These hybridization biases should be considered and controlled in every microarray experiment to ensure reliable results. Application of rigorous quality control and signal correction methods is strongly advised to avoid erroneous findings. Also, incremental refinement of physicochemical models is a promising way to improve signal calibration paralleled with the opportunity to better understand the fundamental processes in microarray hybridization.

Acknowledgments

First of all, I would like to thank my supervisor Hans Binder for giving me the opportunity to work on this interesting research topic. For each of the many challenging theoretical and practical problems I faced in the past years, I could always be sure he would help with his good advice. Without his continuous and great support much this work would not be possible. I also thank my second supervisor Peter F. Stadler for his great support and always useful scientific advice.

I will always remember Jan Bruecker who died much too early in 2011. He was a cheerful and inspiring character, and he contributed to this thesis by developing parts of the software used and by providing many helpful discussions.

Of course I thank all dear colleagues and friends from the IZBI and the Chair of Bioinformatics of the University of Leipzig. I really enjoyed the vivid discussions, the fun events and the good advice and deep knowledge one could always rely on. I particularly thank Corinna and Petra for their always positive attitude and their indispensable support for all administrative activities. Jens provided great technical support and advice, as did David, Henry and Christian with their helpful discussions about R and other technical obstacles. Additionally, I thank the following people who in some way contributed to the success of this work: Anne, Axel, Berni, Christian, Dom, Edith, Gero, Gunnar, Katrin, Lydia, Konstantin, Jana, Markus, Maribel, Sven, Joerg, Stephan, Stephie, Steve, Volkan, Wolfgang and everybody else I've missed. It was really a pleasure to work with you.

Finally, I thank my family and friends who indirectly contributed to this work with their continuous love and encouragement. I particularly thank my beloved Anne, as well as my parents for their support. I express my deep appreciation to all of you.

Related publications

This thesis is partially based on the following publications:

Binder H, Fasold M, Glomb T: **Mismatch and G-stack modulated probe signals on SNP microarrays.** *PloS One* 2009, **4**.

Fasold M, Stadler PF, Binder H: **G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration.** *BMC Bioinformatics* 2010, **11**:207.

Fasold M, Binder H: **Estimating RNA-quality using GeneChip microarrays.** *BMC Genomics* 2012, **13**:186.

Fasold M, Binder H: **AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data**. *Bioinformatics* 2013, **29**:129-131.

Fasold M, Binder H: **Prevalence and impact of technical artifacts in microarray expression data.** *In preparation.*

Contents

1 Introduction

1.1 The role of high-throughput technologies in modern life sciences

When a researcher in the field of molecular biology carried out an experiment in the early 1990s he would need experience, craftsmanship and a lot of time. Assume the researcher was interested in gene expression. For example, he would like to know whether a gene that potentially causes cancer is active in some tumor cells or not. He could employ a technique called Northern blot and follow a long protocol of manual steps involving, amongst other things, production of an agarose gel, RNA separation using gel electrophoresis, transfer of RNAs to a membrane and production of labeled probes. Including proper controls the whole procedure would usually take days up to weeks to complete successfully. At the end, he would know whether his gene of interest is expressed in a single cell line of a single species.

If the same researcher was interested in the same question only 10 years later in the early 2000s, the experiment would run markedly different. He could resort to several commercially fabricated instruments and automated techniques specifically designed to aid in his experiment. For example, he could employ a sensitive scanner device that uses lasers to read signals out of miniaturized DNA microarrays. He would be able to simply order some of the pre-manufactured microarrays that contain probes designed to measure the expression of his gene of interest and many other genes at the same time. And he would be able to buy tailor-made reagents that help him preparing his sample for the assay in a few simple steps. The procedure would take only hours instead of weeks.

It is easy to see why high-throughput technologies like microarrays quickly replaced previous techniques in labs all over the world. They revolutionized the way how researchers could approach the problems they were facing in their particular domain. It allowed them conducting experiments hypothesis-free: The researcher could not only study the expression of one single gene he chose because he hypothesized that it relates to the cancer, but he could instead screen thousands of genes for their expression status in the tumor cells. Also it allowed conducting experiments that could not be done before because of time or money restrictions of the previous techniques. Edward Southern, one of the inventors and early adopters of these automated techniques, later commented on this dramatic development: “Genomics, in its early days, used a range of techniques that were developed to explore the composition and sequence organization of the nuclear DNA. High-throughput methods changed that, and most research in genomics is now done in factory-like laboratories, with robots doing much of the work.” [1]

Today, many areas of life sciences rely on the methodological advances provided by a large toolbox of available high-throughput technologies. Gene expression profiling using microarrays is such a tool - being one of the first and more popular ones it is probably the best known representative for the whole toolbox. These assays are now performed routinely and in large-scale for testing the reaction of cells on different treatments and condition changes. Consider the following numbers: More than 25.000 peer-reviewed papers have been published using microarray technology from a single vendor (Affymetrix) alone [2]. Each of these publications refers to one or more experiments. For some experiments the generated data is made publicly available. Over 28.000 datasets comprising over 850.000 microarray samples have been stored in two public data repositories in the last 5 years alone[1]. The data of many more experiments is not shared in public databases, but is kept secretly, particularly for experiments performed at companies and private institutions.

Every experiment performed using high-throughput technologies has the property of producing large amounts of data that must afterwards be analyzed and interpreted. The analysis of such complex data is no simple task, even for experienced researchers. Without a deep understanding of the limitations of the technology and knowledge about proper statistical analysis it can easily be misinterpreted. Daniel MacArthur notes that "all high-throughput genomic technologies come with error modes and systematic biases that, to the unwary eye, can seem like interesting biology. As a result, researchers who are inexperienced with a technology — and some who should know better — can jump to the wrong conclusion" [3]. The combination of difficult-to-analyze data and the hope of surprising results can lead to so-called 'false positives', erroneous research findings that later had to be revoked after other groups have pointed out flaws in the analysis done by the original authors.

One example about how critical it is to ensure accurateness and rigorousness in high-throughput data analysis is given by a study published in 2007 by Spielman *et al.* [4]. It was previously known that the genetic divergence, the differences in the genetic code, between individuals of our species is drastically small. The human to human nucleotide divergence for example was estimated to be around 0.1% [5]. The study of Spielman *et al.*, which was published in Nature Genetics, sought to find the factors that contributed to the large phenotypic differences between human populations. Their approach was to focus on the variation of gene expression, patterns of genetic activity, rather than on the variation of DNA sequence. Microarray technology was to be used to obtain profiles of genetic activity in lymphoblastoid cell lines from individuals belonging to one of three population groups. The authors found that the expression of about 25% of the tested genes differs significantly

[1] Queried on the ArrayExpress website http://www.ebi.ac.uk/arrayexpress/ on January 7, 2013.

between European and Asian populations. These numbers suggested that phenotypic variability to a large part is reflected in expression variability which constituted an important finding.

However, later concerns about the accuracy of these numbers were raised [6]. Akey *et al.* reanalyzed the data of Spielman *et al.* and found 78% of the genes – a rather unrealistic number - to be significantly differentially expressed. After closer inspection of the microarray data, they found that samples have been processed in groups spanning a time of more than 3 years and that European and Asian samples had been mostly processed at different times. Akey *et al.* then found that 79% of the genes are differentially expressed between processing years but within the same population. This significant variation between the processing groups cannot be explained by biology. They concluded that the data possesses a systematic and confounding technical bias, and that the reliability of the obtained results is therefore at least questionable.

The publication of these spurious results of Spielman *et al.* in one of the most trusted scientific journals illustrates how difficult it can be to control the quality of high-throughput data and to implement its analysis. Errors and biases can be introduced in many steps and at different levels in the course of such an experiment. Differences in sample storage and treatment, reagent composition, lab worker experience, device or program variants and many other factors can lead to different results. These are methodological issues, relating to technical effects of the employed tools. Note that measurement errors are a critical but common element in scientific research methodology since its earliest days. However, for the recent high-throughput technologies the number of ‘error modes’ is drastically higher, and their impact on the complex data multifaceted and therefore hard to detect.

In summary, the powerful high-throughput technologies enjoy a high popularity in research applications, yet there are issues with the accuracy of data generation and analysis. Many factors aside the biological variable of interest influence the measured quantities. Given the critical impact of these technical effects as illustrated for the case of Spielman *et al.* it is imperative to thoroughly study them to better understand their origins and ideally to provide solutions for controlling them. Doing so for the important classes of RNA quantity, RNA quality and sequence effects in the context of common high-density microarray technologies is the main aim of this thesis.

1.2 Physicochemical models for microarray data analysis

An essential task in high-throughput data analysis is the obtainment of accurate estimates of the input quantity (e.g. transcript abundances) from the measurement output (e.g. intensity signals) which is affected by various technical disturbances. This *calibration* step requires a model describing the relationship between both quantities which is subject to the entirety of processes in the experimental system. Note that this modeling of technical processes is complementary to the modeling of the input quantities in their complex biological systems as for example in gene regulatory network models.

Most calibration methods for data originating from high-throughput technologies rely on statistical approaches. A prominent example is the MAS5 algorithm included in the manufacturer software that ships with each Affymetrix microarray device. As the default solution for computing gene expression estimates for various array types it is widely used. This simple method applies a bi-weight estimator to compute a robust mean of the probe signals interrogating one, mostly gene-related transcript [7].

The benefit of such relatively simple approaches is that no prior knowledge of the exact experimental processes is required. The processes involved in a typical microarray measurement, for example, are complex: The hybridization is highly multiplexed with thousands of competing reactions occurring in parallel. The devices are imperfect with manufacturing errors which are hard to detect, for example the probes may vary in length (‘polydispersity’) and sequence. There are a large number of biases and errors that can be introduced during the multi-step assay for sample preparation. Purely statistical approaches here provide a straightforward solution for obtaining fast and effective signal calibrations.

On the other hand, the simplicity of those methods comes with the cost of decreasing accuracy in the obtained results. While it is obviously not feasible to consider all relevant factors, it is possible to incorporate existing knowledge about important processes involved in the measurement. There are accepted physicochemical models that well describe binding of molecules on surfaces as well as the hybridization of nucleic acids, and either of these processes is central in microarray hybridizations. We and a number of peers believe that building upon basal models based on these fundamental physical principles and their incremental refinement will eventually lead to a better high-throughput data analysis. Improving on these models will increase our understanding of these complex technologies and, at the same time, increase our ability to control the data.

1.3 Objectives and outline

The objective of this thesis is to rigorously assess the specifics of microarray technology using Affymetrix GeneChip microarrays as an example. We aim to establish a deeper understanding of the limitations of current technology and to investigate how to make the most of available and future microarray data within these limitations. Particularly we intend to

- objectively assess the quality of microarray experiments (quality-control) and detect and possibly correct for confounding factors affecting the reliability of the obtained results
- evaluate and improve the precision and accuracy of microarray gene expression estimates under varying experimental conditions
- improve the understanding of the basal mechanism of surface hybridization by employing physicochemical models of duplex formation

Particularly critical methodical issues relate to variations in quality and quantity of the RNA used for hybridization as well as to variations in sequence-dependent binding due to changing experimental conditions. These effects lead to systematic changes in the microarray data which are however unrelated to the biological changes under study. Using appropriate experimental designs and newly developed methods we are able to study these technical variations and to investigate the physicochemical principles of the processes involved in microarray measurements.

We here focus on the widely adopted Affymetrix GeneChip type of microarrays. The challenges and limitations are however similar for a wide range of other chip types and to a certain degree also for other technologies that exploit the mechanisms of nucleic acid hybridization in general.

This thesis will be laid out as follows. Chapter 2 will describe microarray technology for gene expression analysis, genotyping and other applications. Chapter 3 will lay the foundations for modeling of microarray signals using physicochemical principles of competitive surface hybridization. We will describe the Hook method and its use for the robust estimation of essential model parameters. In Chapter 4 we investigate whether this methodology can also be applied to other microarray technologies besides Affymetrix GeneChip expression arrays. Chapter 5 focuses on RNA quality as a technical bias in microarray experiments and how it can be determined and corrected within the resulting data. Chapter 6 deals with sequence effects largely referring to changes in the observed probe signals due to molecular interactions of complementary nucleotide strands. We will investigate which models are both adequate and practical for modeling the signal

contribution due to sequence variation. Chapter 7 addresses the important and more general question of the impact and prevalence of technical bias in gene expression experiments. We will use the methodology developed in the previous chapters to study the effect of known sources of batch effects in a meta-study comprising thousands of microarray samples. The final Chapter 8 will discuss and conclude the results of this thesis.

2 Microarray technology

2.1 Microarrays assembly and assay

Microarrays are a powerful technology for the targeted analysis of thousands of DNA or RNA molecules in parallel. The basic principle is the hybridization of a mixture of unknown, but marked, nucleotide strands to a set of known nucleotide strands called *probes*. During the reversible chemical reaction of hybridization, complementary nucleotide strands build up a duplex structure. Quantification of bound nucleotide strands allows then to infer the contents of the mixture.

Microarrays are today available in a wide variety in terms of available instruments and assays, as well as its applications. Possible applications of microarrays include, but are not limited to, gene expression analysis, DNA genotyping, copy-number analysis, isoform expression, microRNA profiling and discovery of novel transcripts or protein/DNA interaction sites. We will here focus on microarrays of the manufacturer Affymetrix with application to gene expression analysis. Other applications and manufacturers differ in the employed protocols, reagents and instruments, but the overall principle is similar for all microarray types. Consider the following four basic elements of a microarray experiment: the microarray with surface-attached probes, the preparation of the target mixture, the scanner device and computational image/data analysis.

The microarray itself refers to a solid surface with attached oligonucleotide probes. Figure 2.1a shows how the surface is separated into thousands of *spots* or *features*. The size of a spot ranges between 5 by 5 square microns (HuExon) and 20 by 20 square microns (HG-U95) [8]. Each spot comprises more than one million oligonucleotides that are, separated by a linker molecule, covalently attached to the surface [9]. The oligonlucleotides are built up one base at a time during fabrication using photolotographic masks [10]. In an ideal production, all oligonucleotides attached to one spot have the same length and identical nucleotide composition termed *probe sequence*.

The mixture sample containing unknown nucleotide strands must be prepared to be suitable for being hybridized to the microarray. Let us consider a target preparation assay for gene expression studies (Affymetrix 3' IVT Express Kit [11]) where one is interested in profiling cellular mRNAs. These assays follow a protocol developed by Van Gelder *et al.* called the ‘Eberwine method’ [12]. After extraction of the total RNA from the cells or tissue of interest, mRNA is reverse-transcribed into complementary DNA (cDNA). The

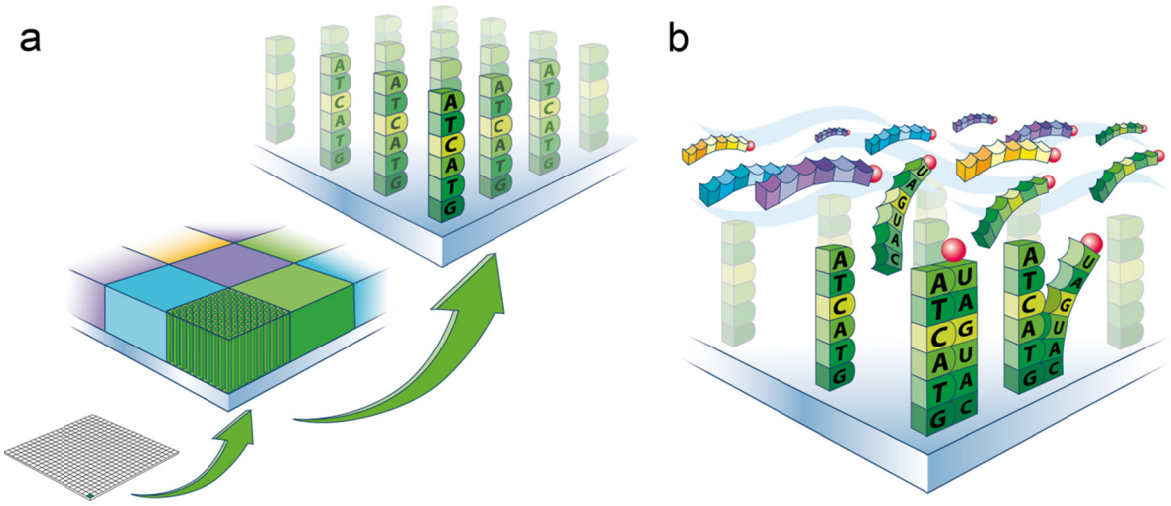

Figure 2.1: Microarray assembly and hybridization. Panel a shows how the microarray is made up of thousands of different features with oligonucleotides of identical sequence attached to the surface of each spot. On panel b marked target nucleotide strands bind to probes with complementary sequence. Image taken from the Affymetrix Image Library [13].

amounts of RNA or DNA required for hybridization to a microarray are typically much higher than the amounts that can be extracted from the cells. Therefore they are amplified using *in-vitro* transcription (IVT) resulting in amplified RNAs (aRNAs or cRNAs). Importantly, the aRNAs are labeled by attaching the marker molecule biotin to a fraction of the nucleotides. The aRNA are then purified and fragmented into shorter nucleotide strands with a typical length between 30 and 200 nt.

The labeled aRNA fragments are then hybridized to the surface-attached probes within a microarray scanner device. This process is allowed to take several hours aiming at reaching an equilibrium state. After that the exceeding sample solution is washed away and the bound aRNAs are stained, that is, large fluorescent molecules (phycoerythein) are attached to the biotin label. A camera then records how laser excites light from the fluorescent molecules. If the target RNA molecules with sequences complementary to a given probe sequence were abundant in the mixture, many aRNAs bind to the respective spot and the fluorescent will shine bright. The light intensity captured by the camera thus relates to the abundance of the targeted RNA.

2.2 3’ expression arrays

Affymetrix GeneChip 3’ expression arrays are among the most widely used microarray types to date - thousands of studies have been carried out on these popular arrays. They are available for over 30 different organisms including human, mouse, rat, zebrafish, yeast, E.coli, tomato, sugar cane and soybean. They are, for example, being used to understand

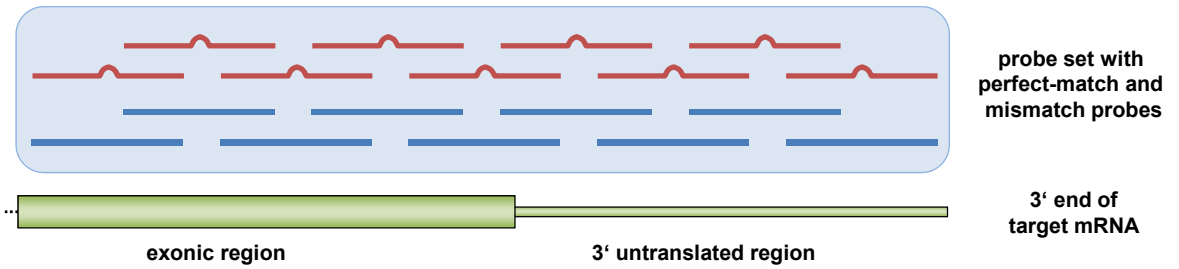

Figure 2.2: Probe design in Affymetrix 3’IVT expression microarrays. Paired perfect-match (blue) and mismatch probes (red) query sequences located towards the 3’ end of the target mRNA (green). Together, these probes form a probe set.

basic mechanisms of molecular biology, to characterize disease states and to classify tumor types, and to assess the transcriptional variation of whole populations.

The length of the spotted oligonucleotides is 25 bases for all types of Affymetrix expression microarrays. These short probes have a relatively low sensitivity for the detection of gene expression changes in complex mixtures [14]. To cope with this shortcoming Affymetrix uses not only a single probe, but instead a *probe set* comprising of 11-16 probes to interrogate each target sequence. The probe set is selected to be “unique to a single transcript or common among a small set of similar transcript variants” [15]. Having multiple intensity measurements for each transcript has several advantages. For example it is hard to predict whether each probe is always fully functioning or if it suffers from deficiencies like strong cross-hybridization to other sequences in the mixture or intra-probe folding. Those errors can be compensated, improving the accuracy of the summarized signal. Furthermore, multiple measurements allow calculating statistics for assessing the confidence in each expression estimate.

The probe sets in 3’ expression arrays are primarily designed to target the 3’ end of the transcripts. Figure 2.2 illustrates how the probes of a probe set interrogate sequences in the 3’ untranslated region (3’ UTR) as well as in the adjacent first exon of a longer transcript. As a result, gene expression estimates from these arrays are necessarily an extrapolation from the 3’UTR abundance of the genes.

Another ‘specialty’ of Affymetrix microarrays is that probes come in pairs: each perfect-match (PM) probe has an accompanying mismatch (MM) probe which has identical sequence except the center base. With the short 25meric oligonucleotides such a single mismatch destabilizes duplex formation between probe and specific target. The ratio behind using mismatch probes is to quantify the sequence-dependent amount of cross-hybridization, which can later be subtracted from the specific signal to improve specificity and sensitivity of the obtained signal [16].

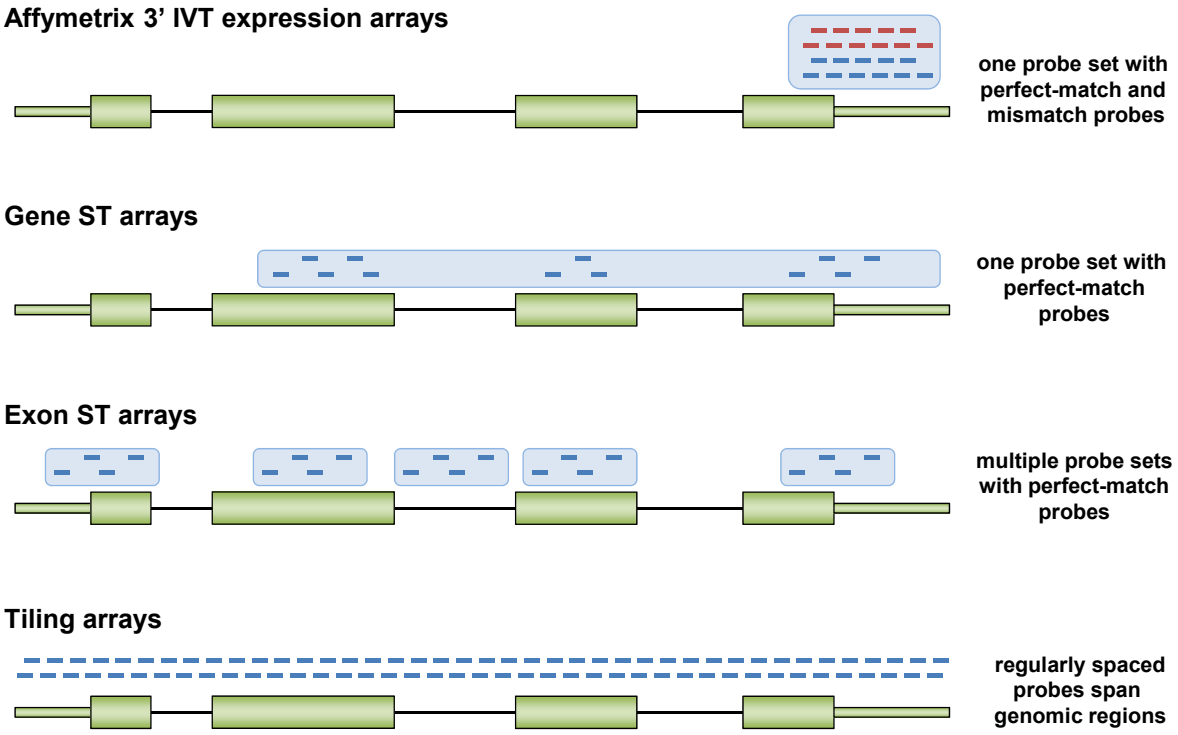

Figure 2.3: Comparison of how probes align to a target gene for various types of Affymetrix microarrays. Whereas probes are located towards the 3' end of the target mRNA (the respective genomic region with exons, introns and UTRs is shown in black and green) in 3' based expression arrays, other array types query sequences in the entire gene. For tiling arrays, probe sets (light blue boxes) are not defined.

2.3 Gene ST and Exon ST arrays

About 40-60% of human genes are not transcribed solely into a single form of mature mRNA [17]. Instead the primary transcripts of these genes are transformed into a number of different isoforms by alternative splicing. Since each splicing isoform can encode for a different, potentially functional protein one is highly interested in their identification and quantification. Affymetrix 3' expression arrays are however by design unable to discriminate splice variants. Gene ST and Exon ST microarrays are designed to overcome these drawbacks.

For one, these *whole transcript expression arrays* employ a different target preparation protocol, typically using the Ambion WT Expression Kit [18]. Synthesis of cDNA strands here is not done using poly-T primers starting at the 3' end of the transcript, but rather using a pool of reverse transcription primers. These bind at various loci in non-ribosomal RNAs to initiate the polymerase reaction. In-vitro transcription is then used to amplify these fragments which span various regions of the available transcripts. Biotinylated sense-strand cDNA, opposed to the cRNA used in 3' IVT expression arrays, is then fragmented and end-labeled for hybridization to the array. The resulting DNA-DNA duplexes between probes and targets have been found to be more specific than DNA-RNA duplexes [19].

The probes of whole transcript arrays interrogate sequences spread across the entire gene with the aim of getting a more complete picture of gene expression. As shown in Figure 2.3, the probe set of a 3' IVT array contains a fixed number of perfect-match and mismatch probes which concentrate at the 3' end of the transcript. A transcript is queried by typically one probe set. For the Exon ST arrays, each exon or non-coding region is interrogated by about four probes. Using these exon-level probe sets allows distinguishing between different splicing isoforms. The probes of multiple exons can be combined, giving about 40 probes per gene and allowing a complementary gene-level expression analysis. The Gene ST arrays are designed as a less expensive alternative to the Exon ST arrays containing only a subset of the probes mainly designed for gene-level analysis. A high concordance has been found between the gene-level estimates of Gene ST, Exon ST and 3' IVT expression arrays [20, 21].

It should be noted that Gene ST arrays are less popular than Affymetrix' 3' expression arrays. McCall *et al.* found that "between 1 June 2010 and 1 June 2011, over 13 000 Affymetrix Human Genome U133 Plus 2.0 samples were added to the Gene Expression Omnibus (GEO)" but "during the same time period, less than 2000 Human Gene 1.0 ST samples were added" [22].

2.4 Genome-wide SNP arrays

Another important application of microarrays is the analysis of genetic variants. In diploid human cells the genetic information is spread on two homologous sets of 23 chromosomes. Alleles are alternative forms of a certain position or region of a chromosome (a locus) that occur between members of a species or within the chromosome set. In the case of the most common type of variation, the single nucleotide polymorphism (SNP), only a single base of DNA is altered. Since there are four possible nucleotides a SNP can have at most four alleles. Most SNPs have however only two alleles [23]. These bi-allelic loci result in three possible states a SNP can take in a diploid chromosome set: either homozygous allele AA with allele A on both chromosomes, homozygous allele BB, or heterozygous AB with two different alleles on both chromosomes. Genotype calling or *genotyping* aims at inferring these states.

Another form of variation measured by microarrays is copy-number variants. These are alterations of chromosome structure in which large segments (> 1 kb) of the DNA are present in variable copy number compared to a reference genome [24]. A duplication of certain segment of the chromosome, for example, would have the effect that all previously unique genes in that section are now present in two copies. About 12% of the human

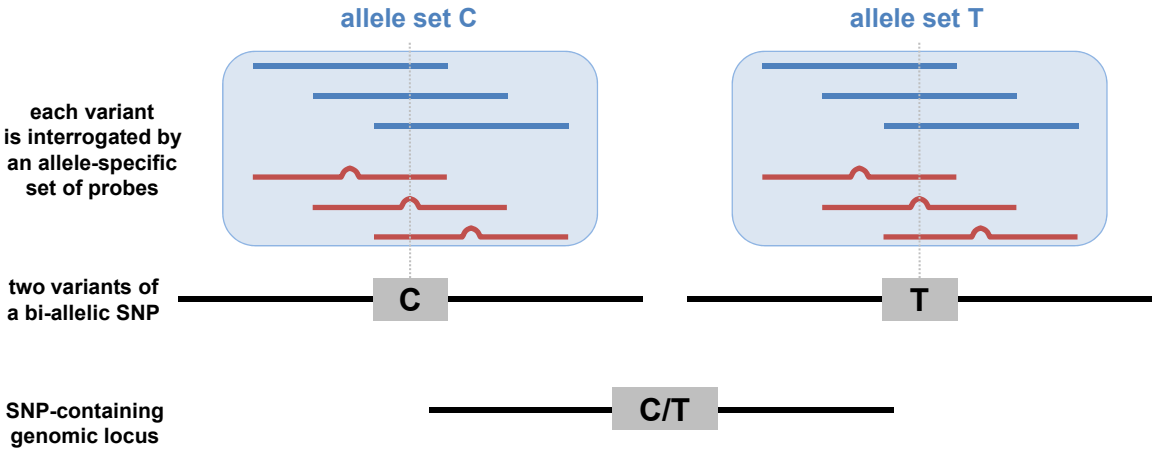

Figure 2.4: Probe design of Affymetrix SNP Arrays. All probes (blue PM probes and red MM probes) interrogate a single SNP located in genomic DNA. The SNP has the two alleles C and T each being interrogated by an allele set of probes.

genome has been found to be covered by copy number variations [25] rendering them an significant source of genome heterogeneity and a potential factor contributing to phenotypic variation and disease states/susceptibility.

Specific target preparation assays and microarray designs are employed to allow detection of genetic variants with high sensitivity. Compared to gene expression experiments, these assays do not target (m)RNA molecules but instead genomic DNA. Total genomic DNA is digested with restriction enzymes (see Genome-Wide Human SNP Nsp/Sty Assay Kit 6.0 documentation [26]). Adapters are ligated to the resulting fragments which are then used for a PCR procedure that has been optimized to amplify fragments of certain size range to reduce complexity of the genomic DNA. The amplified DNA is further fragmented, end-labeled and finally hybridized to the array [27].

The probes are designed to tile around each SNP with slight variations in perfect matches, mismatches, and flanking sequence [28] as shown in Figure 2.4. The Affymetrix GeneChip Human Mapping 100k Array Set, for example, uses 40 different 25meric probes for each SNP. For each of the two interrogated alleles there is an *allele set* consisting of 10 probe pairs: 10 PM probes and 10 corresponding MM probes with a mismatch at the center base, depicted separately in Figure 2.4. The probes include the SNP at the center base or are slightly shifted by some offsets $\delta = -4,..0,..4$. Of the 10 PM probes 3 to 7 target the sense strand whereas the remaining ones target the antisense strand. This design with a large number of probed sequence combinations can be used to study the impact of mismatches and other duplex interactions on probe signals [29]. Some arrays such as the Genome-Wide Human SNP Array 6.0 omit the mismatch probes which makes it possible to capture 1.8 million genetic variants with about 6 million probes.

2.5 Agilent expression arrays

Agilent's manufacturing technology differs from that of the other two major producers of high-density microarrays namely Affymetrix (which use photolithographic masking [10]) and Illumina (which use self-assembling silica beads [30]). Agilent prints its arrays similar to how an inkjet printer prints a document - instead of ink on paper, nucleic acids are printed base by base onto the glass surface [31]. A major advance of this technology is that the features can easily be customized for each microarray: probes are designed to interrogate the targets of interest and then added or removed as desired. This flexibility is not given for Affymetrix technology, where only standardized expression microarrays are available.

The most recent Agilent SurePrint G3 Gene Expression Microarrays comprise more than one million features. The printed oligonucleotides have a length of 60 bp. These 60mer probes were shown to be significantly more sensitive to expression changes in complex mixtures compared to 25mer oligonucleotides [14], according to Agilent between five and eight times [32]. Longer probes are however less specific – 25mers are about 20 times more specific for differentiating a single mismatch [14]. This tolerance with respect to sequence mismatches can however also be an advantage when probing highly polymorphic regions. Agilent arrays support different target preparation assays including two-color and one-color preparations.

2.6 Summary and conclusions

Microarrays come in a diverse set of flavors aiming at different genomics applications ranging from gene expression analysis and profiling over DNA analysis and genotyping to gene regulation analysis. The great utility of microarrays in these fields of applications has driven - and vice versa has been driven by - many developments in the private and in the academic sector resulting in the rapid advancement of the technology since its appearance in the 90s. These improvements in terms of accuracy, coverage, reproducibility, standardization and cost have made microarrays an established tool widely used in research and even in clinical settings [33].

The variety in the set of possible applications is enabled by differences in microarray designs and protocols. Specifically, Affymetrix 3' expression arrays target sequences that reside within the 3' UTR and act as a proxy for the expression of the respective gene; exon arrays interrogate sequences from exons of known splice isoforms, and tiling arrays have their probes distributed uniformly across large fractions of the genome. Additional to these application-specific differences, each microarray manufacturer has its own ways of

production and supports its own instruments and reagents. Affymetrix provides an unparalleled coverage and feature density, as well as a high standardization. Agilent in turn provides highly customizable microarray designs.

3 A model for microarray hybridization

3.1 Modeling microarray intensity signals

The presented technologies share the common mechanism of multiplexed hybridization of fluorescently labeled target molecules against known oligonucleotide probes. The input quantity that one wishes to infer in a microarray experiment is the abundance, or concentration, [S] of specific nucleic acid targets. The measured output quantity is fluorescence signal intensities I for the surface-attached probes. Modeling microarray hybridization with the aim of obtaining accurate signal calibration consequently seeks to identify an adequate functional relationship $I^p = f([S_g])$ between a probe p and the respective target (gene) g.

Several effects in the microarray measurement prevent an accurate description of the input and output quantity via the simple proportional relationship $I \propto [S]$. Firstly, there are technical limitations in the optical recording of the intensity signals using the scanner. Even when no specific transcripts are bound to the probes the scanner reports positive intensity values $I > 0$. An additive optical background term O, i.e. in the form $I = [S] + O$, should therefore be considered in microarray calibration methods [34, 35].

Secondly, several fundamental binding and folding processes can occur at or near the microarray surface as shown in Figure 3.1a. The yield of the interaction between free probes and specific targets is reduced by bulk-dimerization, non-specific hybridization and intra-molecular folding reactions. During non-specific hybridization additional to the fully complementary specific targets other, only partly complementary, DNA or RNA fragments bind to the probes. Due to the large diversity and quantity of target molecules in the complex mixture solution this type of binding typically is considerable [36]. A practicable solution for incorporating non-specific binding in the hybridization model is to summarize the diversity of non-specific transcripts into a single probe-specific term, i.e. $I = [S] + [N]$ (see also [35, 37]).

Thirdly, the kinetics of the reversible binding reactions of targets in excess to limited, surface-attached oligonucleotides can result in a non-linear response of the probe intensity. The binding reactions can be regarded as a Langmuir adsorption process as exemplified in Figure 3.1b. Accordingly, the amount of adsorbed molecules Θ on a surface in dependence

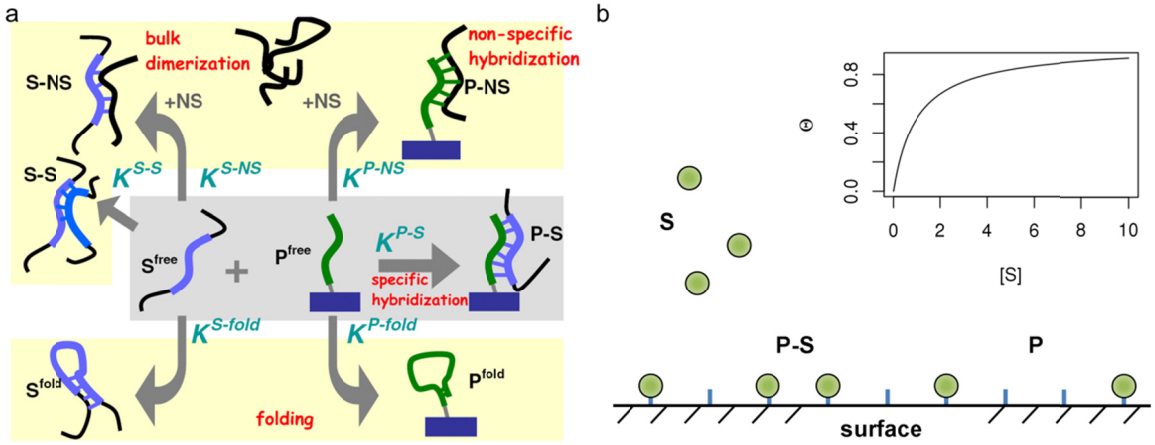

Figure 3.1: Interaction processes and dynamics of surface adsorption on microarrays. Panel a shows possible binding and folding processes of probes and targets (image taken from [37]*). Panel b visualizes the Langmuir adsorption model describing the occupancy Θ of surface sites (e.g. probes P) with particles (e.g. specific transcripts S).*

of the molecule concentration [S] can be described with the Langmuir isotherm

$$\Theta = \frac{K \cdot [S]}{1 + K \cdot [S]} \tag{3.1}$$

where K is the rate constant. In the case of microarray hybridization, Θ represents the fractional coverage of oligonucleotides of a given probe and [S] the concentration of relevant target molecules in the solution. The Langmuir adsorption model has been shown to describe well microarray signals based on experiments with known target concentrations [38–41].

The three described effects relate either to basic technical limitations of the instruments or to physicochemical principles of the hybridization processes and strongly influence the obtained intensity signals of a microarray experiment. One or more of these factors are considered in virtually any calibration method. Importantly, the shown relevance of physicochemical models for describing the basic processes of microarray hybridization suggests that building upon these models and refining them using additional knowledge about relevant mechanisms is a promising strategy for signal calibration and beyond.

3.2 The two-species Langmuir model

We will now introduce the two-species Langmuir model which applies well to microarray expression data [36, 42]. Accordingly, the intensity of a probe p of type $P \in \{PM, MM\}$ measured in a microarray experiment is given to a good approximation by

$$I_p^P = M \cdot \frac{X_p^{P,N} + X_p^{P,S}}{1 + (X_p^{P,N} + X_p^{P,S})} + O \tag{3.2}$$

where M is the maximum intensity upon saturation and O is the optical background intensity. We here assume that the term O can be corrected in a separate step, for example using the Affymetrix zone algorithm [7], and will rely on background-corrected probe intensities if not stated otherwise. The numerator $L_p^P \equiv M \cdot (X_p^{P,N} + X_p^{P,S})$ is also denoted the linearized signal and decomposes into contributions due to non-specific and specific binding (see next section) scaled by M. The binding strengths $X^{P,h}$ linearly scale with the respective concentration of specific and non-specific targets, $X^{P,h} \propto [h]$ with $h \in \{N,S\}$. Only considering the factors described in the previous section, the binding strengths are given as

$$X_p^{P,S} = [S_g] \cdot K_p^{P,S} \quad \text{and} \quad X_p^{P,N} = [N]_{chip} \cdot K_p^{P,N} \tag{3.3}$$

where $K_p^{P,h}$ are the equilibrium constants for the formation of probe/target duplexes.

Two factors not considered in this thesis are washing and target depletion. The washing step that follows hybridization in the microarray assay has been shown to remove probe-bound targets and inversely scales with the respective binding constants [43]. Target depletion in the solution can lead to an underestimation of the concentrations of specific transcripts [44].

3.3 The hook transformation and hybridization modes

The parameters of the Langmuir-type model are not directly accessible given only the intensity signals of the particular microarray hybridization. The target concentrations are unknown in typical applications and the specifics of the hybridization reaction can differ for each microarray experiment. The hook method elegantly solves this challenging problem by using information inherent in the coupled signals of perfect-match (PM) and mismatch (MM) probe pairs [45]. These paired probe signals are transformed in a special mean-difference plot:

$$\begin{aligned} &\Delta^{hook} = mvg\_avg(\Delta_{pset}) \text{ and } \Delta_{pset} = \langle \Delta_p \rangle_{pset} \\ &\Sigma^{hook} = \Sigma_{pset} = \langle \Sigma_p \rangle_{pset} \\ &\text{with } \Delta_p \equiv \left(\log I_k^{PM} - \log I_k^{MM} \right) \quad \text{and} \quad \Sigma_p \equiv \frac{1}{2} \left(\log I_p^{PM} + \log I_p^{MM} \right). \end{aligned} \tag{3.4}$$

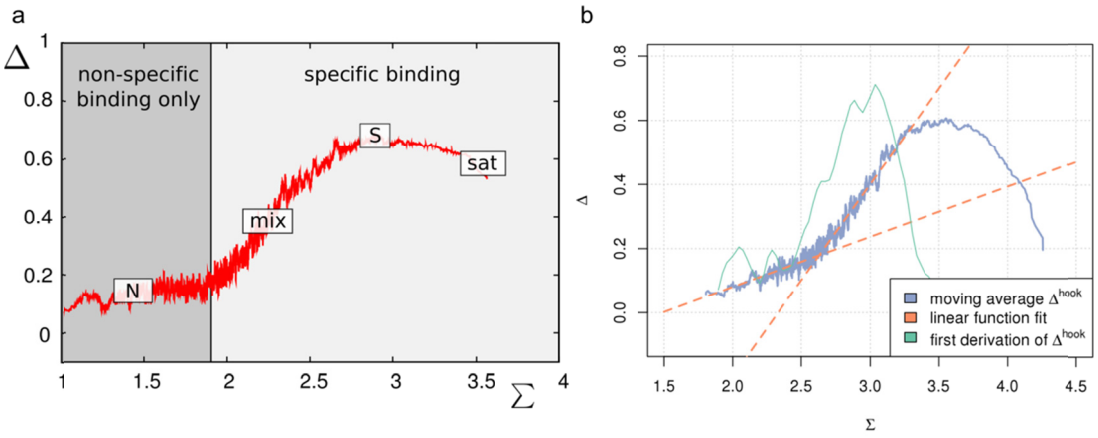

Figure 3.2: Hook curve with different binding regimes, and computation of Σ^{break}. Panel a shows the N-, mix-, S- and sat regimes and the Σ^{break} threshold (black vertical line) separating non-specific binding and the onset of specific binding. Panel b shows the first derivation and the linear fits which are used for the estimation of Σ^{break}.

The Δ and Σ transformations are based on the PM/MM difference and average, respectively, of logged probe intensity values ($\log \equiv \log_{10}$ is the decadic logarithm). $<\ldots>_{pset}$ denotes averaging over all probes within a probe set (typically 11 to 16 probes targeting the same transcript) to obtain robust Δ_{set} and Σ_{set} values. The operator mvg_avg applies a moving average to the Δ_{set} values with a window size of about 100 probe sets for smoothing. Plotting the microarray data into Δ-versus-Σ coordinates provides the hook curve which enables decomposition of the probe signals into contributions due to different modes of hybridization by simple visual inspection.

Particularly, we differentiate between two modes: non-specific (N-) and specific (S-) hybridization. In the S-hybridization mode the probes bind the aRNA fragments of complementary sequence transcribed from mRNA transcripts which they intend to detect. In the N-hybridization mode the probes bind aRNA fragments of partly complementary sequence originating however from mRNA transcripts not referring to the interrogated gene. Probe binding of this type is termed (ubiquitous) cross-hybridization (e.g. [46]), non-specific binding (e.g. [47]) or non-specific hybridization (e.g. [36]).

Figure 3.2a shows the resulting hook plot for a typical GeneChip expression array. Its visual inspection allows the simple and straightforward detection of five hybridization regimes with increasing Σ, namely the N- (virtually only non-specific hybridization contributes to the signals), mix- (combination of non-specific and specific hybridizations), S- (predominantly specific hybridization), sat- (saturation range; the relation between intensity and transcript concentration becomes progressively non-linear) and as- (the intensity reaches its asymptotic saturation level) regime.

Consider the N-regime referring to probe sets with the smallest Σ values. How can the at most weakly increasing Δ values that scatter around 0 be explained? $\Delta \approx 0$ refers to an

equal intensity level of PM and MM probes for probe sets in this Σ-interval. Targets with exact complementary sequence are however expected to have a significantly higher binding strength compared to those with a single mismatch in the middle of the probe-target duplex. The probes consequently do not bind to the interrogated target but to fragments of partly complementary sequence – they bind non-specifically. Probes or probe sets with virtually only non-specific hybridization are called *absent* whereas others are called *present*.

Figure 3.2a also indicates a Σ threshold characterized by a significant increase of the Δ values referring to the onset of specific binding. This threshold Σ^{break} separates the N and mix-regime and consequently separates absent and present probe sets. The ability for this separation allows the estimation of essential parameters of the hybridization model [42].

Estimation of Σ^{break} should be both accurate and robust. The following heuristic method has been shown to improve over previously proposed approaches and delivers reliable results over a large variety of chip-types [42, 48, 49]:

1. Compute the empirical first derivation of the hook plot (by fitting a straight line to 7 subsequent data points of (Σ, Δ) sorted by Σ)
2. Find the point of maximum deviation $\Sigma_{max\text{-}d}$
3. Use linear regression to find the best joint fit of two straight lines ($y = mx + n$) to all data points between the smallest Σ of the hook plot and $\Sigma_{max\text{-}d}$ (see [49] for details about the formulation of the least squares error).
4. The intersection point between the two lines defines $(\Sigma^{break}, \Delta^{break})$

Figure 3.2b illustrates this approach of estimating Σ^{break}. The green line is the empirical first derivation computed from the hook curve (shown in blue). The maximum derivation is located in the mix regime. The best fitting two straight lines (shown in orange) intersect at the threshold $\Sigma^{break} \approx 2.6$.

3.4 Positional-dependent sequence models

3.4.1 Modeling the formation of duplexes

In our model, the binding constants from Eq. (3.3) decompose into

$$K_p^{P,h} = K_0^{P,h} \cdot \exp(\delta A^{P,h}(\xi_p)) \tag{3.5}$$

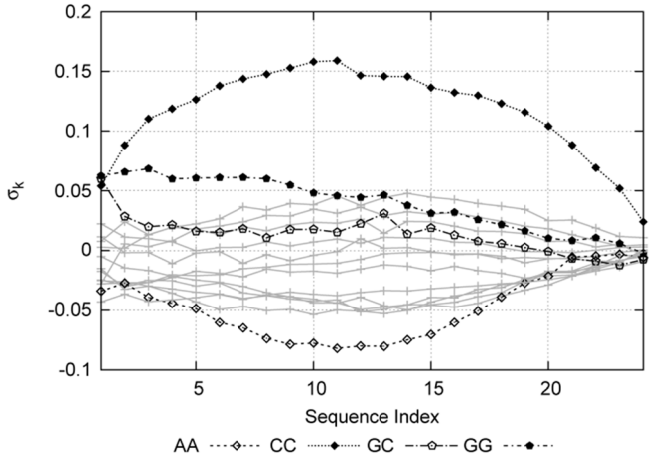

Figure 3.3: Typical sensitivity profiles of rank r = 2 for an Affymetrix HG-U133a microarray. Points indicating positional-dependent sensitivity terms for four selected base tuples (AA,CC, GC, GG) are shown in black, whereas the points and connecting lines of the remaining 12 base tuples are shown in gray.

where $K_0^{P,h}$ is a probe independent contribution and $\xi = \xi^{1,25}$ is the probe sequence string of length 25. We further use the convention $\xi^{k,k+r-1}$ to assign the subsequence of r adjacent nucleotides starting at position k in ξ. The sequence effect δA is modeled using the sum of sensitivity terms over all sequence positions [42, 50, 51]

$$\delta A^{P,h}(\xi_p) = \sum_{k=1}^{25-r+1} \sigma_k^{P,h}(\xi_p^{k,k+r-1}). \tag{3.6}$$

The sensitivity profiles $\sigma_k^{P,h}(b_r)$ depend on base tuples $(b_r)_k = (B_1 \ldots B_r)_k$ (with $B_i \in \{A,T,G,C\}$, $1 \le i \le r$) of length r with its first base at position k of the probe sequence. For example, $(GGG)_1$ denotes a sequence motif containing three adjacent guanines beginning at sequence position 1. The parameter r specifies the rank of the model. Thus, $r = 1 \ldots 4$ refers to the single nucleotide (N), nearest neighbor (NN), next nearest neighbor (NNN) and quadruple (NNNN) models, respectively.

Figure 3.3 shows sensitivity profiles of rank r = 2 estimated from the intensities of non-specific, perfect-match probes (P = PM, h = N) of an Affymetrix HG-U133a microarray sample taken from an experiment by Su *et al.* [52]. The positional-dependent contributions are roughly symmetrical around the middle of the probe sequence, except for GC and GG base tuples where the sensitivities decrease monotonically with increasing sequence index. The sensitivities of all profiles converge towards the surface-attached side at k = 24 but differ at the solution end at k = 1. The base tuples AA and CC exhibit the maximum sensitivity amplitudes.

Integral sensitivities $\sigma_{ku,ko}^{P,h}(b_r)$ are calculated by summing up the positional dependent values either over all sequence positions or over a positional range that was selected, for

example, to exclude the region of the $(GGG)_1$ effect

$$\sigma_{ku,ko}^{P,h}(b_r) = \sum_{k=ku}^{ko} \sigma_k^{P,h}(b_r). \tag{3.7}$$

3.4.2 Different characteristics for specific and non-specific binding

Binding characteristics are known to be different between specific and non-specific hybridization modes [36]. Eq. (3.2) simplifies into $M \cdot (I_p - O) \approx L_p^N$ for the special case of predominantly non-specific binding far below saturation, $L_p^N \ll L_p^S \ll M$. Restricting our basic analysis to this regime, we ensure linearity of the intensity response and homogeneous probe-target interactions. The latter are mainly governed by canonical Watson-Crick pairings [53].

We select the subensemble of probes meeting these conditions using the hook method [42]. Typically more than 40% of all probe sets are called 'absent' in a particular microarray hybridization, providing a sufficient number of probe intensities to adequately fit the model (see also Table 6.1 in Chapter 6).

The ensemble of present (i.e. not-absent) probes refers to signals which partly or completely originate from specific hybridization. We apply the hook method to filter out probe sets which hybridize predominantly with specific transcripts, ($p \in S$), and to correct their intensities for the effect of saturation (see [42] for details).

3.4.3 Estimation of profiles

We define the experimental sensitivity of each probe as the deviation of the logged linearized signal of its average over all probes of the respective probe set [54]

$$Y^{exp} = \log X^{P,h} - \langle \log X^{P,h} \rangle_{pset}. \tag{3.8}$$

After insertion of Eqs. (3.5) and (3.6) into (3.8) and making use of $\log(K_0^h[h]) = \langle \log(K_0^h[h]) \rangle_{pset}$ we obtain the theoretical sensitivity of each probe

$$Y^{theo} = \sum_{k=1}^{25-r+1} \sum_{br} \sigma_k(b_r) \cdot (\delta(b_r, \xi^{k,k+r-1}) - f_k^{pset}(b_r)) \tag{3.9}$$

with the Kroenecker function $\delta(x,y)=1$ for $x=y$ and $\delta(x,y)=0$ otherwise. $f_k^{pset}(b_r)$ is the probability to find motif b_r at sequence position k among the probes of the considered probe set. Note that the transcript concentration (specific and non-specific) is assumed to be constant for each probe set because each probe within the set targets the same transcript. This condition cancels the term $\log(L_0)$ in Eq. (3.5).

The sensitivity profiles are estimated using multiple linear regression. It minimizes the sum of squared residuals [42]

$$SSR(r)=\frac{1}{\#p}\sum_p RES^2=\langle RES^2\rangle \tag{3.10}$$

with $RES=(Y^{exp}-Y^{theo})$ by optimizing $\sigma_k(b_r)$ for all $4^r\cdot(25-r+1)$ base tuples $(b_r)_k$. The sum runs over all relevant probes $p\in N$ or $p\in S$ and #p defines the respective number of probes. The obtained sensitivity terms meet the center condition $\sum_{\text{all } b_r}\sigma_k(b_r)=0$ for each sequence position k.

3.5 Fitting the hybridization model

Application of sequence correction leads to a less noisy and more consistent hook curve. Figure 3.4 shows two versions of the hook curve: before (panel a) and after correction (panel b) of the signal intensities used for the computation of the Δ-Σ-transformation given in Eq. (3.4) with the positional-dependent nearest neighbor model (r = 2) from the previous section. The sequence correction improves the precision of the probe signals: the within-probe set variability is reduced as well as the scattering of probe set averages around the hook curve. Basic features of the hook curve such as the relative positioning of the binding regimes are essentially the same in both versions. The N-regime however differs significantly in its width and slope. In summary, these effects result in an improved hook curve which is sufficiently robust to allow fitting the theoretical hybridization model as described below.

Let us now give a formulation of the two-species Langmuir model that predicts the Δ and Σ coordinates of the hook curve. We define the relative hybridization degree, or S/N ratio R as

$$R_p\equiv\frac{K_p^{PM,S}\cdot[S_g]}{K_p^{PM,N}\cdot[N]_{chip}}. \tag{3.11}$$

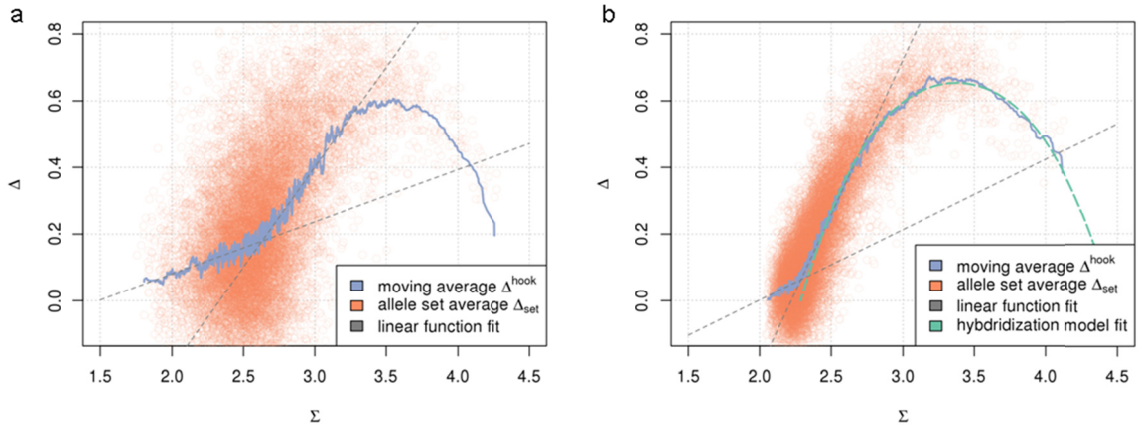

Figure 3.4: Δ and Σ transformations calculated from raw (panel a) and sequence corrected (panel b) probe intensities of a GeneChip Rat Expression Array 230A. Sequence correction leads to an improved hook curve to which we fit the hybridization model.

R is an expression measure that directly relates to the concentration of specific transcripts. We further define pairwise PM/MM ratios of the binding constants

$$s_p = \frac{K_p^{PM,S}}{K_p^{MM,S}} \text{ and } n_p = \frac{K_p^{PM,N}}{K_p^{PM,N}}. \tag{3.12}$$

Inserting Eqs. (3.11) and (3.12) in the basic model given in Eq. (3.2) and afterwards in the hook transformations Δ_p and Σ_p given in Eq. (3.4) we obtain

$$\begin{aligned} &\Delta(R) = \Delta^{start} + \log\left[(R+1)/\left(R\cdot 10^{-\alpha}+1\right)\right] - 1\log\left[B^{PM}(R)/B^{MM}(R)\right] \\ &\text{and} \\ &\Sigma(R) = \Sigma^{start} + \tfrac{1}{2}\log\left[(R+1)\cdot\left(R\cdot 10^{-\alpha}+1\right)\right] - \tfrac{1}{2}\log\left[B^{PM}(R)\cdot B^{MM}(R)\right] \end{aligned} \tag{3.13}$$

where $B^{PM}(R) = 1+10^{-\left(\beta-\frac{1}{2}\Delta^{start}\right)}(R+1)$ and $B^{MM}(R) = 1+10^{-\left(\beta+\frac{1}{2}\Delta^{start}\right)}\left(R\cdot 10^{-\alpha}+1\right)$ are saturation terms.

Plotting the Δ and Σ trajectories in dependence of the relative expression degree R gives the *theoretical hook curve* shown in Figure 3.5. The parameters Δ^{start}, Σ^{start}, α and β characterizing the position and the geometrical dimensions of the theoretical hook curve are given as

$$\begin{aligned} &\alpha = \log\frac{s}{n} \quad \text{and} \quad \beta = \tfrac{1}{2}\log n - \left\langle \log\left(K^{PM,N}\cdot[N]\right)\right\rangle_{chip} \\ &\Delta^{start} = \log n \quad \text{and} \quad \Sigma^{start} = \log M - \beta \end{aligned} \tag{3.14}$$

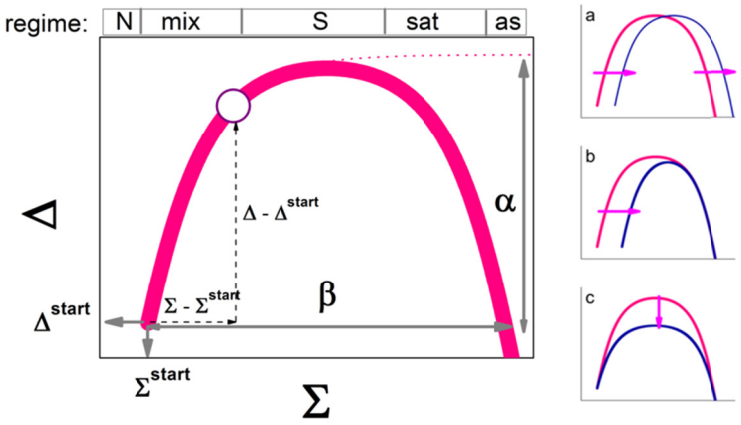

Figure 3.5: Theoretical hook curve and its geometrical dimensions. The start coordinates (Σ^{start}, Δ^{start}) characterize the non-specific background level in intensity units and the N-PM/MM-gain, respectively. (α, β) characterize the width of the hook and its height in the absence of saturation, respectively. Image taken from [55].

where $n = \langle n_p \rangle_{chip}$ and $s = \langle s_p \rangle_{chip}$ are the PM/MM ratios from Eq. (3.12) averaged for all probes on the array. A detailed description of derivation of the model is given in [42].

The alternative formulation of the two-species Langmuir model given in Eq. (3.13) can now be fitted to the experimental Δ-Σ-trajectories as explained in [42, 49]. The respective optimization problem has only a single local maximum and can be solved using a gradient descent algorithm. Figure 3.4b shows that the theoretical hook curve predicted by the hybridization model fits well to Δ and Σ transformations of the sequence-corrected intensity data of a typical GeneChip expression array. The fitting provides chip-specific parameters which can be used to compute estimates of the transcript concentration [S] based on the model Eq. (3.2). The *hook-method* for signal calibration based on this fitting and reversal of the two-species Langmuir model is described in detail in [42].

3.6 Chip summary measures characterize RNA quantity

Additional to their utility for the calibration of microarray signals, the basic hook parameters from the previous section (Δ^{start}, Σ^{start}, α and β) characterize a particular microarray hybridization and can be used to compare samples, or groups of samples, of an experiment, for example to identify laboratory effects. These parameters are important indicators for sources of technical variability which are difficult to detect by other means. This section shows how selected chip-specific summary measures that can be obtained alone from the microarray signals relate to the non-biological variable of the amount of hybridized RNA.

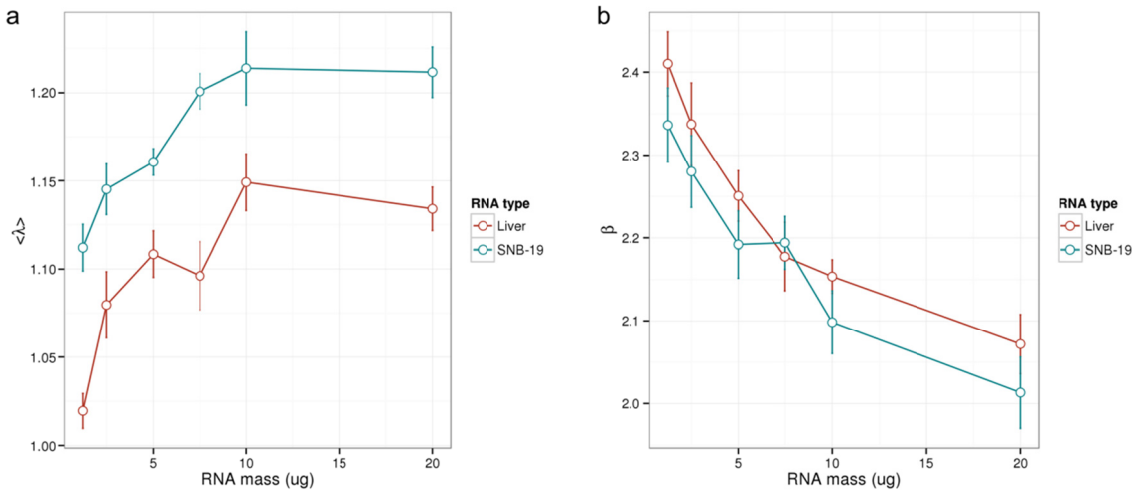

Figure 3.6: Chip-specific parameters <λ> and β in dependence of the amount of hybridized RNA. We computed both parameters for the samples of Gene Logic's dilution data set where two types of RNA (liver and SNB-19) have been hybridized at varying concentrations with 5 replicate samples for each concentration. In Panel a <λ> increases roughly linear with increasing RNA mass between 1 and 10ug (Pearson correlations of r > 0.7), but saturates at 20ug. Panel b shows how β decreases with increasing RNA mass. An amount of 10ug aRNA is recommended for the employed HG-U95A platform.

The S/N ratio, R, from Eq. (3.11) characterizes the specific hybridization level of a transcript relative to the baseline of non-specific binding. Averaging these R values over all probe sets of a microarray exceeding a given expression threshold provides the *relative specific transcript level*, or mean log S/N ratio,

$$\langle \lambda \rangle = \langle \log(R+1) \rangle_{R>0.5;\text{chip}} . \tag{3.15}$$

This chip-specific measure is typically in the range of $0.2 < \langle \lambda \rangle < 1.5$ and also describes the R-range over which the density of expression values decays by one order of magnitude [45].

We computed β and $\langle \lambda \rangle$ summary measures for all samples of a dilution experiment conducted by Gene Logic Inc. [56]. In this experiment two distinct types of RNA samples, liver tissue and CNS Cell Line SNB-19, have been hybridized to Affymetrix Human Genome U95A arrays at varying concentrations[2]. Multiple samples have been prepared from total RNA according to the manufacturers' protocol and the resulting aRNA has been collected into one master solution for each of the two RNA types. The master solutions, whose RNA concentrations have been determined using an electropherometer (at 260nm), were then diluted to generate solutions with nominal aRNA masses between 1.25 and 20ug. Additional dilutions containing mixed RNA of both liver and SNB-19 are available but have been omitted here. Five technical replicates were processed for each

[2] A detailed description of the study design is given in the white paper accompanied by the data which can be ordered from Gene Logic Inc.

concentration, leaving a total of 50 samples. This study design enables an assessment of the effect of technical variation within the replicates and between the dilutions.

Panel a of Figure 3.6 displays the obtained $\langle\lambda\rangle$ parameters in dependence of RNA mass for the 50 microarray samples. $\langle\lambda\rangle$ increases with increasing RNA mass between 1 and 10ug with Pearson correlation coefficients of $r = 0.71$ for liver tissue and $r = 0.78$ for SNB-19. However, $\langle\lambda\rangle$ does not increase further for a RNA mass of 20ug which can be explained by the up-down effect: increasing RNA concentrations result in a larger non-specific background accompanied by a smaller effective specific binding constant due to bulk dimerization [57]. The $\langle\lambda\rangle$ summary measure which averages the ratio S/N of specific and non-specific binding (see Eqs. (3.15) and (3.11)) is therefore not collinear with RNA mass as the effect of bulk dimerization is not considered in the hybridization model (Eq. (3.2)). In summary, $\langle\lambda\rangle$ describes the amounts of aRNA in a particular microarray hybridization in a non-linear, yet for typical RNA ranges sensitive fashion.

The β parameter from Eq. (3.14) characterizes the width of the theoretic hook and typically is in the range $2.0 < \beta < 3.2$. Since $\beta = \log M - \Sigma^{start}$ (Eq. (3.14)), the width is limited by the saturation level M and the start of the theoretic hook Σ^{start} at the onset of specific binding, and thus describes the measuring range of specific signals [45]. As shown in Figure 3.6b the parameter β decreases with increasing RNA mass in the dilution experiment. The increasing concentration of RNA in the hybridization solution here results in an increased signal contribution due to non-specific binding and thus in a non-linear, negative effect on the measuring range β.

3.7 Summary and conclusions

Multiplexed hybridization reactions between nucleic acids on a surface can be well described using the two-species Langmuir model. This model is based on fundamental physical principals of surface adsorption and can be easily refined to incorporate additional factors such as sequence-dependent affinities, washing and degradation. Critical is the fitting of the model to the intensity data which should ideally be performed separately for each microarray hybridization due to a significant variation in the described biological and technical factors. The Δ-Σ-transformations provide a practical way to fit the model without prior knowledge of target concentrations.

The Δ-Σ-transformations of probe signals result in the typical hook curve which allows identification of different hybridization regimes. We distinguish between N-, mix-, S- and sat regime depending on the effect of non-specific binding, specific binding and of saturation. The threshold Σ^{break} separating N- and mix- regime is of particular importance

as it allows independent characterization of specific and non-specific binding. A robust and accurate method to estimate Σ^{break} is presented.

Additional to their utility for signal calibration, the hybridization model and the hook approach allow studying the technical factors and physicochemical processes involved in microarray hybridization. Fitting an alternative formulation of the two-species Langmuir model provides summary parameters which are of great utility for the assessment of the non-biological variation among the samples of an experiment. Examples are the $\langle\lambda\rangle$ and β parameters which relate to the abundance of hybridized RNA.

4 Hook analysis applied to different types of microarrays

In the previous chapter we presented a model for microarray hybridization and showed that it adequately predicts microarray data based on the Σ–Δ-transformations of intensity signals. We showed this exemplary for Affymetrix GeneChip 3' IVT expression arrays and it remains an open question whether the same approach can also be applied to other microarray types: those that interrogate DNA instead of RNA targets, those that employ different protocols or different designs, and those produced by other manufactures relying on their own proprietary technologies. The ability to obtain suited Σ-Δ-transformations provides the basis for successful application of the methods described in this thesis to these microarray types.

4.1 Genome-wide SNP arrays

In Section 2.4 we defined so-called allele sets for SNP arrays in analogy to the probe set in expression arrays. All probes within an allele set interrogate a unique variant of a target nucleotide strand, here however referring to fragments of genomic DNA containing a particular SNP. We calculated Δ and Σ transformations according to Eq. (3.4) for a 50K Array Xba 240 from the Human Mapping 100k Array Set as shown in Figure 4.1a. Note that these arrays contain both perfect-match and mismatch probes. Basic features of the obtained SNP hook curve are strikingly similar to those of expression arrays (compare Figure 3.2a): at $\Sigma = 2.3$ the curve starts with small values of $\Delta \approx 0$ which, due to an increasing contribution of specific binding, increases monotonously to $\Delta = 0.6$. The onset of saturation then results in a decrease of the Δ values with increasing Σ up to the highest probe signals at $\Sigma = 4.1$.

A noteworthy difference between both array types is that expression arrays have a distinct N-regime containing a substantial amount of probe sets with $\Delta \approx 0$. A significant change of the slope of the hook curve at Σ^{break} separates this region from the subsequent mix-regime. In the SNP arrays no change of slope can be observed. The distribution of probe sets in Δ-Σ-coordinates (orange circles in Figure 3.2 and Figure 4.1) shows that a large fraction of probe sets are in the N-regime in expression arrays, whereas in SNP arrays only few probe sets are in this regime of predominant non-specific binding. This is not surprising because by design each allele set should contain probes exactly complementary to the present genomic DNA variant.

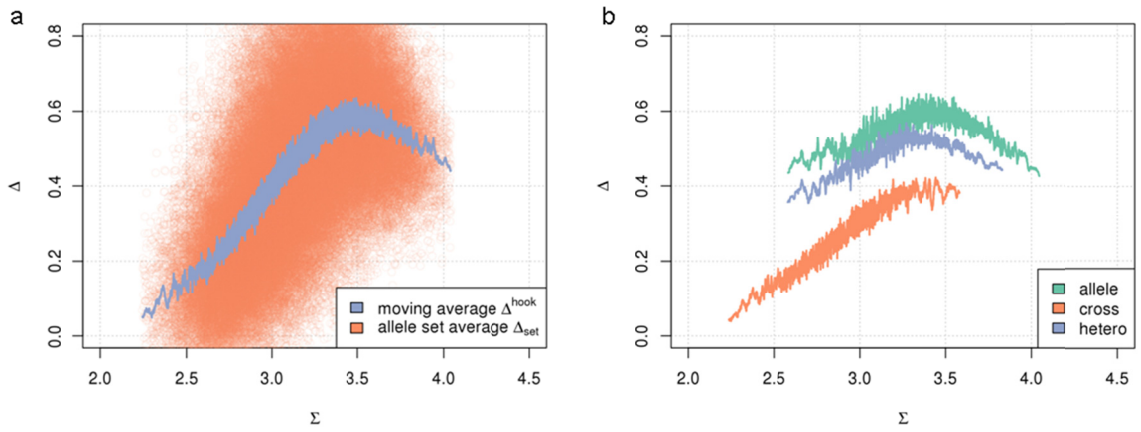

Figure 4.1: Hook plots for a Mapping50K_Xba SNP microarray. In panel a all allele sets of the SNP array have been used to compute Δ and Σ values. In panel b allele-sets have been separated for the three different binding modes: allele, cross and hetero.

In summary, Δ and Σ transformations of SNP array data show the same basic hybridization characteristics as in expression data and the differences can be well explained by the different targets and by the different probe design. The characteristic shape, the hook curve, is predicted by the two-species Langmuir model. Oligonucleotide probes on SNP arrays are therefore subject to essentially the same model of competitive surface hybridization. Consequently, the intensities of the PM and MM probes are governed by several hybridization regimes.

For a more fine-grained view of hybridization on SNP arrays let us now address allele sets with different binding modes. Affymetrix SNP arrays consider only bi-allelic SNPs, hence the genotype call for a SNP can be either homozygous allele A, homozygous allele B or heterozygous AB. Therefore each allele set can take three possible binding modes: it may bind the SNP variant which is present in the genotyped individual (*allele* binding mode), or the targeted variant is not present but instead the other variant cross-binds to the allele set (*cross*). In the third option both alleles are heterozygously present in the diploid genome (*hetero*).

Figure 4.1b shows the obtained hook plots for the same array as in Figure 4.1a. The SNPs have been called using the genotyping algorithm GTYPE implemented in the manufacturer software [58], allowing us to classify the allele-sets into one of the three binding modes. The hook curve of the 'allele' binding mode starts at high values $\Delta > 0.5$ confirming the expected substantial specific binding for the 'present' SNP variant. Δ increases to the maximum value $\Delta = 0.7$ which agrees well with the typical Δ values for perfect matched probe-target binding observed in expression arrays [42]. The probe intensities are more affected by saturation than in the other binding modes: a substantial number of probes are in the saturation range $\Sigma > 3.4$.

In the ‘cross’ binding mode the hook curve starts closely to $\Delta = 0$ and is characterized by smaller Σ and Δ values compared to the ‘allele’ binding mode. The perfectly complementary binding of PM probes to the allele is replaced by a mismatch at the SNP locus. Consequently, the PM probes in ‘cross’ allele sets exhibit a single mismatch and MM probes exhibit either two (in the case of $\delta \neq 0$, see Figure 2.4) or one ($\delta = 0$) mismatch with respect to their genomic target. Differences in the hook curves of both binding modes thus characterize the effect of incremental total mismatches in the probe-target binding.

In the ‘hetero’ binding mode both alleles are present. The hook curve is a superposition of ‘allele’ and ‘cross’ binding modes, but more resembles the latter one due to the logarithmic signal transformations in the Δ and Σ values. Only few probes are affected by saturation in the ‘allele’ and ‘hetero’ binding modes.

In conclusion, intensity values of probes referring to the ‘allele’ and ‘cross’ binding modes give rise to different hook curves because of the different mismatch configurations. Conversely, the effect of incremental mismatches can be studied using the specific design of the SNP arrays. The hook curve of the heterozygous binding mode can be understood as a superposition of ‘allele’ and ‘cross’ binding modes.

4.2 Gene ST and Exon ST arrays

Gene ST and Exon ST microarrays by design do not include mismatch probes and will thus subsequently be termed *PM-only arrays*. The key question, therefore, is how the Δ and Σ values of the hook transformation be computed for these arrays. As the average intensity of all probes in a probe set, including perfect matches and mismatches, the Σ values represent a measure of the overall expression level of the target transcript. The use of all perfect-match probes of a probe set should provide a similar measure of the expression level. The Δ values, on the other hand, represent the spread between target abundance measurements of high sensitivity and of lower sensitivity targeting the same transcript. The MM probes with their mismatch position at the center base are expected to have an about one order of magnitude decreased sensitivity compared to the PM probes. While such well-defined sensitivity differences are not given within the PM probes, significant intensity differences between probes of the same probe set nonetheless exist, for example due to variations in

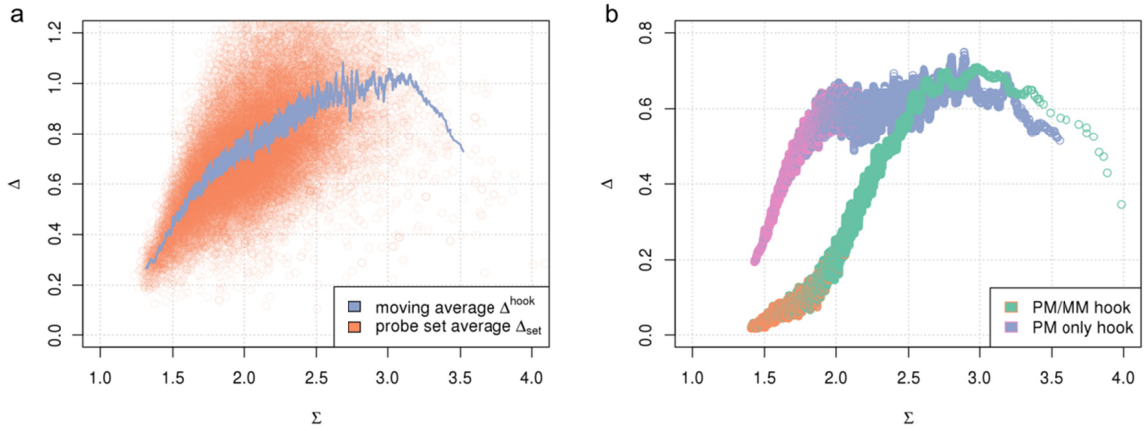

Figure 4.2: Panel a) Hook plot for a Gene ST microarray obtained by Eq. (4.1). The curve starts with the highest slope in the non-specific binding mode to the far left, then increases to its maximum at Σ = 3.0 where it decreases with the onset of saturation. Panel b) Comparison of PM-only hook (in green/orange) and PM/MM hook (in blue/magenta) for a HG-U133_Plus2 array (GSM175849). Probe sets classified as 'absent' (using hook intensity corrected signals) have been colored orange and magenta, respectively, in each curve. Both curves change their slope with the onset of specific binding and both curves decrease at Σ = 2.9 due to saturation.

duplex stability of probes with different sequences. We consequently define the following Σ–Δ-transformation for PM-only arrays

$$
\begin{aligned}
&\Delta^{\text{hook}} = \text{mvg\_avg}(\Delta_{\text{pset}}) \text{ and } \Delta_{\text{pset}} = \left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-high}} - \left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-low}} \\
&\Sigma^{\text{hook}} = \Sigma_{\text{pset}} = \left\langle \log \text{I}_{\text{p}} \right\rangle_{\text{pset}} \\
&\text{with } \left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-low}} \equiv \left\langle \log \text{I}^{*}_{\text{k}} \right\rangle_{\text{k=1..}\lceil \text{n}/3 \rceil} \quad \text{and} \quad \left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-high}} \equiv \left\langle \log \text{I}^{*}_{\text{k}} \right\rangle_{\text{k=}\lfloor \text{n}-\text{n}/3 \rfloor\text{..n}} \\
&\text{where } \text{I}^{*}_{1} \le \text{I}^{*}_{2} \le \ldots \le \text{I}^{*}_{\text{n}} \text{ and } (\text{I}^{*}_{1}, \text{I}^{*}_{2}, \ldots, \text{I}^{*}_{\text{n}}) \text{ is a permutation of } (\text{I}_{\text{p1}}, \text{I}_{\text{p2}}, \ldots, \text{I}_{\text{pn}})
\end{aligned}
\tag{4.1}
$$

where Δ values are based on the difference between the highest third, $\left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-high}}$, and the lowest third, $\left\langle \Delta_{\text{p}} \right\rangle_{\text{pset-low}}$, of the perfect-match probe intensities of each probe set.

The resulting curve shown in Figure 4.2a has characteristics that can be interpreted in analogy to the PM/MM hook. It starts with a steep slope in the interval $1.3 \le \Sigma \le 1.6$ and thereafter continues with a slower rise. After $\Sigma = 3.0$ the Δ values decrease which can be attributed to the onset of saturation where competitive binding of transcripts reduces the impact of sensitivity differences on the probe intensity.

We suspect that the change of slope at $\Sigma \approx 1.6$ relates to the onset of specific binding. To test this hypothesis we compare the PM-only hook curve to the PM/MM variant in Figure 4.2b. Both curves have been computed for a GeneChip expression array of type HG-U133_Plus2 using the PM/MM hook as defined in Eq. (3.4) and the PM-only hook as defined in Eq. (4.1). We highlighted all 'absent' probe sets as classified by the hook

method in different colors. Interestingly, both hook variants show a similar change point at $\Sigma^{break} \approx 1.9$ separating probe sets mainly governed by the N-regime ($\Sigma < 1.9$) and the onset of specific binding in the mix-regime ($\Sigma > 1.9$).

The sensitivity differences captured in the Δ values vary in the N- and S-regime, in total spanning a range of $0 < \Delta < 1.0$ in Gene ST arrays which agrees well to 3' expression arrays. While the sensitivity differences between PM and MM probes of 3' expression arrays are relatively small in the N-regime, they nonetheless increase to $\Delta = 1.0$ with increasing average expression indicated by Σ. The sensitivity differences *within* PM probes increase up to $\Delta = 0.6$ in the N-regime. In the S-regime, where Σ increments are accompanied by increments of target abundance, Δ values remain on a flat plateau (as in 3' expression arrays, Figure 4.2a) or increase only weakly (as in GeneST arrays, Figure 4.2b). The flat plateau is observed also for other PM-only microarray types as for example the Human Exon ST (data not shown).

In summary, the Δ and Σ transformations of the probe signals of PM-only arrays provide a suited approach for characterizing microarray hybridizations. Although of different shape, visual inspection allows detection of the different hybridization regimes in analogy to the PM/MM hook curve analysis. Gene ST and Exon ST show essentially the same hybridization characteristics as 3' expression arrays.

4.3 Agilent expression arrays

As discussed in Section 2.5 Agilents inking technology allows for full customization of the probe design of ordered microarrays. Predesigned arrays with typically one probe per gene (e.g. SurePrint Human Gene Expression Microarray) or one probe per exon (e.g. SurePrint Human Exon Microarray) are available. However, appropriate Σ–Δ-transformation of probe signals require multiple signals per transcript.

We here employ a custom microarray designed to quantify the expression of splice isoforms. It is the aim of the respective study to assess and compare the technical performance of microarray and high-throughput sequencing data by independently measuring the same RNA samples by both technologies. The microarray probes target 877 different genes which are known to be present in the RNA samples studied and which have a total of 5797 known splice isoforms. Also included are probes targeting 96 external RNA control transcripts from the ERCC initiative [59]. The RNA controls here refer to prepared mixes containing polyadenylated transcripts from the ERCC plasmid reference library. Each target is interrogated by several probes following a tiling design where probes query sequences at regular genomic intervals (compare Figure 2.2).

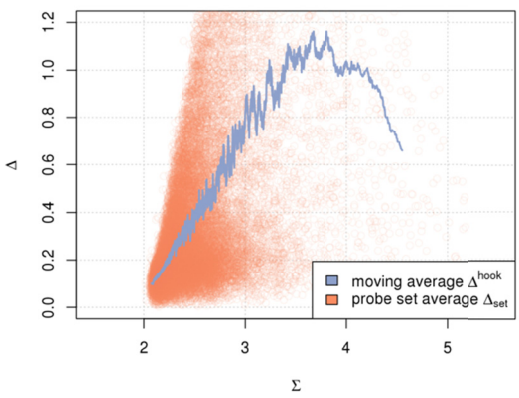

Figure 4.3: Hook plot for a custom Agilent expression microarray. Δ-Σ transformations are computed using Eq. (4.1) with 'equivalence regions' as probe sets. Note the similarity to the hook curve of SNP microarrays shown in Figure 4.1.

An appropriate definition of a probe set is required to compute Σ–Δ-transformations using the same PM-only signal transformations as for Exon ST and Gene ST arrays given in Eq. (4.1). Ideally, this probe set should include a sufficiently large number of probes of different affinity, which however target a unique set of transcripts with identical total concentration. Collecting all probes of a gene into a probe set would not be optimal: the probes spanning the entire length of gene would only give valid signals if the gene is transcribed into a single isoform. A different approach is used here where a probe set collects subsequent probes spanning DNA sections ('equivalence regions') intersected neither by known splice junctions nor known intro/exon boundaries. Only probe sets containing at least four probes are considered.

We computed the respective hook curve which is shown in Figure 4.3. It is eminent that there is only a small N-regime at the far left of the hook curve starting at $\Sigma = 2$. This can be explained by the particular design of the array and the experiment: probes have been primarily designed against genes which are expected to be present in the mixture. The curve exhibits the typical peak characteristic for the S- and sat-regime similar is other microarray types, showing that saturation is present in these Agilent expression arrays as well. The onset of saturation is around $\Sigma = 3.5$ referring to intensity values of $10^{3.5}$. Between 2 and 5 percent of the probes are affected by saturation in the 16 array samples of this experiment. This suggests that a correction of saturation effects is strongly advised to obtain unbiased expression estimates from these data.

Note that the height of the hook curve with $\Delta \approx 1.0$ is comparable with the height of the one derived from (uncorrected) probe intensities of Gene ST arrays in the previous section. These arrays use spotted oligunucleotides of different lengths: Agilent uses 60meric probes compared to the 25mers on Affymetrix arrays. The effect of sequence variations on duplex stability, and thus on intensity variability, is expected to decrease with increasing sequence length. The comparable levels of Δ values in Agilent expression data suggest that

nonetheless strong probe effects with a significant intensity variation are present here. Such high variations are however not uncommon for microarrays with a tiling design since the probes have not been previously optimized.

4.4 Summary and conclusions

We showed that appropriate Σ–Δ-transformations of probe signals can be found for other types of microarrays as well. The necessary modifications to cope with the specifics of each microarray type relate mainly to different definitions of probe sets and the lack of mismatch probes. Genome-wide SNP arrays require substitution of the 'probe set' concept, referring to a set of probes with a common mRNA transcript, with the so-called allele-set, referring to a set of probes targeting a common allele at a particular genomic location. The allele sets can be further split up to allele sets with particular binding modes, resulting in a fine-grained view of hybridization on SNP arrays. The lack of mismatch probes in Gene ST and Exon ST arrays is coped with by using existing affinity variations among the probes of each probe set. No probe sets are given by design for the presented Agilent expression array but we showed here how appropriate ones can be defined.

Using the modified Σ–Δ-transformations of probe signals we obtain characteristically shaped hook curves that can be analyzed in a similar fashion to 3' expression arrays. The differences can be well explained by the specifics of the particular microarray type and the probed targets. For example, there is essentially no N- hybridization regime in SNP chips which can be explained by the expected specific binding of at least one of the two SNP variants in the genomic DNA fragment captured by the allele set. The PM-only hook has a slightly different shape, for example in Gene ST arrays where the slope in the N-regime is higher than that of the mix regime. However, we could show that the change point between these slopes again separates the N- and mix hybridization regimes.

In summary, we showed that it is possible to apply the hook analysis to Affymetrix SNP arrays, Gene ST and Exon ST arrays and a custom Agilent expression array. This creates the possibility to study the characteristics of the different microarray technologies using the methodology presented in this thesis.

5 RNA quality effects

5.1 RNA amplification and degradation in microarray experiments

In this chapter we investigate the effect of varying RNA quality as an 'unwanted' covariate inducing potential artifacts in microarray data. Measurement of gene expression is based on the assumption that an analyzed RNA sample closely represents the amount of transcripts *in vivo*. Several effects can distort the abundance of RNA transcripts during extraction and preparation before RNA analytics using, e.g., microarrays. The first problem concerns the degradation of the RNA *in vitro* [60–63]: The quality of purified RNA is variable and after the extraction during storage rather unstable (see [64] and the references cited therein). Especially long mRNA fragments up to 10 kb are very sensitive to degradation through cleavage of RNAses introduced by handling with RNA samples. Moreover, transcripts show stability differences of up to two orders of magnitude *in vivo*, raising the possibility that partial degradation during cell lysis could cause a variable extent of bias in quantification of different transcripts [65]. The second problem concerns amplification of RNA in samples analyzed on microarrays giving rise to the decrease in the length of products that are reverse transcribed and amplified using T7 polymerase [66, 67]. The multiple rounds of *in vitro* transcription that are used to generate samples from small amounts of RNA thus induce a decrease in transcript yield and length.

The screening of nearly three thousand publicly available GeneChip array data suggests that there are noticeable degradation effects in the majority of the data files and that 2% of the files were even so severely degraded that their worth was questionable [68]. Working with low-quality RNA may strongly compromise the experimental results and lead to erroneous biological conclusions. It is therefore recommended that the highest quality RNA be used for genomic analyses. However, in some cases, such as human autopsy samples or paraffin embedded tissues, high quality RNA samples may not be available [69–71]. It is therefore important to understand how RNA quality affects microarray results and also how reliable current quality measures are at indicating RNA quality issues. The assessment of RNA integrity is a critical first step in obtaining meaningful gene expression data. A second step comprises developing methods to quantify degradation and, most importantly, to correct the induced degradation bias in the data and thereby provide more coherent expression measures.

Several RNA quality measures are established based on conventional wet lab techniques such as gel optical density measurement or denaturating agarose gel-

electrophoresis (see [61, 64] for a review). More novel lab-on-chip gel electrophoresis techniques like Agilents Bioanalyzer are now state of the art. In combination with sophisticated analysis algorithms processing the shape of the electropherogram (and, particularly, the 28S/18S rRNA ratio) they provide accepted integrity measures such as the DegFac-RQS (degradation factor RNA quality scale) [65] or the RIN (RNA integrity number) [72] which have been validated independently using qRT-PCR [64].

Importantly, microarray intensity data itself contains information about the RNA quality used for hybridization due to the 3'/5'-gradient of transcript abundance [73]. On microarrays of the GeneChip-type this gradient is typically measured using either specially-designed control probes or exploiting the specifics of the Affymetrix probe design. Both options estimate transcript abundance at close and more distant positions towards the 3'-end based on the hybridization signal [74, 75].

Although proven in many applications, these measures are based on probe intensities which, in general, are non-linear functions of transcript abundance [37, 38, 40, 54]. The signals can be strongly distorted by effects not related to transcript concentration such as saturation and non-specific background hybridization. Intensity-based RNA quality measures are therefore potentially prone to systematic errors which, in worst case, can provide diametrically opposed information in assessing apparently good RNA quality in samples with largely degraded RNA (see below). Moreover, the important task of correcting microarray signals for RNA degradation effects remained unsolved at least in single chip applications. A linear correction model requiring both expression and RNA quality data from a series of arrays has recently been published [63].

This section addresses the following tasks to overcome these problems: Firstly, we adapt non-linear hybridization theory described in Chapter 3 to the special case of truncated transcripts due to RNA degradation. We will show that our approach consistently explains previous observations such as the effect of RNA quality on transcript intensity level [63] and correlations between probe intensity and probe position along the transcripts and their effect on expression measures [76]. Analysis of the probe signals in terms of this model enables us to define unbiased (in the frame of the hybridization model used) measures of RNA integrity. Secondly, we compare these new measures with established ones. We demonstrate that methods such as *affyslope* or the RNA-integrity control probes can provide systematically false information on RNA quality. Thirdly, we propose a simple correction method which aims at removing the degradation bias from the probe intensities and which can be integrated into standard preprocessing pipelines.

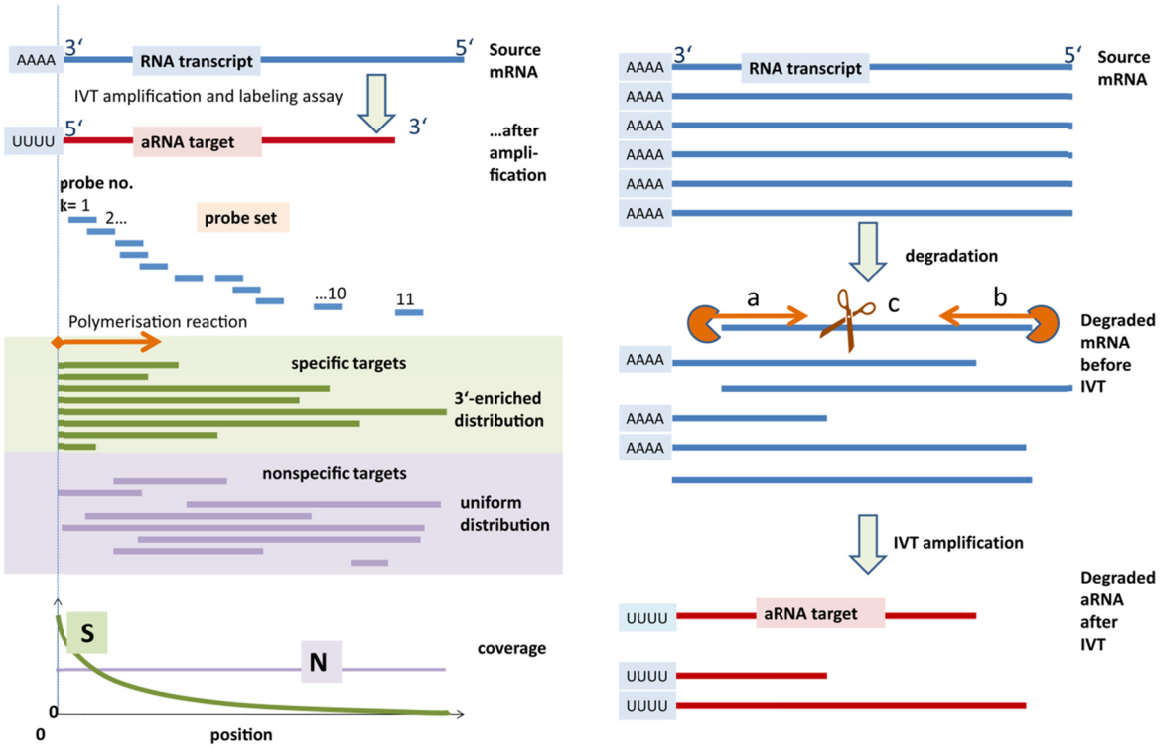

Figure 5.1: The 3'-bias of transcript abundance can be caused by in vitro transcription (left part) and degradation (right part) of source mRNA. Left part: Specific targets hybridize to the probes along the interrogated transcript with decreasing frequency due to incomplete amplification starting at the primers attached to the 3'-poly-A motif of source mRNA. In contrast, cross-hybridization of non-specific targets is not associated with the 3'-end of the transcripts giving rise to uniform coverage. Right part: Degradation of source mRNA due to RNases from both ends (a and b) and/or fragmentation at randomly chosen positions (c) also result in a 3'-enriched length distribution of amplified RNA giving rise to a similar coverage of the probes as shown in the left part. aRNA fragments are shown in 3'→5' direction (from left to right) in contrast to convention to agree with probe numbering used (k = 1, 2...) and the intensity decays introduced below.

5.1.1 3'-biased transcript coverage of microarray probes after RNA amplification and degradation

Affymetrix expression microarrays typically use a 3'-biased probe location which is motivated by the specifics of target preparation prior to hybridization (compare also Section 2.1 and 2.2). The preparation step applies IVT protocols according to the Eberwine method [12]. It starts with first-strand cDNA synthesis from source mRNA using T7 oligo(dA) primers followed by second strand cDNA synthesis [66, 67]. The double-stranded cDNA fragments are subsequently transcribed into amplified antisense RNA (aRNA) which, after labeling, is hybridized on the arrays.

First-strand cDNA polymerization is primed at the 3'-end of mRNA and proceeds towards the 5'-end (see Figure 5.1). Due to incomplete polymerization this method produces truncated transcripts of variable length which are however characterized by a common 3'-start site with respect to the respective fragment of source mRNA [73] (to avoid confusion

we will strictly refer to the 3'- and 5'-ends of the source mRNA and not to that of the product aRNA). In consequence, the resulting distribution of transcript lengths gives rise to a 3'-enriched, decaying towards the 5'-end coverage of the probes of the probe sets interrogating the respective transcript with increasing probe index (for convenience we will count the probes in direction towards the 5'-end in contrast to Affymetrix counting the probes in the opposite direction). Subsequent fragmentation of these aRNA targets into pieces of typically between 30 and 200 nt in length before hybridization leaves the 3'-bias of probe coverage unaffected.

Importantly, the decaying coverage of the probes is expected to apply to specific but not to non-specific hybridization. In the N-hybridization mode the probes bind aRNA fragments of partly complementary sequence originating however from mRNA transcripts not referring to the interrogated gene. Trivially, these non-specific transcripts lack a common start position with respect to the intended target and, as a consequence, they, on the average, uniformly cover the probes of each probe set (see Figure 5.1a). Specific hybridization competes with non-specific one and both hybridization modes contribute to the measured probe intensities. The consequences of different probe coverages for the measured signal will be discussed below.

Also degradation of mRNA, e.g. upon storage, can produce 3'-biased probe coverages of fragmented aRNA by endonuclease activity that cuts RNA internally, or by means of exonucleases [77]. In the first case, the poly(A) tail is removed by a deadenylase activity, followed by two mechanisms that degrade the mRNA: either decapping followed by a 5'-to-3' decay or a 3'-to-5' decay. Once the mRNA poly(A) tail is removed, reverse transcription reaction will not proceed, resulting in low concentrations of truncated transcripts (see Figure 5.1b). Several studies have identified RNA degradation to be a major cause of microarray expression measure variability [63, 65, 68–71].

5.1.2 Probing transcript abundance using GeneChip arrays

In this section we investigate the details of design and annotation of the probes of 3' expression arrays. Affymetrix constructs their probe sets by selecting the probes from a longer target sequence according to various optimization criteria. The original sequences used at design time are of one of three types: consensus, exemplar and control sequences. According to Affymetrix, "A consensus sequence results from base-calling algorithms that align and combine sequence data into groups. An exemplar sequence is a representative cDNA sequence for each gene" [78]. Each probe set refers to one and only one of these sequences. For each 3' expression array, we have downloaded the consensus, exemplar and

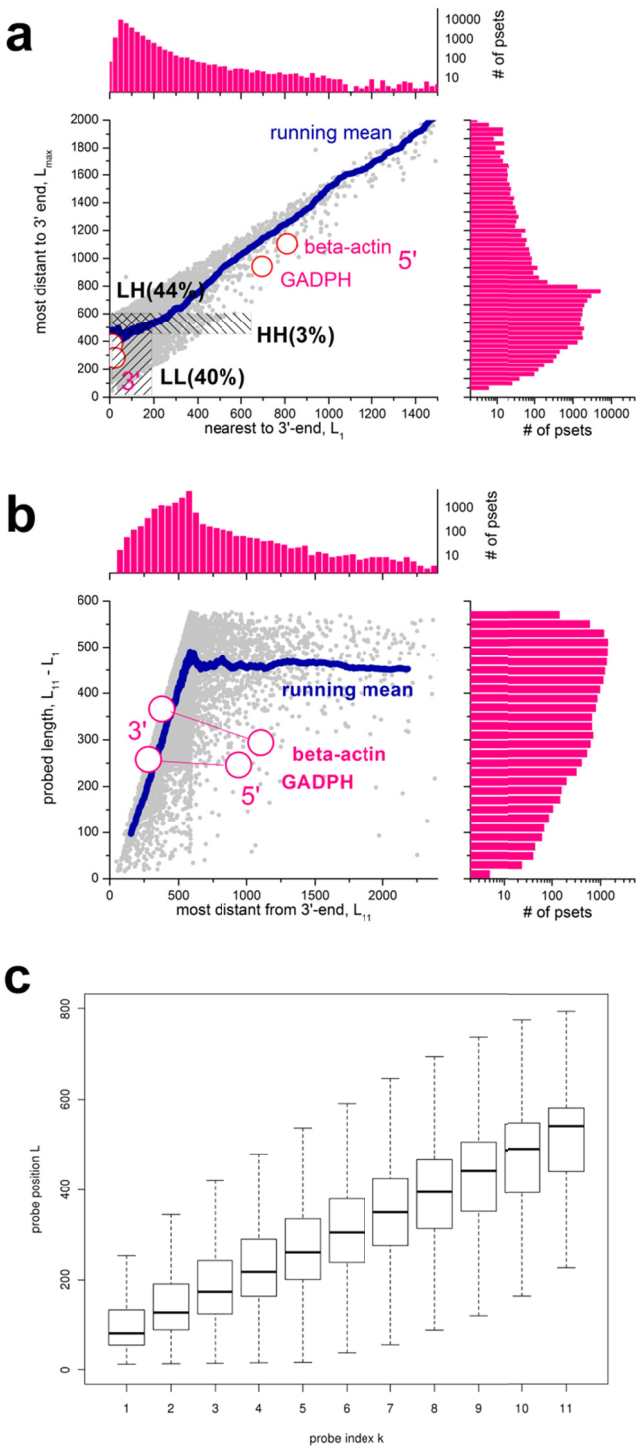

Figure 5.2: Probe and probe set characteristics of the RAE230 GeneChip array: Panel a correlates the position of the 11th (nearest the 5' end of the transcripts) and of the 1st (nearest the 3' end) probe of each probe set and shows the respective number distributions. Most probe sets accumulate in the LH (low L_1, high L_{11}) and LL ranges whereas only a few sets are found in the HH range. Panel b shows the coverage size of the probe sets ($\Delta L = L_{11} - L_1$) as a function of the position of the 11th probe set together with the respective number distributions. The mean ΔL value nearly linearly increases until $k_{11} \approx 600$ and then it remains virtually constant with $<\Delta L> \approx 460$. The most probe sets cover a transcript range of 400 – 550 nucleotides. The open circles refer to the 3'- and 5'-control probe sets. The boxplot in panel c correlates the probe index k with the probe position L. The median position per index (see the horizontal bar in each box) nearly linearly increases with k. The slope provides the ΔL-value of the array (~ 50 nucleotide positions per index increment).

control sequences together with the probe sequences as provided by Affymetrix (see www.affymetrix.com). Probe distances to the intended 3'-end of the transcript, L_p, were computed by aligning the probe sequences to the respective transcript sequences. The position of each probe L_p (p = 1, 2...) is then defined as the number of nucleotides counted between the 3'-end of the transcript and the first (i.e. nearest) base of the 25meric probe sequence[3]. The ordering of probes according to increasing distances L_p defines the probe index k within each probe set.

The probes of each probe set cover transcript lengths which largely exceed the length of the individual probes. This design is well suited to study length-dependent alterations of transcript abundance due to RNA degradation and imperfect amplification. Figure 5.2a shows that the majority of probe sets start (first probe with index k = 1) within the first $L_1 = 100$ - 200 nucleotides nearest to their 3'-end and end at position $L_{11} = 250$ - 600 for the last probe (index k = 11). Only about 5% of all probe sets are located beyond the range of 600 nucleotides. Within this range, the sets can be roughly classified into 'low (i.e., more 3') L_1 and low L_{11}' (LL), 'low L_1 and high (i.e., more 5') L_{11}' (LH) and 'high L_1 and high L_{11}' (HH) sets where low refers to distances close to the 3' end and high refers to distances farther towards the 5' end (see Figure 5.2a). The mean length of the covered transcript range ($\Delta L = L_{11} - L_1$) nearly linearly increases with the position of the 11th probe up to $L_{11} \approx 600$, and then it remains virtually constant $\Delta L \approx 460$ (Figure 5.2b). Hence, short probe sets with $\Delta L < 300$ accumulate near the 3' end of the transcripts whereas more distant probe sets typically cover a wider length range of the transcripts ($350 < \Delta L < 600$).

The mean position of all probes on the array with a given index k = 1...11 linearly correlates with k to a good approximation (Figure 5.2c). The obtained slope characterizes the mean distance between two neighbored probes. It can be interpreted as the probe sensitivity per index increment and depends on the probe design of the particular array type,

$$\langle \Delta L \rangle = \frac{\langle L \rangle_{array} - \langle L_1 \rangle_{array}}{\langle k \rangle_{array} - 1} \approx \frac{\langle L \rangle_{array}}{\langle k \rangle_{array}} \qquad (5.1)$$

$\langle ... \rangle_{array}$ denotes averaging over all probes of the array. The approximation in the right part assumes a vanishing intercept in good agreement with the data (see Figure 5.2c).

[3] Precomputed probe distances for most GeneChip microarrays are available on our website http://www.izbi.uni-leipzig.de/downloads_ links/programs/rna_integrity.php

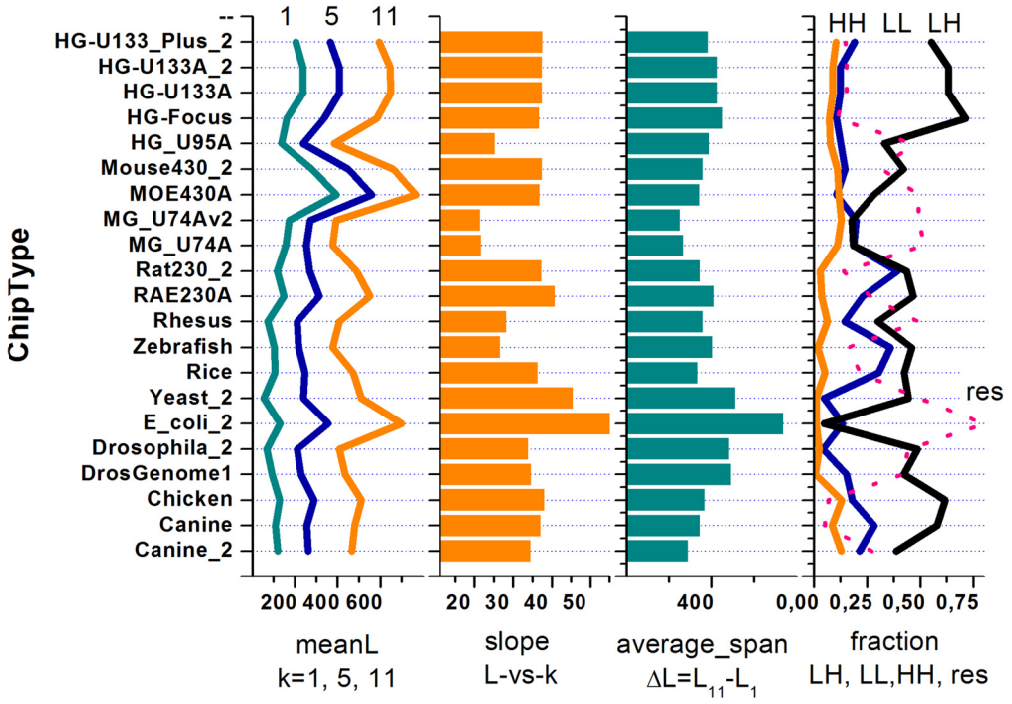

Figure 5.3: Probe and probe set characteristics for different GeneChip microarrays: the mean positions of the 1th (nearest to the 3' end of the transcript), 5th and 11th (nearest to the 5' end) probe averaged over all probe sets; the slope of the linear regression of the mean probe position L versus the respective probe index for all probes (<ΔL>); the average transcript range covered by the probe sets (average span) and the fraction of probe sets from the LH, LL and HH ranges and the residual fraction not contained in one of the three ranges (see the main paper for definitions).

Figure 5.3 provides an overview over selected probe design characteristics of different GeneChip types. It shows that the mean position of the first and of the last probe in the probe sets can strongly vary between the different chip types giving rise to a wide range of <ΔL>-values which can change between about 25 and about 60 nucleotides per index increment. These differences refer in first instance to arrays of older and newer generations (e.g., the human genome HGU95a and HG133a arrays and the mouse genome MG74a and MOE430a arrays, respectively). On the other hand, the average span covered by the probe sets is relatively constant for all chip types considered.

Affymetrix GeneChip arrays include a small number of control probe sets designed to estimate the RNA quality in terms of the 3'/5' bias. They target the 3' end, the 5'-end and the middle (m) of relatively long transcripts coding, e.g., beta-actin and glyceraldehyde-3-phosphate dehydrogenase (GADPH) using 20 probes per set. Figure 5.2a and b shows that the 3'- and the 5' probe sets of the controls together cover the range of about 700 nucleotides between $L_{20} = 281$ and $L_{20} = 378$ for the 3'-probe sets and $L_{20} = 942$ and $L_{20} = 1104$ for 5'-probe sets of GADPD and beta-actin, respectively.

5.1.3 Used expression data

Affymetrix microarray raw intensity data (CEL-file format) were downloaded from the public repositories Gene expression omnibus [79] or Array Express [80]. In this section we study the following data sets.

The *Human tissue* dataset (GSE7307, see supplementary text for the detailed list of samples used) comprises 677 samples taken from over 90 distinct tissue types hybridized to Affymetrix HG-U133 plus 2.0 arrays.

The *RatQC* (rat quality control) dataset (E-MEXP-1069) from [69] was generated to systemically explore how RNA quality affects microarray results. It consists of 36 rat liver RNA samples hybridized to Affymetrix RAE230A expression arrays. The progressive change in RNA quality was generated either by thawing frozen tissue or by *ex vivo* incubation of fresh tissue. Each sample was characterized by the RNA integrity number (RIN) and mean transcript length in [69].

The *RNeasy* data set consists of five pairs of HG-U133A GeneChips which were hybridized with RNA extracted from ovarian cancer samples and processed in two different ways namely with and without a cleanup step using RNeasy reagents [62]. The RNeasy cleanup should lead to good-quality RNA whereas lack of the cleanup step should yield poorer-quality RNA. The RNeasy data set was used in previous work aiming at judging RNA-quality from microarray data [81, 82].

5.2 Degradation and hybridization mode

5.2.1 Intensity-based degradation metrics

In this section we discuss the consequences of the 3'-enriched probe coverage on the observed probe intensities. In the following we will subsume the 3'-bias of probe coverage as 'degradation effect' independent of its origin (IVT amplification or degradation) for the sake of convenience. Let us first define the probe-specific and the mean degradation ratio averaged over all probes of the array,

$$d_{g,k} \equiv \frac{[S_{g,k}]}{[S_g]_{\text{target}}} \quad \text{and} \quad d = \left\langle d_{g,k} \right\rangle_{\text{all probes,all genes}} \tag{5.2}$$

respectively, which characterize the decrease of the transcript concentration due to the degradation effect. $[S_g]_{\text{target}}$ is the (true) expression degree of a selected gene g given as the total concentration of the target transcripts in the hybridization solution independent of

their length. It refers to the target concentration in the absence of degradation and presumes that RNA processing proceeds without 3'/5' bias. Contrarily, $[S]_{g,k}$ denotes the apparent expression degree reported by probe with index $k = 1,...,N_{set}$ designed to interrogate target g. It is given as the concentration of the RNA fragments which specifically bind to this probe. It consequently refers to the probe coverage which decays with increasing distance of the probe to the 3'-end of the target. Angular brackets $<\ldots>_{\text{all probes}}$ denote averaging over all probes of the array. One expects $[S]_{g,k} \le [S_g]_{\text{target}}$ and thus $d_{g,k} \le 1$ owing to the 3'-enrichment after incomplete amplification and degradation of the fragments. The probe specific degradation index, $d_{g,k}$, thus characterizes the loss of mRNA material at a given probe position along the transcribed region of the gene. The mean degradation index d averages the single probe effects over all probes. It estimates the total loss of RNA probed by the microarray in a given preparation prior to hybridization.

The probe intensities measured in the microarray experiment are described by the Langmuir model in Eq. (3.2) where the probe index $p \equiv g,k$ subsumes the gene and probe index explicitly used in Eq. (5.2). We here consider the reduction of the concentrations $[S_g]_{\text{target}}$ and $[N]_{\text{chip}}$ after incomplete amplification and/or degradation as an effect of the binding strengths due to specific and non-specific hybridization (compare Eq. (3.3))

$$X_p^{P,S} = \left[S_p\right] \cdot K_p^{P,S} = d_p \cdot \left[S_g\right]_{\text{target}} \cdot K_p^{P,S} \quad \text{and} \quad X_p^{P,N} = d \cdot \left[N\right]_{\text{chip}} \cdot K_p^{P,N} \tag{5.3}$$

respectively. Non-specific hybridization is related to the total amount of RNA used for hybridization [57]. $[N]_{\text{chip}}$ is consequently reduced by a factor given by the mean degradation factor d.

The probe-specific degradation index d_p defines the decrease of transcript concentration after amplification and degradation (Eq. (5.2)). In the next step we define the apparent degradation index as the intensity ratio of probes located at different positions along the target sequence, for example near its 5'- and 3'-end of one selected target,

$$r_{5'/3'}^{app} \equiv \frac{I_{5'}^{P}}{I_{3'}^{P}} \tag{5.4}$$

where the intensities are given by Eqs. (3.2) and (5.3) with the respective degradation ratios $d_{5'}$ and $d_{3'}$, respectively.

Let us consider two special cases if the probes hybridize either far from saturation in the linear range ($X_p^{P,S}, X_p^{P,N} << 1$) or in the range of saturation of specific hybridization ($X_p^{P,S} > 1 > X_p^{P,N}$). The apparent degradation index becomes

$$r_{lin}^{app} = \frac{\left(d_{5'} \cdot x_{5'}^{S} + d \cdot x_{5'}^{N}\right)}{\left(d_{3'} \cdot x_{3'}^{S} + d \cdot x_{3'}^{N}\right)} = \begin{cases} r_{5'/3'}^{true} \cdot (K_{5'}^{P,S} \cdot w_{5'}^{P,S} / K_{3'}^{P,S} \cdot w_{3'}^{P,S}) & \text{for} \quad x^S >> x^N \quad \text{(specific)} \\ (K_{5'}^{P,N} \cdot w_{5'}^{P,N} / K_{3'}^{P,N} \cdot w_{3'}^{P,N}) & \text{for} \quad x^S << x^N \quad \text{(non-specific)} \end{cases}$$

$$\text{and} \quad r_{sat}^{app} = \frac{w_{5'}^{P,S}}{w_{3'}^{P,S}} \qquad \text{(saturation)} \tag{5.5}$$

respectively. The lower case x defines the hybridization strengths at 'ideal' transcript concentrations (see Eq. (5.3) with $d = d_p = 1$: $x_p^S \equiv [S]_{target} \cdot K_p^{P,S} \cdot w_p^{P,S}$ and $x_p^N \equiv [N]_{chip} \cdot K_p^{P,N} \cdot w_p^{P,N}$) and $r_{5'/3'}^{true} \equiv d_{5'} / d_{3'}$ denotes the 'true' relative degradation index between 5' and 3' probes, respectively. Eq. (5.5) shows that the apparent degradation index is proportional to the true one ($r_{lin}^{app} \propto r_{5'/3'}^{true}$) in the special situation of dominating specific hybridization ($x^S \gg x^N$) far from saturation only. It however scales with the ratio of the specific binding and washing constants of the 3'- and 5'-probes, which might be larger or smaller than unity depending on the sequences of the particular probes (see [43] for details). At dominating non-specific binding or saturation one gets apparent degradation indices which are completely independent of the true one. Their values again depend on the probe sequences and can be larger or smaller than unity. Hence, the use of intensity-based degradation metrics raises problems because they reflect the degradation bias of transcript abundance in special situations only.

On the other hand, two intensity-based degradation measures are well established for quality control of GeneChip arrays: (i) The slope of a linear function fitted to the so-called 'RNA degradation plot', $r_{5'/3'}^{slope}$. This RNA degradation plot displays the mean logged intensity averaged over all probes with the same index k, taken from one array, as a function of k [75]. (ii) The intensity ratio $r_{5'/3'}^{control}$ of special control probe sets targeting the 5'- and the 3'-end of relatively long transcripts such as beta-actin and GADPH. A threshold of the 3'/5'-signal intensity ratio of the GADPH controls less than 3 (in logarithmic scale $\log_{10} 3 = 0.48$) is recommended for good quality RNA [83, 84].

In view of the discussed problems of intensity-based degradation measures we will revise theses estimates and judge their suitability for determining RNA quality. Large values of $r_{5'/3'}^{slope}$ and/or $r_{5'/3'}^{control}$ near unity are generally thought to indicate small degradation bias and thus good RNA quality. Note that reciprocal values of these measures are often used in practice estimating the respective 3'/5'-ratios. Here we consequently use 5'/3'-ratios to ensure direct comparability between the various measures.

In summary, probes located nearer to the 3'-end of the interrogated transcripts potentially shine brighter than more distant probes due to the 3'-enrichment of probe coverage giving rise to expected 'true' intensity ratios $r_{5'/3'} < 1$. However, this rule applies only to conditions of specific hybridization far from saturation. RNA quality measures based on

the 5'/3'-intensity ratio consequently require consideration and evaluation of the hybridization mode of the chosen probes. Moreover, the potential dependence of the probe intensities on the degree of degradation gives rise to systematic errors of the estimated expression degree of the transcripts which requires appropriate correction.

5.2.2 Degradation Hook and Tongs Plot

In analogy to the Δ and Σ transformations given in Eq. (3.4) we define the following modified hook representations

$$\begin{aligned} &\Delta\Sigma_{s3'/s5'} \equiv \left\langle \Sigma_p \right\rangle_{s3'} - \left\langle \Sigma_p \right\rangle_{s5'} \quad \text{(degradation hook)} \\ &\Delta\Sigma_s \equiv \left\langle \Sigma_p \right\rangle_s - \left\langle \Sigma_p \right\rangle_{pset} \qquad \text{(tongs plot)} \\ &\text{with} \quad \left\langle \Sigma_p \right\rangle_s \equiv \frac{1}{3}\sum_{k=i}^{i+2} \Sigma_k \quad \text{and} \quad \Sigma_k \equiv \tfrac{1}{2}\left(\log I_k^{PM} + \log I_k^{MM} \right) \end{aligned} \tag{5.6}$$

where the subscript s = s3', s5' denotes a subset of three consecutive probes within the probe set of size N_{pset} nearest to (s3', i = 1) or most distant from (s5', i = N_{pset} - 2) the 3'-end of the transcript, or centered around its middle probe (s = m). The so-called tongs-plot shows the three positional-dependent values $\Delta\Sigma_{3'}$, $\Delta\Sigma_{5'}$ and optionally $\Delta\Sigma_m$ as a function of Σ whereas the 'degradation hook' plots $\Delta\Sigma_{3'/5'}$-versus- Σ. These plots use the same abscissa as the hook curve and they also smooth the noisy data using a running window of 500 - 1000 probes. Both the 'degradation hook' and the 'tongs plot' estimate the 3'-enrichment of the probes and thus their degradation level in dependence on the hybridization mode. Examples for both plots are shown in Figure 5.5 in the next section.

The two-species Langmuir hybridization isotherm predicts the theoretical hook-curve which was previously used to fit the experimental curves and to extract characteristic chip-related parameters. Here we modify the hook formalism to take into account the effect of incomplete transcript amplification and degradation in terms of the degradation ratios defined in Eq. (5.2). We thus define the probe-specific S/N ratio similar to Eq. (3.11) under consideration of the subset s of probes

$$R_s = \frac{\left\langle d_p \right\rangle_s}{d} \cdot \frac{X^{PM,S}}{X^{PM,N}} = \frac{\left\langle d_p \right\rangle_s}{d} \cdot \frac{[S]_{target}}{[N]_{chip}} \cdot \frac{\left\langle K_p^{PM,S} \right\rangle_{chip}}{\left\langle K_p^{PM,N} \right\rangle_{chip}} \tag{5.7}$$

It scales with $\left\langle d_p \right\rangle_s / d$, the probe specific 3'-bias of the actual transcript abundance averaged over the subset s and divided by the mean degradation index of the selected chip, d. Similarly to Eq. (3.13), the theoretical expressions for the hook coordinates for the

subset s of probes are obtained by inserting Eqs. (3.2) and (5.3) with P = PM and MM into Eq. (5.6)

$$\begin{aligned}&\left\langle \Delta_p \right\rangle_s \equiv \Delta(R_S) = \log\left\{(R_S+1)/(R_S\cdot 10^{-\alpha}+1)\right\} - \log\left\{B^{PM}(R_S)/B^{MM}(R_S)\right\}\\ &\text{and}\\ &\left\langle \Sigma_p \right\rangle_s \equiv \Sigma(R_S)\\ &\quad = \Sigma^{start} + \frac{1}{2}\log\left\{(R_S+1)/(R_S\cdot 10^{-\alpha}+1)\right\} - \frac{1}{2}\log\left\{B^{PM}(R_S)/B^{MM}(R_S)\right\}\end{aligned} \tag{5.8}$$

with the saturation terms

$$\begin{aligned}&B^{PM}(R_S) = 1 + 10^{-(\beta+\frac{1}{2}\Delta_{start})}\cdot(R_S+1) \quad \text{and}\\ &B^{MM}(R_S) = 1 + 10^{-(\beta-\frac{1}{2}\Delta_{start})}\cdot(R_S\cdot 10^{-\alpha}+1)\end{aligned}$$

The vertical and horizontal dimensions of the hook curve and its start coordinates are given as

$$\begin{aligned}&\alpha \approx \log\frac{\left\langle K_p^{PM,S}\right\rangle_{chip}}{\left\langle K_p^{MM,S}\right\rangle_{chip}} \quad , \quad \beta \approx -\left(\log d + \left\langle \log X_p^{PM,N}\right\rangle_{chip}\right) \quad \text{and}\\ &\Sigma^{start} = \log M + \left\langle \log X_p^{PM,N}\right\rangle_{chip} + \log d\end{aligned} \tag{5.9}$$

respectively. Note that the width and the start coordinate of the hook curve, β and Σ_{start}, change with the mean degradation index d whereas the height of the hook α doesn't depend on degradation.

The mean expression index characterizes the mean expression level of present probes of the chip,

$$\varphi \equiv \left\langle \log(d_p^{PM,S}\cdot X_p^{PM,S})\right\rangle_{chip} \approx \left\langle \log(R) + \log X_p^{PM,N} + \log d_p^{PM,S}\right\rangle_{chip} \tag{5.10}$$

The ordinate values of the degradation plots are obtained by inserting Eq. (5.8) into Eq. (5.6),

$$\Delta\Sigma_s(R) \equiv \Sigma(R_s) - \Sigma(R) \quad \text{and} \quad \Delta\Sigma_{s1/s2}(R) \equiv \Sigma(R_{s1}) - \Sigma(R_{s2}) \tag{5.11}$$

One gets after explicit consideration of Eq. (5.8)

$$\Delta\Sigma_{s1/s2}(R) = \tfrac{1}{2}\log\frac{(R_{s1}+1)\cdot\left(R_{s1}\cdot 10^{-\alpha}+1\right)}{(R_{s2}+1)\cdot\left(R_{s2}\cdot 10^{-\alpha}+1\right)} - \tfrac{1}{2}\log\frac{B^{PM}(R_{s1})\cdot B^{MM}(R_{s1})}{B^{PM}(R_{s2})\cdot B_p^{MM}(R_{s2})} \tag{5.12}$$

where $\Delta\Sigma_{s1}(R)$ refers to the special case $r_{s2} = 1$. The parameters

$$\begin{aligned}\gamma_s &\equiv \log r_s = \log\frac{\langle[S]\rangle_s}{\langle[S]\rangle_{chip}} \quad \text{with} \quad s = 3', 5', m \quad \text{and}\\ \Delta\gamma_{3'/5'} &\equiv \gamma_{3'} - \gamma_{5'} = \log\frac{\langle[S]\rangle_{3'}}{\langle[S]\rangle_{5'}}\end{aligned} \tag{5.13}$$

define the 3'-bias of transcript abundance (see also Eqs. (5.5) and (5.7)). Particularly, $\Delta\gamma_{3'/5'}$ provides the logged fold change of the probe specific transcript concentrations between probes located nearer the 3'- and 5'-ends of the transcript. The mean transcript concentration averaged over all probes can be estimated as the geometric mean over the 3' and 5' transcript concentrations,

$$\langle S\rangle_{chip} \approx \langle[S]\rangle_{3'} 10^{-\frac{1}{2}\Delta\gamma_{3'/5'}} \approx [S]_{3'}\cdot\sqrt{\frac{[S]_{5'}}{[S]_{3'}}} \approx \sqrt{[S]_{3'}\cdot[S]_{5'}} \tag{5.14}$$

if one assumes uniformly distributed probes along the relevant transcript regions. With $[S]_{target} \approx \langle[S]\rangle_{3'}$ and Eq. (5.2) one gets

$$\log d \approx -0.5\cdot\Delta\gamma_{3'/5'} \tag{5.15}$$

Hence, the mean amount of RNA (Eq. (5.2)) is directly related to the 3'/5'-difference of transcript abundance.

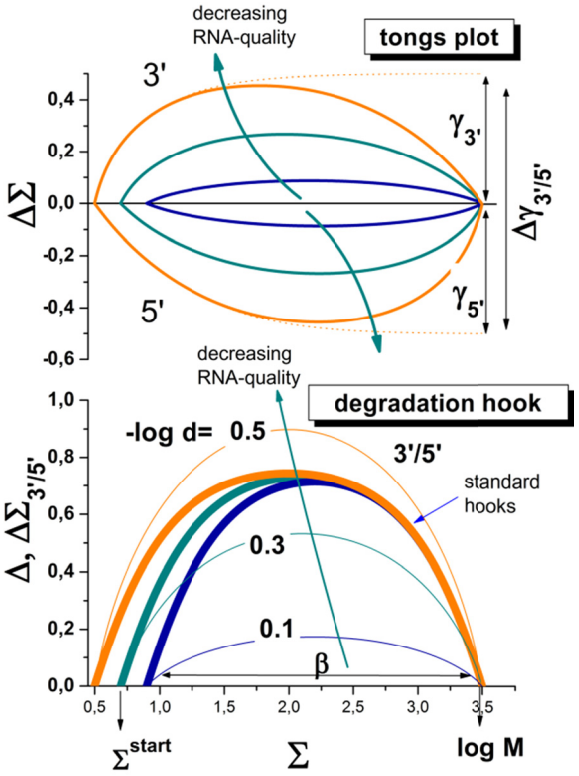

Figure 5.4: Theoretical hook curve (Eq. (5.8), thick curves), degradation hook (thin curves) and tongs plot (panel above; Eqs. (5.11) and (5.12)) for different degradation levels log d. With increasing degradation the positive and negative amplitudes of the tongs plot (the tongs opening $\Delta\gamma_{3'/5'}$) and the height of the degradation hook increase, accompanied by the shift of its increasing branch towards the left which widens the curves (parameter β). The curves are calculated with $\gamma_{3'} = -\gamma_{5'} = 0.1$, 0.3 and 0.5, respectively. The dotted curves in the part above are calculated neglecting the saturation term in Eq. (5.12). The geometrical meaning of selected parameters is indicated by arrows (see text).

Typical examples of the hook curve (Δ-versus-Σ, Eq. (5.8), thick curves), the degradation hook ($\Delta\Sigma_{3'/5'}$-versus-Σ, Eq. (5.12), thin curves) and the tongs-plot (($\Delta\Sigma$-versus-Σ,, Eq. (5.12), panel above) as predicted by theory are shown in Figure 5.4 for different degradation levels. Increasing degradation increases the opening of the tongs and widens the hook. Both changes are governed by the degradation ratio d and r_s and their logarithmic transformations (see Eqs. (5.2), (5.9), (5.13) and (5.15)). The widening of the hook by log d reflects the decrease of the mean transcript concentration due to incomplete amplification and degradation. This trend is equivalent with the decrease of the mean level of non-specific background hybridization which in turn increases the mean binding constant of specific binding [57]. The consequences of this so-called up-down effect are discussed above.

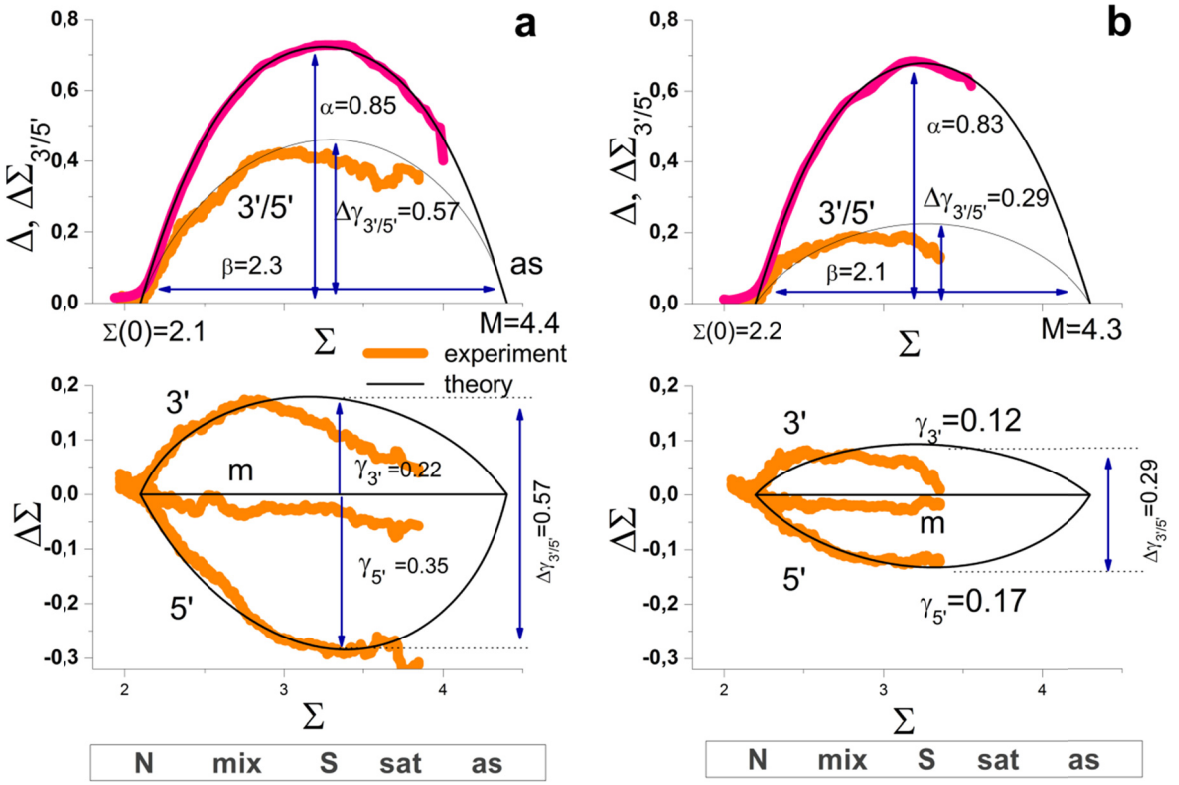

Figure 5.5: Hook- and degradation hook (above) and tongs-plot (below) of two selected chip hybridization taken form the human body index data set (muscle, GEO accession numbers GSM176301 in panel a and skin, GSM175967 in panel b) referring to large and smaller degradation effects, respectively. Note that all plots use the same abscissa scaling (Σ, see Eq.(5.6)) which is related to the expression degree of the respective probes. The hook curve reveals the changing hybridization mode with increasing sigma: non-specific (N), mixed N and S (mix), specific (S), saturation (sat) and asymptotic (as) ranges. The degradation hook and the tongs-plot reveal the mean 3'/5'-intensity bias of the probes. The three branches of the tongs plot refer to three probes nearest to the 3'-end (upper branch), nearest to the 5'-end (lower branch) and located in the middle in-between (middle branch). Note that the different branches split maximally in the S-range of hybridization whereas no bias is observed in the N-range as predicted by theory (lines, see Eqs. (5.8) and (5.12) for the hook and tongs plot, respectively). The theoretical curves are calculated using the formulae given in the previous section using the parameters given in the figure. The hook dimensions (α, 'height' of the hook, see Eq. (5.9); β, 'width' of the hook; Σ(0), 'start' point; M, 'end'-point) are very similar for both arrays whereas the logarithmic 3'- and 5'-degradation levels (Eq. (5.15)) are markedly different. The size of the moving window is decreased towards the right end of the tongs plot to compensate the reduced number of probe sets in saturation range. As a consequence, the part of the curves beyond of the maximum is prone to increasing error.

5.2.3 The 3'-intensity bias depends on the hybridization mode

Figure 5.5 shows the hook curves for two selected microarray hybridizations of differently degraded RNA together with the respective degradation hook (panel above) and tongs plot (panel below). The degradation hook shows essentially the same shape as the standard hook. The curves reflect however different effects: The standard hook plots the mean logged intensity difference between paired PM and MM probes. It consequently estimates the intensity penalty of one mismatched base pairing in the respective probe/target-duplexes. Contrarily, the degradation hook judges the logged mean intensity difference

between probes located nearer and farther to the 3'end of the transcripts, and thus the 3'/5'-bias of the probe intensities in the probe sets due to the degradation effect.

Interestingly, the different hybridization modes analogously affect the intensity differences in the standard and the degradation hook as well. For example, upon non-specific hybridization both, the PM/MM difference and the 3'-bias essentially disappear because both effects, the MM-penalty and the 3'/5'-bias, require duplexing of the probes with the intended targets. Non-specific binding doesn't meet this criterion because the binding of non-specific transcripts is indifferent with respect to the mismatched pairing of the middle base of the MM probes and with respect to the degradation bias as well. Vice versa, both hook-versions show their maximum in the S-range because specific binding is associated with the intended intensity penalty of the MM-probes and of probes located more distant from the 3'-end, respectively.

Note that the two standard hook plots shown in panel a and b of Figure 5.5 are of virtually equal height owing to the similar MM-penalty ($\alpha = 0.83 - 0.85$) whereas the respective degradation hooks markedly differ in this respect ($\Delta\gamma_{3'/5'} = 0.57$ and 0.29, respectively) revealing marked differences in the degradation level between both samples. Comparison of the heights of both hook-types shows that strong degradation can affect the probe intensities nearly by the same order of magnitude as one mismatched base pairing.

The tongs plots explicitly estimate the intensity bias at three positions of the probe sets and thus it illustrates the progression of degradation with increasing probe index. The $\Delta\Sigma_s$ curves of all three subsets (s = 3', 5' and m) degenerate in the N-hybridzation range indicating the absence of the 3'-bias for non-specific binding as discussed above (see also Eq. (5.5) for $x^S \ll x^N$). In the mix-range the $\Delta\Sigma_s$-curves split into three branches which progressively diverge with increasing sigma and thus with increasing contribution of specific hybridization. The 'opening of the tongs', i.e. the split between the 3'- and 5'-branches, reaches its maximum in the S-range of hybridization in parallel with the maximum of the hook curve and of the degradation hook. Subsequently, the different branches start to converge as predicted for the range of saturation (see Eq. (5.5)). Both, the experimental degradation hook and the tongs plot are well described by theoretical curves based on the Langmuir-model of array hybridization in Eq. (5.12). The split parameter $\Delta\gamma_{3'/5'}$ characterizes the height of the degradation hook, or equivalently, the 'tongs opening' serving as a measure of the maximum vertical difference between the 5'- and the 3'-branches of the tongs, respectively. $\Delta\gamma_{3'/5'}$ estimates the 5'-depletion of probe coverage in terms of the logged concentration increment between the targets covering the 5'- and 3'-probes (Eq. (5.15)). The examples shown in Figure 5.5 a and b refer to relatively strong and weak depletion of targets with 5'/3'-concentration ratios of $d_{tongs} = 10^{-0.57} = 0.27$ and

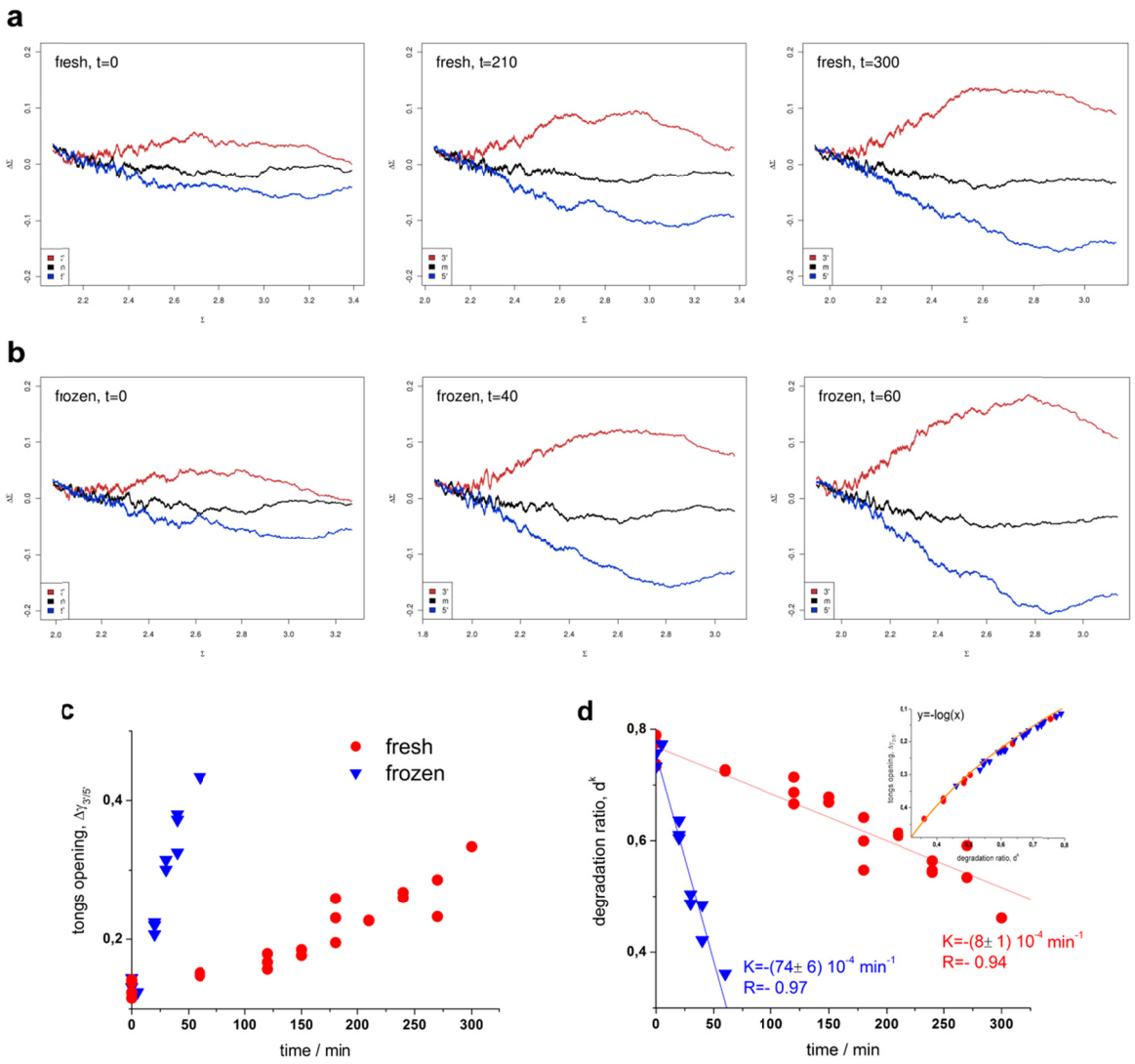

Figure 5.6: Collection of tongs plots taken from the ratQC data set. The RNA was extracted from liver samples either after ex vivo incubation of fresh tissue (panel a, incubation time 0, 210 and 300 min) or after thawing frozen tissue (b, incubation time 0, 40 and 60 min). The plots in panel c and d show the tongs-opening and the d^k parameter of both series as a function of the incubation time, respectively. RNA prepared from frozen samples degrades much faster than RNA from fresh samples. The insert in part b correlates the d^k and tongs opening parameters. Their relation follows a logarithmic function.

$10^{-0.29} = 0.51$, respectively (see Eq. (5.15)). This analysis shows that degradation can reduce the transcript concentration to less than one third of the initial transcript abundance.

Figure 5.6 shows a collection of tongs plots taken from the *RatQC* dataset [69] characterizing the level of degradation of rat liver RNA under two conditions, namely after incubation of fresh tissue (panel a) or after thawing frozen tissue (b). With incubation time the opening of the tongs increases indicating progressive degradation of the RNA. The time dependence reveals that RNA degradation in thawed tissue proceeds much faster: Particularly, its degradation level after 50 min exceeds that of incubated fresh tissue after 300 min in units of the tongs opening parameter, $\Delta\gamma_{3'/5'}$ (Figure 5.6c). It has been argued that freezing disrupts tissue structure, rendering the tissue highly sensitive to RNA

degradation whereas autolysis of fresh liver tissue appeared to be a much slower process [69].

In summary, the 3'/5'-bias of probe intensities essentially disappears for probes which hybridize predominantly non-specifically and it markedly decreases for probes which are strongly saturated with specific transcripts. The 3'/5'-bias consequently provides a suited metrics for RNA-quality only in the linear range of specific hybridization in agreement with the theoretical predictions made in Section 5.2.1.

5.2.4 Short 3'-probe sets are prone to non-specific hybridization

In the next step we selected the probe sets from the non-specific and specific hybridization ranges of the hook curve and calculated their frequency histograms as a function of L_1 and L_{max}, the position of the nearest and of the most distant probe from the 3'-end in each probe set.

Figure 5.7 shows the distribution of the fraction of probe sets of either hybridization range normalized with respect to the total number of probe sets in the respective group. Probe sets which cover the range near the 3'-end with $L_1 < 100$ and $L_{max} < 500$ are more prone to non-specific hybridization than probe sets located at larger distances from the 3'-end with $L_1 > 100$ and $L_{max} > 500$ which are more affected by specific hybridization on the average. The relative difference of the fractions in both groups is large: For example, the fraction of N-hybridized probe sets exceeds that of S-hybridized ones by about 50% at small $L_{max} < 300$. Vice versa, at large $L_{max} > 700$, the S-hybridized fraction considerably exceeds the N-fraction. The observed distributions are very similar for the different arrays of the Rat-QC data set showing that the positional-dependent variation of the hybridization mode is virtually insensitive to the degree of RNA-degradation.

We suspect that the increased fraction of non-specific hybridization towards the 3' end of the transcripts is caused by inaccurate assignment of the 3'-transcript end upon probe design and/or by variations of the 3'-end of the transcripts, e.g. due to effects such as alternative polyadenylation as discussed previously [85, 86]. Alternative polyadenylation leads to transcript isoforms with differences in the 3' UTR length. In these situations the 'true' 3'-end of the transcript can be located at L3'> 0 and all probes at positions closer to the apparent transcript end, L3' > L > 1, will hybridize exclusively non-specifically owing to the absence of specific transcripts. In consequence, the mean fraction of non-specific hybridization of probes at small L will exceed that of specific hybridization on relative scale, as observed. A very similar plot as shown in Figure 5.7 for the rat genome array RG230A was obtained for alternative array types such the human genome HGU133A plus2 (see Additional File of [87]).

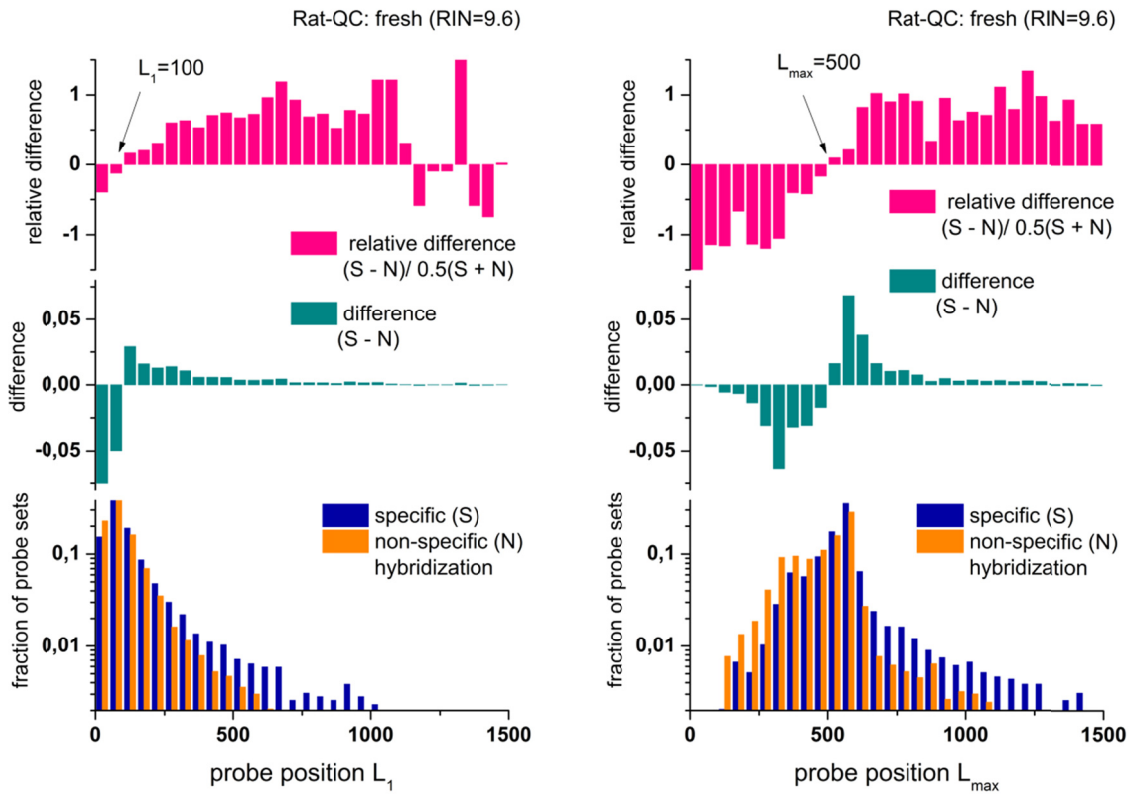

Figure 5.7: Distribution of probe sets hybridized predominantly with specific and non-specific transcripts as a function of the position of the first (left part) and last (right part) probe of each probe set. The graphs show that 'short' probe sets located nearer to the 3'-end of the transcript are more prone to bind non-specific transcripts at ($L_1 < 100$ and $L_{max} < 500$) than specific ones. The part in the middle shows the difference between the respective fractions of probe sets whereas the part above normalizes this difference with respect to the mean fraction of probe sets in both, S- and N-groups. Accordingly, the differential binding refers to about 50% of all probe sets. The distributions are calculated using the ratQC data set.

An alternative option that potentially explains the increase of the relative contribution of N-hybridization near the 3'-end can be sought in depletion of the respective targets in solution due to the high number of probes with partly overlapping probe sequences in this L-range. In consequence, a larger number of probes can be thought to compete for each transcript-fragment than at larger distances L. This competition for the same target can deplete its concentration in solution. Such target depletion effectively reduces the binding affinity of the respective probes for specific binding [44]. In consequence this change can increase the relative contribution of non-specific hybridization as observed in this L-range. On the other hand, it has been shown that depletion is clearly governed by the binding affinity of the probes which exponentially affects the abundance of targets whereas the accumulation of partly overlapping probes near the 3'-end can be assumed to affect target concentrations in a linear and thus much weaker fashion. We therefore suspect that target depletion is, if at all, of secondary importance for explaining the high relative contribution of non-specific hybridization at small L.

In summary, we found a biased distribution of specific and non-specific hybridization along the targeted transcripts: Non-specific binding is more heavily weighted near the

3'-end, presumably owing to its inaccurate assignment and to transcript isoforms with variable 3' UTR lengths.

5.3 Metrics for RNA quality

5.3.1 Positional-dependent intensity decays

The degradation hook and the tongs plot shown in Figure 5.5 highly resolve the 3'-bias of probe intensities in dependence on the hybridization mode. These plots allow classifying each probe set into one of five different hybridization regimes within a microarray experiment. However, this approach only coarsely resolves the positional bias along the transcripts by collecting together three probe intensity values at two or three selected positions only (3', 5' and m).

In this subsection we describe an orthogonal method which uses a more coarse graduation of the hybridization mode while highly resolving the 3'-bias with respect to the probe position. Particularly, we select two groups of probe sets taken either from the N- or the S-hybridization range of the hook curve. We then calculated the logged mean intensities of the selected PM-probes as a function of two alternative arguments, namely their probe index k in the probe set or their probe distance L relative to the 3'-end given in units of the number of nucleotides,

$$\log I^{h}(k) = \left\langle \log I_{p}^{h} \right\rangle\Big|_{p=k} \quad \text{and} \quad \log I^{h}(L) = \left\langle \log I_{p}^{h} \right\rangle\Big|_{L_p = L \pm \delta L} \quad \text{with} \quad h = S, N \tag{5.16}$$

respectively. The angular brackets denote averaging either over all probes with the same index k or over all probes with the same absolute position within a moving window $L-\delta L < L_p \le L + \delta L$.

Figure 5.8a shows the obtained intensity profiles for the example shown in Figure 5.5a. The mean intensity due to specific hybridization markedly decays with increasing distance of the probes from the 3'-end of the transcripts whereas the intensity due to non-specific binding is much smaller and remains virtually constant, as expected. The decay due to specific hybridization can be approximated with a distant-dependent degradation index, $d_p^{P,S} = d^S(L)$ which is given by an 'exponential plus constant' decay law in analogy with Eq. (5.17) (see below). The obtained curves well describe the intensity decay in the intermediate L-range and its flattening at small and large L-values (see dotted curve b in Figure 5.8a).

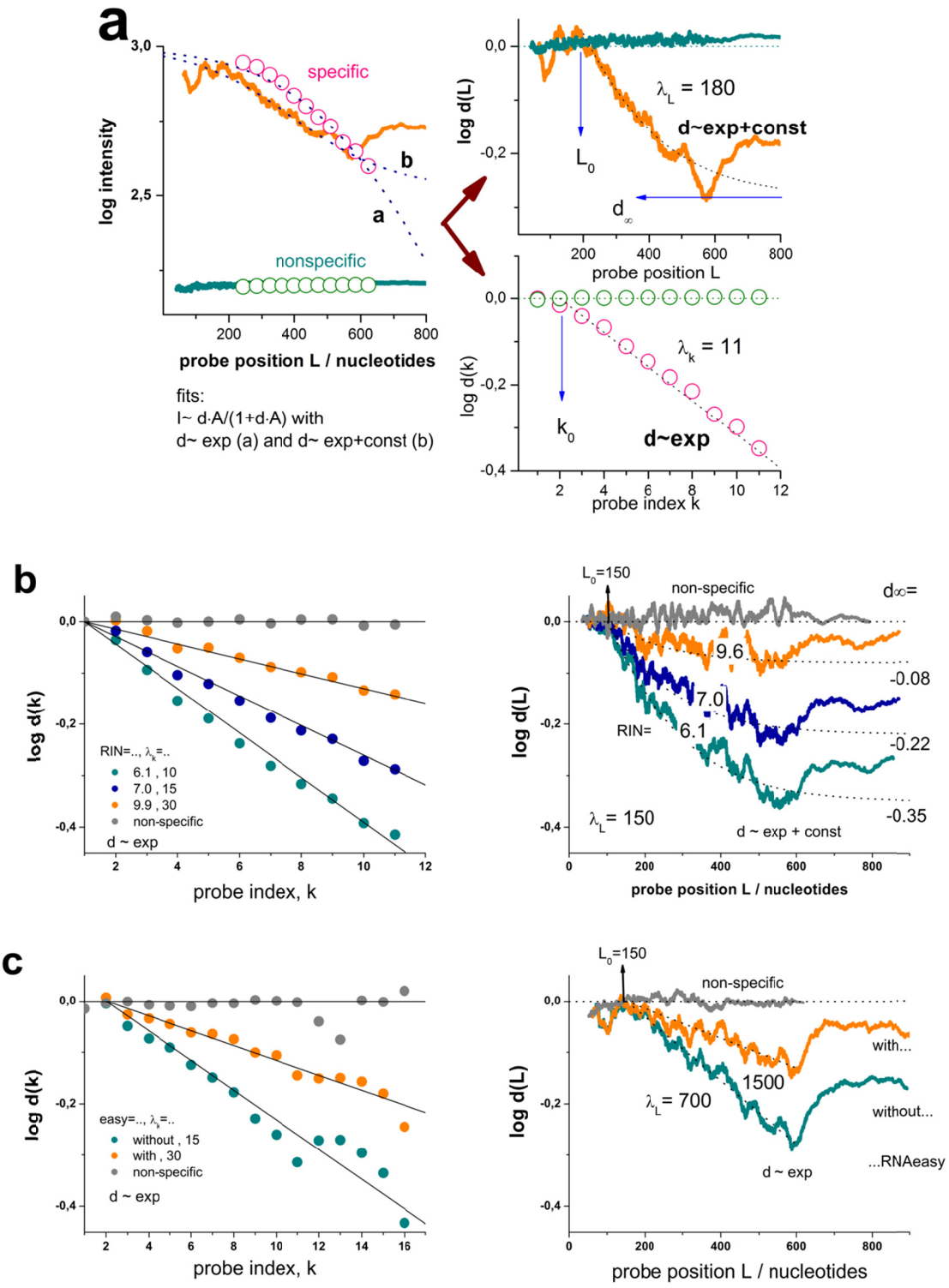

Figure 5.8: Positional dependent intensity decays in relative and absolute scale. Panel a) Mean intensity decays of specifically and non-specifically hybridized probes (Eq. (5.16)) referring to the data shown in Figure 5.5a. The circles denote index-based averages which are plotted as a function of the mean position per index (left part). The decays are normalized according to Eq. (5.17) (right part of the figure). The dotted curves in part a show fits using different functions: Exponential plus constant (a) and exponential (b) intensity decays which consider saturation without initial shift (Eq. (5.18)); exponential plus constant (c) and exponential (d) decays with initial shifts (Eq. (5.18)). Panel b and c) Representative decays are taken from the Rat-QC (b) and the RNeasy cleanup (c) data sets. The index-scaled decays in the left part and the L-scaled decay in the right part of panel c are fit using simple exponential decays ($d_\infty = 0$) whereas the L-scaled decays in part b in addition use a constant $d_\infty > 0$.

This approach attributes the flattening of the decay near the 3'-end to the saturation of the probes with bound transcripts. However this effect becomes relevant usually at large intensity values only (log I ~ log M > 4; see Figure 5.5). The observed mean initial intensity values of the decays are however much smaller (log I(3') ~ 3). We conclude that another effect and not saturation causes the flattening of the decays at small L-values. The decrease of the relative contribution of specific hybridization near the 3'-end discussed in the previous subsection well explains the observed trend: Non-specific hybridization still adds a small residual contribution to the specific decays due to imperfect decomposition of the different hybridization modes. The decrease of the contribution of specific binding presumably due to inaccurate assignment and transcript isoforms then effectively increases the relative weight of non-specific binding and adds a constant component to the decays at small distances from the 3'end which in consequence flattens the initial decay.

To account for this effect we pursue a simple approach which neglects saturation and normalizes the decays with respect to their maximum intensity level near the 3' end of the transcripts,

$$d^h(x) = I^h(x) / I^h(3') \quad \text{with} \quad x = k, L \quad \text{and} \quad h = N, S \tag{5.17}$$

The obtained degradation index due to non-specific hybridization is given by a constant, $d^N(x) \approx 1$, to a good approximation (Figure 5.8). The degradation decays due to specific hybridization are well described using a 'shifted exponential plus constant' functions of the form,

$$d(x) = d^S(x) \approx \left(1 - d^x_\infty\right) \cdot \exp\left(-\frac{x - x_0}{\lambda_x}\right) + d^x_\infty \tag{5.18}$$

as illustrated by the dotted curves in Figure 5.8. The obtained decay length λ characterizes the mean slope of the 3'-bias in units of the number of probes (λ_k) or nucleotides (λ_L) after which the variable contribution of the intensity decays to 1/e of its initial value. The constant d^x_∞ defines the residual constant intensity level at large distances from the 3'-end. The shift-parameters $x_0 = k_0$, L_0 account for the potential flattening of the decay at small arguments discussed above. Both decay constants are linked via the <ΔL>-value, i.e.

$$\lambda_k \approx \lambda_L \cdot \langle \Delta L \rangle \tag{5.19}$$

Panel b and c of Figure 5.8 show selected examples taken from the rat-QC and the RNeasy cleanup data sets which refer to different array types (RAE 230A and HG-U133A, respectively). With decreasing RNA quality the decays become steeper paralleled by

increasing absolute values of the limiting intensity levels but almost constant initial shift parameters $L_0 \approx 150 - 200$ and $k_0 = 1 - 2$. Index- and nucleotide-based length scales give rise to similar trends (compare the right and the left parts in Figure 5.8b and c). The L-scale in units of nucleotides is associated with a slightly more flat and smaller asymptotic level than the relative k-scale using the probe indices as argument. Note that about 95% of the probes of the arrays are positioned with similar frequencies in the range $100 < L < 600$ whereas only less than 5% of them are found at larger distances, however with a broad distribution over the range $600 < L < 2800$ (Figure 5.2). Most of the more distant probes refer to the probe indices $k = 10, 11$. This assignment effectively compresses the asymptotic region to the last two probes with indices $k = 10$ and 11. As a consequence the decays in relative k-scale can be described with sufficient accuracy using a 'single exponential' decays (Eq. (5.18) with $d_\infty^k = 0$) where the values d(10) and d(11) roughly refer to the limiting decay level obtained in the fits using the L-scale, d_∞^L.

The L-decays of specifically hybridized probes obviously behave differently for $L > 600$ showing a less pronounced loss of intensity than for $L < 600$. The origin of this difference is unknown. The standard error of the experimental decays roughly agrees with the symbol size (k-dependencies, left part of Figure 5.8b and c) or it slightly exceeds line thickness (L-dependencies, right part of Figure 5.8b and c). The small oscillations in the decays and the relative increase at $L > 600$ thus reflect systematic effects presumably due to differences of the probe properties in the different subensembles of probes referring to each data point such as their binding affinity and also their degradation degree. Recall that the number of probes drastically decreases at $L > 600$ which makes this range less relevant for correcting purposes of the majority of probes. We exclude this range therefore from curve fitting.

Our fits show that the values of the decay parameters systematically depend on the chosen decay function and strongly correlate each with another. To illustrate this correlation we show fits with variable d_∞^L but constant $\lambda_L = 150$ in Figure 5.8b (right part) and fits with constant $d_\infty^L = 0$ but variable λ_L in Figure 5.8c (right part). The values of the variable parameters d_∞^L and λ_L systematically decrease with progressive degradation. Both options equally well describe the decaying part of d(L) in the range $100 < L < 600$.

To obtain a robust decay characteristics we substitute the exponential fit functions in Eq.(5.18) by a simple two-point estimate

$$\log d^k = \left\langle \log I^S \right\rangle_{k=10,11} - \left\langle \log I^S \right\rangle_{k=1,2} \qquad (5.20)$$

This logged degradation ratio characterizes the intensity decay in the index-range $k_{start} - k_{end} = 2\text{-}10$, or equivalently, in the positional range $L_{start} - L_{end} = \langle L \rangle_{1,2} -$

$\langle L \rangle_{10,11} \sim 150 - 550$ which comprises the majority of more than 95% of all probes. The degradation ratio can be transformed into estimates of the decay length of the exponential decays: $\lambda_k \approx 8 / \ln d^k$ and $\lambda_L \approx 8 \cdot \langle \Delta L \rangle / \ln d^k$.

Please note that the decay function defined in Eq. (5.17) estimates the fold change of transcript abundance at position x and x_0, to a good approximation, i.e.

$$d(x) \approx [S]_x / [S]_{x_0} \tag{5.21}$$

The degradation ratio (Eq. (5.20)) consequently estimates the mean fold change of transcript abundance reported by the probes positioned near the 5'- and 3'-ends of the probed range. It represents an alternative estimate of the tongs opening parameter introduced above, $d^k \propto d_{tongs} = 10^{\Delta\gamma_{3'/5'}}$ (Eq. (5.13)). Figure 5.6c shows the time course of RNA degradation in the Rat QC experiment using the tongs opening (panel c) and the d^k (panel d) parameters. Both measures strongly correlate (see insertion in Figure 5.6d) and essentially reflect the same degradation behavior of the samples studied.

In summary, the effect of degradation can be described as a function of the probe position in terms of a 'shifted exponential decay plus constant'-function using either the probe index or the 'absolute' probe position as argument. This information can be further condensed into a single degradation ratio parameter characterizing the fold change of transcript abundance over the length of the DNA region interrogated by the probes.

5.3.2 3'/5'-controls are affected by the hybridization mode

It was previously shown that the 3'/5' intensity ratios of special control probe sets interrogating long transcripts such as GADPH and beta-actin might not represent a sufficient measure of the degradation bias at small expression degrees because non-specific binding leads to an underestimation of the 3'/5'-bias [45]. Here we show that the controls are often prone to saturation which also leads to the systematic underestimation of the 3'/5'-bias (see also Eq.(5.5)).

The threshold hook represents a modified version of the degradation hook described in Section 5.2.2. It defines a threshold of the 3'/5-intensity ratio of the probe sets used to assess RNA-quality such a GADPH or beta-actin. The threshold hook accounts for the fact that the probe signals are affected by non-specific binding and by saturation. Both effects give rise to an intensity-dependent threshold for estimating good RNA-quality.

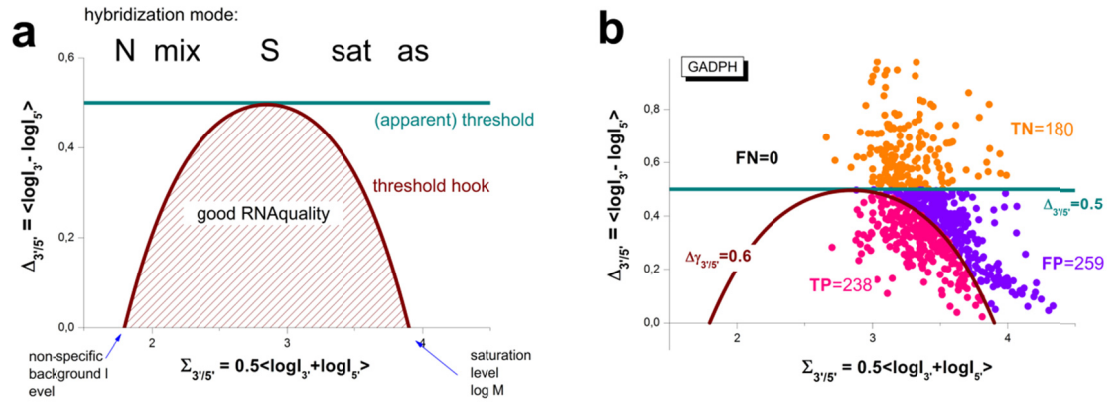

Figure 5.9: Threshold hook for estimating good RNA quality using control probe sets: (a) Constant (apparent) and variable (threshold hook) RNA quality threshold. The true threshold depends on the hybridization regime and vanished upon non-specific hybridization and upon saturation. (b) Error estimates of GADPH-controls taken from the tissue data set (see text).

In the first step one transforms the intensity values of the control probes into hook coordinates

$$\Sigma_{3'+5'}^{\text{control}} = \frac{1}{2}\left(\left\langle \log \mathrm{I}^{\mathrm{PM}} \right\rangle_{3'\text{probeset}} + \left\langle \log \mathrm{I}^{\mathrm{PM}} \right\rangle_{5'\text{probeset}}\right)$$
$$\log \mathrm{d}_{3'/5'}^{\text{control}} \equiv \Delta_{3'/5'}^{\text{control}} = \left\langle \log \mathrm{I}^{\mathrm{PM}} \right\rangle_{3'\text{probeset}} - \left\langle \log \mathrm{I}^{\mathrm{PM}} \right\rangle_{5'\text{probeset}} \qquad (5.22)$$

In contrast to the standard hook (Eq. (3.4)) we here use only the intensities of PM-probes.

In the second step, one calculates the threshold hook as a $\Delta_{3'/5'}$-vs-$\Sigma_{3'/5'}$-plot under the condition that both the 5'- and 3'-probes hybridize according to the hyperbolic Langmuir isotherm (see Eq. (3.2)), however using different specific binding activities due to the different degradation indices of the respective transcripts, $d_{5'} < d_{3'}$(Eq. (5.3)). The delta-value is expressed as a function of sigma using the degradation hook-formalism described in the previous subsection after neglecting the MM-probes (use Eq. (5.12) with $\alpha = -\infty$). The start and end point of the threshold hook are taken from the standard hook analysis which provides these data with relatively high accuracy.

The obtained hook curve thus describes the 'trajectory' of a pairing of 3'/5-probe sets upon changing expression degree of the respective transcript (see Figure 5.9a for illustration). Note that the delta-coordinate directly provides the apparent logged degradation ratio ($\Delta_{3'/5'} = -\log(r_{5'/3'}^{\text{app}})$, see Eq. (5.4)) whereas the 'true' degradation index is given by the height parameter used in the fits ($\Delta\gamma_{3'/5'} = \log(r_{5'/3'}^{\text{true}})$, see Eq. (5.13)). The latter true degradation index is adjusted in such a way that the maximum value of the apparent degradation ratio agrees with the empirical RNA-quality threshold of the chosen control probe. Hence, the threshold hook transforms the constant RNA-quality threshold into a variable one which depends on the hybridization mode of the controls. In consequence,

different data points residing along one hook curve refer to identical true degradation levels irrespective of their different delta coordinates characterizing their apparent degradation level.

The application of a constant threshold instead of the variable one will cause false quality estimates. We estimated the error of the 3'/5'-intensity ratio of the GADPH-control taken from the tissue data set as example: Figure 5.9b shows the hook-coordinates of the GADPH control probe sets of the 677 samples of the human tissue data set (see Eq. (5.22)). The threshold hook and the horizontal line provide the true and the apparent (false) thresholds for good RNA quality in terms of the logged 3'/5'-intensity ratios, $\Delta_{3'/5'} < \Delta_{threshold}$. The constant threshold is assumed to agree with that of the hook curve in the range of specific hybridization. It consequently forms the tangent of the hook-curve at its maximum referring to the S-range of hybridization. The hook curve describes $\Delta_{threshold}$ under the realistic assumption of saturation whereas the constant threshold neglects this effect. As a consequence, data located between both thresholds (colored in blue) define false positives (FP) with respect to the constant threshold whereas data below the hook and above the line are true positives (TP, red) and true negatives (TN, orange), respectively. The number of false negatives (FN) is zero because the hook threshold remains below the constant one. The positive predictive value (PPV = TP/(TP+FP)) and the specificity (Sp = TN/(FP+TN)) are 0.48 and 0.79, respectively, meaning that less than 50% of the 3'/5' controls properly estimate the quality of RNA in terms of good and degraded one. This particular example assumes that the 3'/5' signal ratio for GADPH for good RNA is of no more than 3, or in our notation $\Delta_{3'/5'} < \log(3) \approx 0.5$.

To assess the effect of the hybridization mode on the 3'/5' controls we first estimated the hybridization regime of the GADPH and beta-actin controls of the rat-QC and the human tissue data sets using modified hook plots (Figure 5.10). They depict the logged PM-intensity ratio of the 3'- and 5'-probe sets of the controls ($\Delta^{control}_{3'/5'}$, Eq. (5.22)) along the horizontal coordinate and either the sigma coordinate of each probe set (Σ, Eq. (3.4)) or the mean sigma of both probe sets ($\Sigma^{control}_{3'+5'}$, Eq. (5.22)) along the vertical coordinate axis. In the former plots, each control (GADPH and beta-actin) thus provides two data points per array referring to the 3'- and 5'-probe sets, respectively (see green and blue dots in Figure 5.10). In the latter plots both data points are merged together to illustrate the mean intensity trend of the controls as a function of the degradation index.

To judge the hybridization mode we also depict the sigma coordinates of the non-specific background intensity (N, red dots) and of the asymptotic saturation level (as, black dots) obtained from the standard hook analysis of each of the arrays. Recall, that the sigma-values of the N- and the as-mode limit the range of possible probe intensities. They

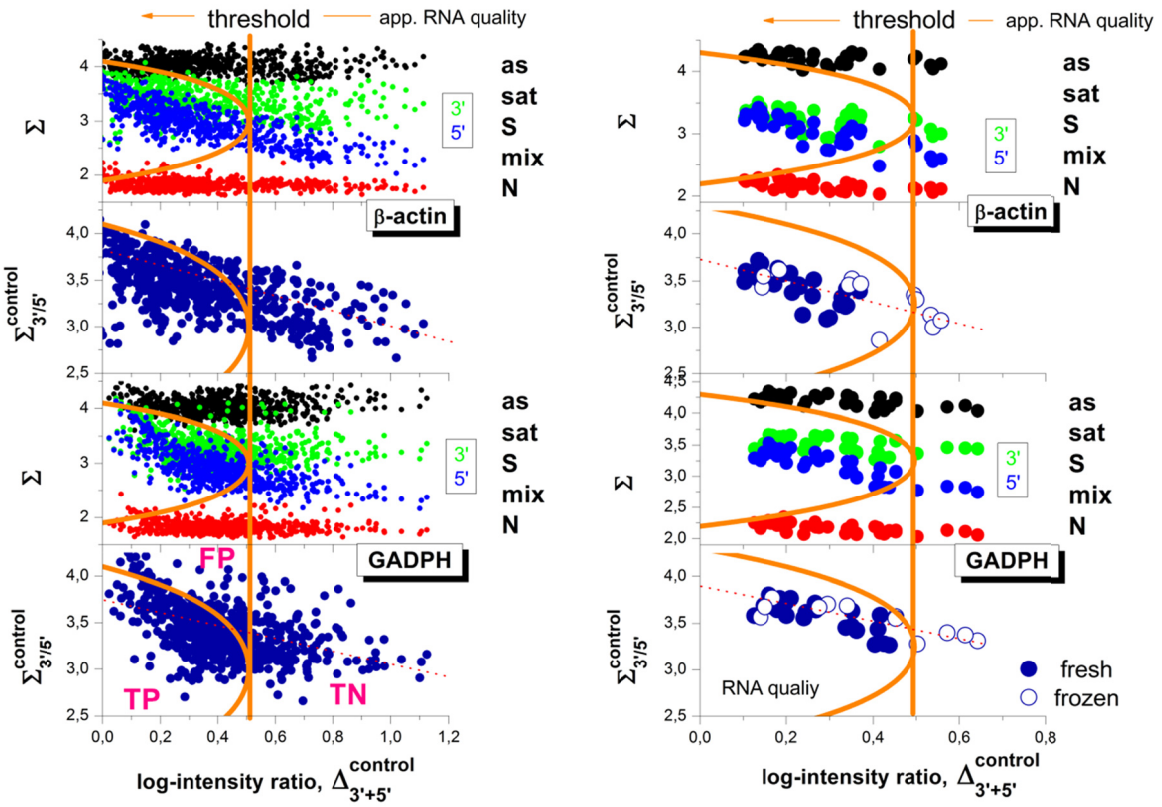

Figure 5.10: Hybridization and RNA-quality characteristics of the GADPH and beta-actin control probe sets in the tissue (left) and rat-QC (right) data sets. Each data point refers to one array of the respective series. The abscissa provides the degradation level in units of the logged 3'/5'- mean intensity ratio of the respective control data sets. The vertical axes plot either sigma coordinates of the 3'- (green dots) and 5'- (blue dots) probe sets of the controls, or their mean (dark blue circles). The red and black dots mark the respective sigma-levels of non-specific binding and of saturation, respectively. The vertical orange lines indicate the constant quality threshold separating good and poor apparent RNA quality. The 'threshold' hooks (orange) refer to the same quality threshold. They however explicitly consider its decrease in the N- and sat-ranges of hybridization. Application of the constant threshold thus produces false positives together with true positives and true negatives.

consequently constitute an intrinsic metrics allowing to assign the probes to one of the five possible hybridization modes as indicated in the figure (see also Figure 5.5). It turns out that small values of the degradation index ($\Delta^{control}_{3',5'} < 0.2$) are often associated with sigma-values near the asymptotic saturation limit of the intensities especially in the tissue data set. We argued before that intensity-based degradation measures are not suited to exactly estimate the degradation level in the saturation limit of the probes. In best case they underestimate the true degradation level; in worst case the intensity ratios become meaningless.

It has been recommended that good-quality samples should have a 3'/5' signal ratio for GADPH and beta-actin of no more than three, or in our notation of $\Delta^{control}_{3',5'} < \log(3) = 0.47$ [83]. We display this threshold as the vertical orange line in both parts of Figure 5.10. It consequently divides the data points of each data set into (apparently) bad and good ones for $\Delta^{control}_{3',5'} >$ threshold and $\Delta^{control}_{3',5'} <$ threshold, respectively.

The 3'/5'-intensity ratio of the probe sets is however not constant for a given RNA-quality level. Instead it depends on the hybridization mode (see above and Eq. (5.5)). Particularly, the 3'/5'-intensity ratio referring to a constant RNA-quality level follows the degradation hook shown in Figure 5.5: It is maximal in the S-hybridization range and vanishes near the N- and as-ranges of hybridization. We plot representative degradation hook curves in Figure 5.10 (see the orange curves; note that the x- and y-axes are exchanged in comparison with Figure 5.5) which are calculated using the threshold value of good RNA-quality ($\Delta\gamma_{3'/5'} = 0.47$) the mean sigma-levels in the N- and as-ranges of the respective data sets. Hence, the degradation hook illustrates that the threshold value of the 3'/5'-intensity ratio strongly decreases in the mix- and sat-ranges due to the progressive effects of non-specific hybridization and of saturation, respectively. It consequently defines a variable, sigma-dependent threshold-curve which allows differentiating between bad and good RNA quality data independent of the particular hybridization mode of the respective probes. In other words, it is more appropriate to apply this variable 3'/5'-'threshold hook' for quality assessment beyond the linear hybridization range instead of using a constant threshold value of the 3'/5'-intensity ratio.

For example, a large fraction of the GADPH- and beta-actin intensity ratios of the tissue data set meet the constant quality criterion, $\Delta^{control}_{3'/5'} < \text{threshold} = 0.47$, indicating apparently good RNA quality (Figure 5.10, right part). Consideration of the hybridization-dependent 'threshold hook' divides this region further into true positive estimates ($\Delta^{control}_{3'/5'} < \text{hook}_{threshold}$) and false positives ($\text{hook}_{threshold} < \Delta^{control}_{3'/5'} < \text{threshold}$), where the latter data are located between the curved and linear thresholds as shown in Figure 5.10. We estimated a positive predictive value for GADPH controls of about 0.48 which reflects overestimation of RNA-quality for about 50% of all 677 arrays of the tissue data set. Note also that strong saturation of the probes can completely prevent detection of poor RNA-quality samples because the respective intensity ratio levels off to $\Delta^{control}_{3'/5'} = 0$.

The mean sigma coordinates ($\Sigma^{control}_{3'+5'}$) of the Rat-QC data set are found approximately halfway between the respective N- and as- levels indicating that the controls are predominantly hybridized in the S-range (Figure 5.10, left part). Application of a constant quality threshold seems appropriate for this data.

The sigma values of both data sets studied clearly indicate the decrease of the mean intensity of the controls with decreasing RNA quality due to the loss of material assumed, e.g. in Eq. (5.21). In consequence, the hybridization regime of the controls can shift with changing RNA-quality. Note also that GADPH is associated with slightly larger probe signals than beta-actin in both data sets. Beta-actin controls are consequently less prone to saturation than GADPH controls.

In summary, control probes can overestimate RNA-quality if one uses a constant threshold criterion because the true threshold level strongly decays for saturated probes. The problem can be fixed either by using an intensity dependent 'threshold' hook or by using alternative RNA-quality estimates such as the degradation ratio d^k.

5.3.3 Affy-slope is affected by absent probes

A widely applied metric for RNA quality is the 'RNA degradation plot' provided with the R package *affy* [15-16]. The RNA degradation plot displays the mean log intensity averaged over all probes with the same index k of one microarray as a function of the probe index, $k = 1,...,N_{pset}$. The slope of the regression line then provides a summary measure to characterize the mean degree of RNA-degradation in a chip-specific fashion. Note that the affy-slope parameter originally does not intend to serve as an absolute RNA quality measure per se but instead, represents a relative measure for comparing RNA quality between different chips in a particular series of measurements.

However, the affy-slope degradation plot is virtually identical with the reciprocal positional dependent degradation index introduced above in Eq. (5.17) ($d^h(k)^{-1}$) except the fact that it considers all probes of the array whereas our approach separately averages over the N- and S-subensembles referring either to the S- or N-hybridization regimes, respectively. The affy-slope estimates are expected to underestimate the degradation level owing to the inclusion of predominantly non-specifically hybridized probes (so-called absent probes) which do not respond to RNA quality as shown above. More importantly, the chip-to-chip variability of the fraction of absent probes (%N; as determined by methods such as MAS5 or hook) is expected to affect the affy-slope measures by factors which are not or only weakly related to RNA quality.

To illustrate this effect, a series of affy-slope curves referring to different degradation levels are shown in Figure 5.11a. Panel b of the figure plots our degradation profiles $d^S(k)$ of the specifically hybridized probes for the same arrays. Both presentations provide similar trends for the microarrays with similar %N-values. However, affy-slope and our degradation plot provide different results for arrays with marked differences of %N, as expected. Particularly, affy-slope tends to underestimate the slope for large %N values and thus to overestimate RNA quality.

Hence, the apparent degradation ratio derived from the simple affy-slope intensity measures is strongly modulated by the fraction of non-specifically hybridized 'absent' probes leading potentially to the systematic overestimation of RNA quality. Contrarily, the proposed use of specifically hybridized probes largely removes this bias from the data and

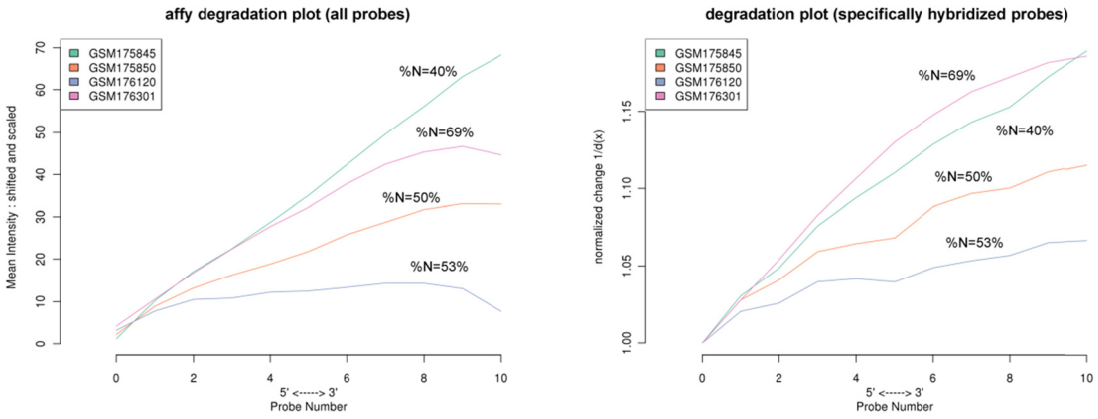

Figure 5.11: RNA degradation plot of all probes (panel a) and degradation profile of specifically hybridized probes (b) for microarrays selected from the human tissue data set. Panel a shows the plots obtained using the affy package [75] *whereas the curves in panel b are given by the inverse of the degradation function* $d^S(k)^{-1}$ *(see Eq. (5.17)). The slopes of most of the curves rank in the same order in both panels, except the two curves of steepest slope which reverse order in both parts of the figure owing to the different percentage of absent probes. The percentage of absent probes are %N = 40% (GSM175845), 69% (GSM176301), 50% (GSM175850) and 53% (GSM176120) as determined by the hook method.*

provides a reliable measure of the degradation degree which can be consistently compared between different arrays.

5.3.4 Array-degradation metrics correlate with RIN

The RNA Integrity Number (RIN) provides a numerical value for the assessment of RNA quality based on the electropherogram trace of a RNA sample captured with the Agilent Bioanalyzer [72]. The RIN is widely used and its scale ranges from 1 to 10 (most to least degraded). A RIN-cutoff of RIN ≥ 7 is recommended for obtaining good-quality RNA for microarray analysis [84]. Figure 5.12a compares our d^k degradation measure with the RIN reference values obtained in the ratQC experiment. Both measures correlate strongly, however, the two samples and incubation conditions result in different slopes of the regression lines. In other words, each microarray degradation parameter does not unambiguously transform into one RIN-value especially at larger degradation levels. Instead, the two different samples and incubation conditions reflect a bimodal relation between the two types of measures: each RIN value splits into two d^k options and vice versa. Note also that correlation coefficient between RIN and our improved d^k-degradation measure exceeds that between RIN and affyslope (r = 0.95 vs 0.92, RatQC fresh) owing to the reasons discussed in the previous subsection.

In panel b of Figure 5.12 we re-plot the d^k degradation parameter as a function of the mean transcript length measured independently using the Bioanalyzer [69]. The two branches of the d^k -vs-RIN plot merge into one within the error limits. This result confirms that our

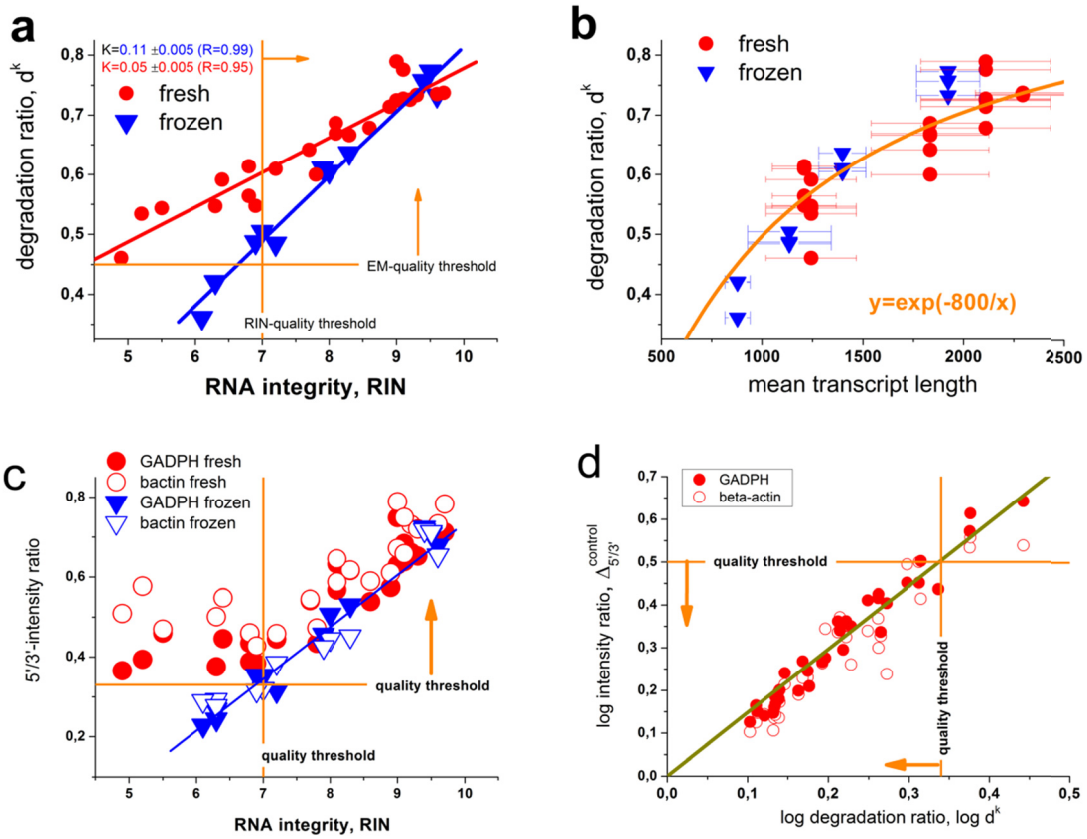

Figure 5.12: Comparison of microarray degradation measures (d^k and 5'/3'-ratio of the hybridization controls) with the RNA integrity number (RIN, panel a and c) and with the mean length of the transcripts (panel b) obtained in the ratQC experiment [69]. *The d^k parameter split into two branches for the two sample treatments when plotted as a function of RIN whereas the d^k data virtually merge into one branch if plotted as a function of transcript length. Panel d correlates the d^k and the 5'/3'- intensity ratio of the control probes in logarithmic scale. The vertical and horizontal orange lines indicate the respective quality thresholds. Good RNA-quality probes are found in direction of the arrows.*

microarray-based degradation measure more directly relates to the mean transcript length and thus to the state of the mRNAs that the microarray experiment intends to quantify. The RIN however represents an alternative integrity measure capturing a series of electropherogram features that are indicative also for additional properties of the RNA solution such as the ratio of larger to smaller molecules and how far the degradation process has proceeded [69]. The correlation of the 5'/3' ratios of the degradation controls with the RIN-numbers reveals subtle differences compared with the behavior of the d^k parameter (Figure 5.12c). Particularly, the hybridization control-measures taken from the 'fresh' samples are virtually independent of degradation at RIN<7 whereas for RIN>7 both treatments give rise to similar behavior of the controls. Recall that the GADPH and beta-actin controls cover a slightly wider range of the transcripts (from about 200 to about 1000 nt, see Figure 5.2b) than the d^k degradation ratio which probes the range from about 100 to about 550, on the average (see Figure 5.2c). This difference presumably explains the smaller values of the control ratios at larger expression degrees. More importantly, the d^k parameter is calculated as the average over a large number of probes. Presumably both

types of parameters respond differently to changes of the length distribution of the transcripts due to degradation. Below we address this issue more in detail.

The regression line between the d^k parameter and the 5'/3'-intensity ratio of the controls allows to transform the quality threshold of the latter ratio into a d^k threshold (Figure 5.12d, see orange lines). We replot these thresholds into panel a and c of Figure 5.12: The RIN threshold is clearly more restrictive assigning more samples to bad RNA-quality than the threshold of the microarray-based control probes.

5.4 Degradation reduces total transcript abundance

We so far estimated RNA quality in relative units using suited 5'/3'-intensity metrics which reflect the decrease of transcript abundance with increasing distance from the 3'-end. Trivially, this effect is expected to reduce the total amount of mRNA used for hybridization. The decrease of the mean intensity of the control probe sets with increasing degradation ratio as shown in Figure 5.10 confirms the decrease of the total amount of the respective specific transcripts with progressive degradation. Figure 5.10 also shows the non-specific intensity level of each of the arrays studied (red dots), which tends to decrease with increasing degradation.

The hook method enables the independent estimation of the mean levels of non-specific 'background' hybridization and that of specific expression using the simple summary measures β (see Eq. (3.14)) and φ (Eq. (5.10)) which are based on large ensembles of probe sets on each array. Particularly, the width of the hook curve β has been shown to relate to the total amount of RNA material [45, 57]. Figure 5.13a shows how β decreases with progressive degradation. The observed decrement indicates that the amount of RNA material decreases by about 40% in the Rat-QC experiment.

The degree of specific binding drops upon degradation, however to a considerable smaller degree than the amount of non-specific binding (Figure 5.13b). This discrepancy surprises because naively one expects that the loss of material similarly affects specific and non-specific binding on the average, i.e. $d^N \approx d^S$. The mean hybridization levels of specific and non-specific binding are however directly related also to the respective mean binding constants, $\langle K^{P,S} \rangle$ and $\langle K^N \rangle$, respectively (Eq. (5.3)). We have previously shown, that the decrease of RNA-material used for hybridization increases the specific binding constant due to weaker bulk hybridization and vice versa [57]. In consequence, this so-called up-down effect will partly compensate the decrease of the concentration of specific transcripts giving rise to the smaller decrease of the specific hybridization strength upon RNA degradation.

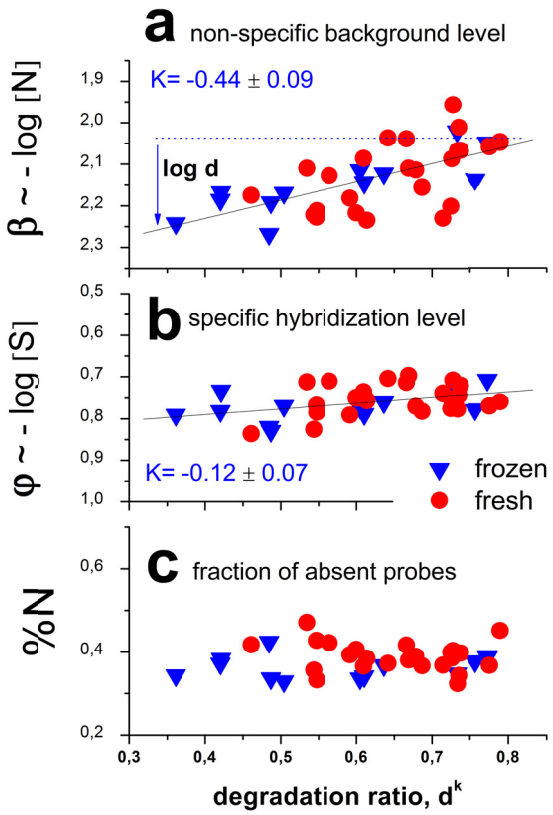

Figure 5.13: Hook-hybridization characteristics of the arrays of the ratQC data set. (a) The width of the hook curves β increases with progressive degradation indicating the decrease of the non-specific background due to the loss of material (see Eq. (3.14)). Log d is the mean degradation index (Eq. (5.2)) and K the slope of the regression line. The mean level of specific hybridization changes only weakly with degradation (b). The fraction of absent probes is virtually unaffected by degradation (c).

Part c of Figure 5.13 depicts the percentage of absent probes detected on each of the arrays. It remains essentially unaffected by RNA degradation. This result shows that the loss of material does not affect the detection threshold of the array experiment for specific binding.

5.5 Correction of the 3'/5' bias

5.5.1 RNA-quality scaling of gene expression

It has been previously found that, although moderate levels of RNA degradation are tolerated by differential expression analysis, beyond a threshold especially long targets provide erroneous expression results [69, 88]. Systematic large-scale microarray analyses reveal that the expression values of up to 30% of all probed genes significantly correlate with degradation quality measures such as the 3'/5'-ratios of control genes [63, 76]. The observed correlations can be well explained on the basis of the results presented here: For example, it is found that the expression values of weakly expressed genes negatively

correlate with the quality of their transcripts [63]. The authors explain this '…the worse the quality the stronger the signal…'-effect by either the enrichment of low quality RNA in the low signal range due to nonspecific hybridization or by compensating effects due to chip-to-chip normalization. The former interpretation disagrees with our results presented in the Section 5.4. We found that progressive degradation dilutes the sample and by this way decreases the amount of nonspecific hybridization. On the other hand, the observed negative correlations also mean '…the better the (apparent) quality the weaker the signal…' in agreement with our results: For low intensity signals the 3'/5'-ratio indeed improves with decreasing intensity suggesting better RNA-quality. We demonstrated that this trend is however caused by the increasing amount of nonspecific hybridization and not by improved RNA-integrity.

Considering also correlations between 3'/5'-quality measures and signal values (called LEV, 'labeling extension values'), Lee *et al.* [76] found that LEV are typically small at low expression values but step-wisely increase beyond a certain expression threshold. The authors hypothesized that the positional 3'/5'-bias is less notable for low abundant transcripts due to inefficient reverse transcription. However, according to our results, the observed trend can be explained by the dominance of non-specific hybridization lacking positional 3'-bias at small expression levels. These two examples demonstrate advantages of model-based expression analysis using physicochemical hybridization theory compared with simple correlation analysis.

The aim is therefore to use the degradation model for correcting the 3'-probe intensity bias to provide (largely) unbiased probe signals for downstream analysis. One expects that the loss of RNA material in general and particularly, RNA-fragments probed far away from the 3'-end, systematically decreases the apparent expression degree extracted from microarray probe intensities.

5.5.2 Correcting the 3'/5' bias of probe intensities

Two main factors related to RNA quality potentially affect the intensities of the probes : (i) the distance of a probe relative to the 3'-end of the transcript, L (or, alternatively, the probe index in the probe set, k, which counts the probes in direction away from the 3'-end of the transcript) and (ii), the hybridization mode [87]. The specific hybridization regime below saturation is particularly prone to biased intensities as opposed to non-specific hybridization and specific hybridization in the asymptotic saturation range.

Under consideration of these factors, the raw probe intensities of each sample are corrected

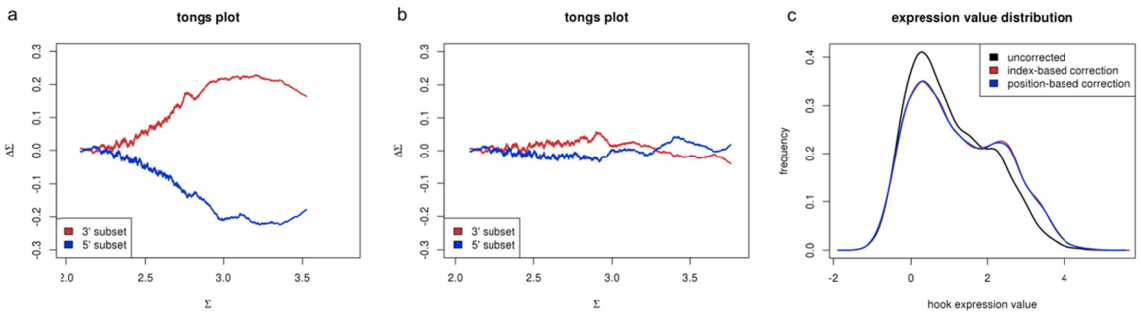

Figure 5.14: Tongs plot for a single array hybridized with strongly degraded RNA in the RatQC experiment (RIN = 6.1) before (panel a) and after (panel b) correction of the probe intensities for the 3'-bias (compare Figure 5.6). Panel c shows the respective distributions of expression values before correction and after index- and position-based correction (see next section). The right flank clearly shifts towards larger expression values after correction.

by the following algorithm:

1. Calculate the degradation hook $\Delta\Sigma_{3'/5'}$-vs-Σ using all perfect-match probe intensities for the current array as described in Section 5.2.2.
2. Probes are considered as specifically hybridized if the sigma-value of the respective probe set meets the condition $\Sigma^{max}_{3'/5'} - 0.4 < \Delta\Sigma_{3'/5'} < \Sigma^{max}_{3'/5'} + 0.2$ where $\Sigma^{max}_{3'/5'} = \arg\max\{\Delta\Sigma_{3'/5'}(y)\}$.
3. The decay function $d^S(x)$ (x = k, L) is calculated as described in Section 5.3.1 using the subensemble of all specifically hybridized probes.
4. The mean fraction of probe intensities due to specific hybridization is estimated for each probe set as, $f^S(y) = \Delta\Sigma_{3'/5'}(y) / \Delta\Sigma^{max}_{3'/5'}$.
5. The correction function is calculated as weighted sum of the decay functions due to specific and non-specific hybridization where the latter one is simply set to unity, $d^N(x) = 1$, i.e. $C(x,y) = d^S(x) \cdot f^S(y) + d^N(x) \cdot \left(1 - f^S(y)\right) = d^S(x) \cdot f^S(y) + \left(1 - f^S(y)\right)$
6. The biased probe intensities are then corrected using the inverse of the correction function, $I_p^{P,x-corr} = I_p^P / C(x,y)$.

Each probe intensity is rescaled according to value of the mean intensity decay at its position (x = k or L) and according to its hybridization mode as indicated by the abscissa-value of its probe set y. Consequently, probe intensities taken from the non-specific hybridization range remain uncorrected. With increasing degree of specific hybridization the probes are progressively scaled up with increasing distance from the 3'-end of the transcript. The maximum correction applies to probe sets in the S-hybridization range. MM probe intensities are scaled using the mean logged MM-intensity of the probe set as argument.

Figure 5.14 illustrates the effect of the correction by replicating the tongs plot for a sample from the RatQC data set with strong degradation (RIN = 6.1, also compare Figure 5.6) before (panel a) and after (panel b) correction with the algorithm described above. It

demonstrates that the tongs opening, and thus the degradation bias affecting particularly specifically binding probes, is largely removed from the intensity data. Figure 5.14c shows the frequency distributions of expression values obtained after hook calibration of uncorrected, index- and position-based 3'-corrected intensity data (see below). The correction shifts the right flank towards larger expression values. Both correction methods affect the distribution nearly identically. The distribution reflects the mean correction amplitude without emphasis on the individual probe sets.

5.5.3 Index and position based correction

We here compare the effect of the correction on the obtained expression values using either on the absolute probe position ('L-correction') or on the relative probe position (index-based, 'k-correction') relative to the 3' transcript end. Consider the special case of predominant specifically hybridized probes below saturation (S-regime, $f_s = 1$) which implies that expression and intensity values roughly agree owing to the small effect of non-specific hybridization. We also assume an exponentially decaying correction function for sake of simplicity ($d_\infty^x = 0$ in Eq.(5.18)). The logged mean intensity averaged over a selected probe set then becomes after correction (compare the correction algorithm in Section 5.5.2 with $f_s = 1$)

$$\left\langle \log I_p^{P,x-corr} \right\rangle_{pset} = \left\langle \log I_p^{P} \right\rangle_{pset} + \left\langle x \right\rangle_{pset} / \lambda_x \cdot \ln 10. \qquad (5.23)$$

Let us first consider the index-based correction. The probe set averaged mean index is identical with the array-related mean index averaged over all probe sets of the array, i.e. $\left\langle k \right\rangle_{pset} = \left\langle k \right\rangle_{array}$, if all probe sets contain the same number of probes. This applies to GeneChip microarrays to a good approximation because the overwhelming majority of probe sets contains the same number of probes per set (usually $k_{max} = 11$ and thus $<k>_{array} = 5.5$). The index-based correction consequently scales the intensity values referring to specific hybridization ($f_S = 1$) of one array by a constant factor, or, in log-scale, adds the increment term $\sim \left\langle k \right\rangle_{array} / \lambda_k$ (see Eq.(5.23)).

Contrarily, the position-based correction applies a specific correction $\sim \left\langle L \right\rangle_{pset} / \lambda_L$ to each probe set. The mean position of the probes of each probe set varies from set to set and thus it usually deviates from the mean value averaged over all probe sets on the array, i.e. $\left\langle L \right\rangle_{pset} \neq \left\langle L \right\rangle_{array}$.

Using the previously defined $<\Delta L>$ (Eq. (5.1)) then allows to link the index- and position-corrected mean intensities

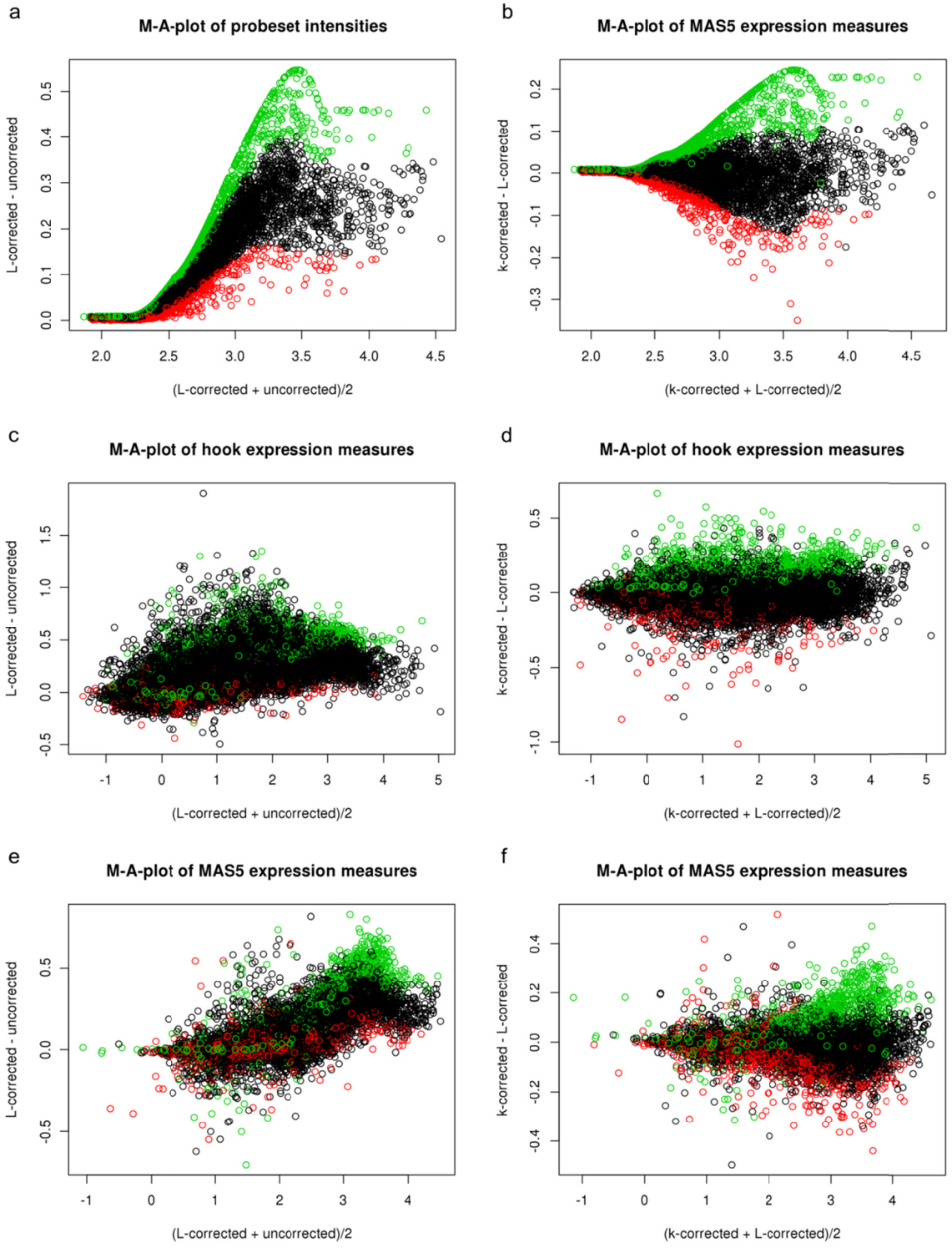

Figure 5.15: M-A plots of the L-corrected versus uncorrected intensities (panel a), and L-corrected-versus-k-corrected intensities (b) of the RIN = 6.1 sample of the RatQC series. Each symbol refers to the mean of the logged probe intensities averaged over the probe set. Panel c – f show the respective M-A plots of expression values obtained after intensity calibration using the hook method (panel c and d) and MAS5 (e and f). The lower quartile of probe sets which are located closer to the 3'-end of the transcripts are colored in red and the upper quartile of probe sets located far away from the 3'-end are colored in green.

$$\left\langle \log I_p^{P,L-corr} \right\rangle_{pset} \approx \left\langle \log I_p^{P,k-corr} \right\rangle_{pset} + \left(\frac{\langle L \rangle_{pset}}{\langle L \rangle_{array}} - 1 \right) \cdot \langle k \rangle_{array} / \lambda_k \cdot \ln 10. \qquad (5.24)$$

Eq. (5.24) shows that both correction types agree if the set averaged mean position of the probes agrees with the respective total array average, $\langle L \rangle_{pset} = \langle L \rangle_{array}$. For probe sets with a mean position nearer the 3'-transcript end (i.e. $\langle L \rangle_{pset} < \langle L \rangle_{array}$) the index-based correction exceeds that of the position-based correction whereas for $\langle L \rangle_{pset} > \langle L \rangle_{array}$ this relation reverses. The analysis of a series of different array types shows that the 25%- and 75%-percentiles of the distributions of $<L>_{pset}$ provide correction factors $\left(\langle L \rangle_{pset} / \langle L \rangle_{array} - 1 \right)$ = - 0.4 – -0.5 and +0.4 - +0.5, respectively (compare Figure 5.3). Hence, the position-specific correction deviates from the index-based correction by more or less than +0.1/-0.1 for 50% of the probe sets if one assumes $\lambda_k = 10$ referring to relatively strong degradation (e.g. RIN = 6.1 of the ratQC data set, see Figure 5.15b). In the mix-range one expects the same qualitative relations between both options for correcting the 3'-bias of expression values, however with a systematically reduced amplitude due to the down-weighting of the effect ($f_S < 1$).

Hence, the index-based correction effectively applies the same factor to all probe sets which is scaled solely by the degree of specific hybridization whereas the positional correction applies a specific factor to each probe-set. The M-A–plot in Figure 5.15b shows the difference between L-corrected and k-corrected intensities. Each point represents the mean of the logged probe intensities over a probe set using the same strongly degraded microarray sample as in Figure 5.14 (RIN = 6.1 from the RatQC experiment). The points have been colored according to the average location <L> of the probes within each probe set: the lower quartile of probe sets located closer to the 3'-end of the transcripts are colored in red and the upper quartile of probe sets located far away from the 3'-end are colored in green. Figure 5.15a shows that the log-scale correction increment increases with increasing intensity level of the respective probes set, with the strongest corrections of $\Delta \log I = 0.55$ for the probe sets which are more distant to the 3' end. A comparison of the red and green symbol in Figure 5.15b shows that probe sets located on the average nearer to the 3'-end of the transcripts are corrected to a less degree than probe sets located more distant from the 3'-end of the transcripts for the position-based correction compared to the index-based corrections.

Panels c-f of Figure 5.15 show similar M-A-plots as in panels a and b, but this time based on expression values as computed with MAS5 and with the hook calibration methods. The normalization and summarization steps applied to the probe intensities result in a more heterogeneous effect of the corrections which however shows the same general trend as discussed for the intensity data.

In summary, the k-correction applies the same positional factor to all probe sets. In consequence, the probe set-specificity of the correction is solely determined by the degree of specific hybridization. Contrarily, the L-correction applies a specific factor to each probe-set depending on the particular location of its probes. Comparison of both correction methods shows that probe sets located on the average nearer to 3'-end of the transcript are corrected to a less degree using their absolute position than probe sets located more distant from the 3' transcript end. Hence, the L-correction is more specific with respect to each particular probe set. On the other hand, the k-correction is more robust with respect to outliers.

We recommend use of absolute probe positions to cope with the effect of differently distributed probes. In practice the intensity changes due to index-based and position-based correction differ only slightly with, in general, small differences in the resulting expression values.

5.6 An R package for the analysis and correction of RNA quality effects

We developed the R package *AffyRNADegradation* that facilitates the analysis of RNA quality of Affymetrix expression data. It provides programmatic access to the RNA quality measure described in Section 5.3.1 that overcomes the drawbacks of existing methods by strictly referring to specific hybridization. Furthermore, it enables correction of the 3'-probe intensity bias for improved downstream analysis. We will here illustrate the functionality of the *AffyRNADegradation* packages using the RNeasy data set where the same cell extract has been used for multiple microarray hybridizations, however either prepared with RNeasy to remove RNA degrading enzymes, or not [81].

The first package functionality addresses the analysis of the effects of RNA degradation and amplification on the microarray signals. The degradation hook-plot, shown in Figure 5.16a and b, displays this 3'/5' intensity difference in dependence on the mean logged probe intensity approximating the expression degree of the respective gene (see Section 5.2.2). Cross-hybridization of partly matching targets of other genes causes nearly equal intensities for weakly expressed genes [37]. With increasing expression competitive binding of specific targets progressively unmasks their actual 3'/5' gradient, until probe saturation sets in. Desirable would be equal intensities for 3' and 5' probes for all expression levels. The maximum height of the hook-plot reflects the relevant 3'/5'-intensity gradient of the selected array enabling the unbiased comparison of differentially

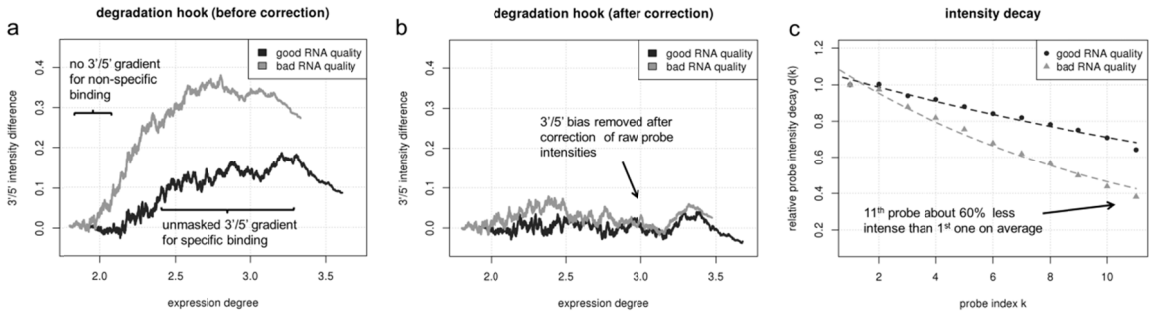

Figure 5.16: Degradation hook plots referring to strongly and weakly degraded RNA taken from the RNeasy data set before (panel a) and after (panel b) correction using AffyRNADegradation. The height of the hook curve increases with increasing degradation level. Panel c shows the respective probe positional decays d(x) as plotted by the AffyRNADegradation package: the worse the RNA quality, the steeper is the respective decay.

expressed genes under variable RNA quality. The hook-plot is accessible using the `PlotDegradationHook` function in the package. A complementary representation is the Tongs Plot (see Section 5.2.2) which is accessible using the `PlotTongs` function.

The second package functionality addresses the estimation of RNA quality of a particular sample. We found in Section 5.2.3 that one should only use specifically hybridized probes for the estimation of RNA quality because of the 3'/5' gradient of the intensity as a function of the expression degree. For these probes we compute the mean probe intensity separately for each probe index k = 1…11 starting from the 3' end of the target transcript. Figure 5.16c shows the resulting probe positional intensity decay after normalization with respect to the mean intensity for the first probe k = 1. Alternatively the intensity decay can be calculated as a function of the distance L of the probes given in units of nucleotides from the 3'-transcript end (see also Figure 5.8).

We determine the decay-length parameter d from the mean intensity decays of all specifically hybridized probes. As we showed in Section 5.3, it provides an accurate estimate for the RNA quality of a particular array hybridization improving other array-based metrices. The d(x = k,L) plot is available via the `PlotDx` function and the RNA quality estimate is available via the `d` function in the *AffyRNADegradation* package.

The third package functionality addresses the correction of the RNA quality bias. Differences in RNA quality and the resulting probe positional intensity decay are technical artifacts which can affect expression measures and the results of differential expression analysis. We here aim at removing the systematic differences in probe positional intensities between different conditions. Figure 5.16a shows two such conditions in the example data relating to degraded transcripts due to increased presence of RNases not removed by RNeasy treatment. *AffyRNADegradation* uses a correction function that reverses the probe positional intensity decay d(x) after applying the expression level dependency of the hybridization mode as described in Section 5.5.2. Optionally, the correction can be

performed based on probe indices k as well as probe distances L. Figure 5.16b shows the degradation hook after application of the correction using probe indices k: The 3'/5' bias is almost completely removed. Corrected probe intensities are available via the `afbatch` function.

The *AffyRNADegradation* package extends the Bioconductor package *affy* [75] and integrates well in a typical microarray analysis workflow. All calculations are performed directly on the `AffyBatch` object and carried out separately for each particular microarray hybridization in a single-chip approach. Our approach corrects the 3'/5'-bias on the level of raw probe intensities which can afterwards be processed with any method. The runtime is about 2 minutes and 3 minutes per sample for index and distance based corrections, respectively. Since each chip is processed independently, arbitrarily large data sets can be processed.

5.7 Summary and conclusions

Amplification of RNA-material using primed in-vitro transcription protocols and degradation of RNA during extraction, storage and processing of the samples affects RNA-quality in microarray experiments with consequences for expression estimates and their interpretation. We systematically analyzed the effect of varying RNA quality on microarray probe intensities using a physicochemical hybridization model and propose (i) new measures to assess RNA quality and (ii), a method to correct probe intensities for the degradation bias.

Particularly, it is shown that poor RNA quality is associated with a 3'-bias of transcript abundance which affects only the probe signal due to specific hybridization. Estimation and correction of the resulting signal bias of each particular probe requires consideration of its hybridization mode (specific, non-specific or a superposition of both) and of the positional effect of probe intensity along the respective gene due to truncated transcripts. The former issue is solved by applying a modified 'hook'-approach of data analysis based on Langmuir hybridization theory. The latter effect is taken into account by estimating the mean positional intensity decay on each array as a function of either the probe index or the probe's distance to the 3' end of its target transcript.

RNA quality is estimated in terms of the 3'/5'-intensity gradient of specifically hybridized probes. In addition to appropriate quality values (such as the 'tongs opening'-parameter and the degradation ratio d) we introduce graphical characteristics allowing assessment of RNA quality of each single array ('tongs plot' and 'degradation hook'). The parameters have a well-defined physical meaning related to the fold change of transcript abundance

along the genes. 'Poor' RNA quality is characterized roughly by a decay of the mean specific signal by a factor of less than 0.5 between probes near the 3'-end and probes located about 600 nt away.

Our approach improves established RNA-integrity measures such as 'affyslope' and the 3'/5'-intensity ratio of degradation control probe sets. Both methods are prone to overestimate RNA quality if the signals are dominated by non-specific hybridization (affyslope) and/or saturation (controls). Our microarray-based quality estimate correlates well with the RNA integrity number (RIN) which, in addition, is affected by more complex properties of RNA degradation not uniquely related to transcript length. Short probe sets near the 3'-end are prone to non-specific hybridization presumably because of uncertainties in 3'UTR length owing to inaccurate assignment of the 3'-end and transcript isoforms.

Poor RNA quality is associated with a decreased amount of RNA material hybridized on the array paralleled by a decreased total signal level. Additionally, it causes a gene-specific loss of signal due to the positional bias of transcript abundance which requires an individual, gene-specific correction. The former total effect can decrease the overall signal level of an array by the factor of 0.5 - 0.7 in the case of poor RNA quality (RIN < 7). The latter local effect can be more pronounced with a penalty in expression measures by a factor of 0.3 - 0.4 or even less in worst cases.

The functionality to assess and to correct RNA quality effects in GeneChip expression data has been implemented in the software package *AffyRNADegradation*. It provides programmatic access to the degradation measures d^k and d^L as well to the tongs and degradation plot visualizations which help to assess RNA quality. Furthermore, it allows correcting probe intensities for the degradation bias for more reliable downstream expression analysis. The *AffyRNADegradation* package is implemented in R and freely available via the Bioconductor software repository[4].

[4] http://www.bioconductor.org/

6 Sequence effects

6.1 Probe sequence affects intensities and expression values

The mechanism of nucleic acid binding on solid surfaces is the basic principle of microarrays and a significant number of other technologies widely used in life sciences. Yet there are many unknowns among the factors affecting the binding process itself. It is known that the base composition has a large effect on duplex yield in solution, particularly for short oligonucleotides as used in microarrays [89]. This is mainly due to the higher stability of duplexes containing GC base pairs compared to AT base pairs. Besides these probe-target interactions there is a substantial number of additional interactions that occur on microarray surfaces: probe-probe interactions [89], probe-folding [90], non-specific binding [47], intra-target RNA folding [44], target-target interactions [91], steric crowding [92], sequence-specific fluorescence marking and more [37]. These interactions alter the effective binding of marked nucleic acids to the probes, and thus the observed fluorescent intensity signal. They must therefore be studied thoroughly in order to improve the specificity and sensitivity of those signals and to fully understand the dependence of intensity signals on factors like probe sequence.

Figure 6.1 shows the surface image of a hybridized Affymetrix GeneChip expression array. The image clearly reveals dark and bright horizontal stripes which correlate with the non-random arrangement of probe sequences on the chip: Firstly, the vertical position of perfect match probes (PM) alternates with that of paired mismatch (MM) probes. The intensity of the former ones exceeds that of the latter ones on the average due to their altered middle base which mismatches the target. Secondly and more importantly, the probe sequences arrange in rows with respect to short motifs. In particular, the position of most of the probes possessing triple degenerated guanines at the solution end of their sequence ($(GGG)_1$) are found within a horizontal band which exactly matches the brightest stripe of the chip image. The respective intensities exceed the average intensity level of the array typically by a factor of two to ten. It seems unlikely that these strong intensity values are associated with extraordinary large expression levels of the respective target genes. Instead the bright intensities can be attributed to probe effects which typically reflect the sequence specifics of probe/target interactions [54]. Such probe effects must be removed from the data to obtain accurate expression values.

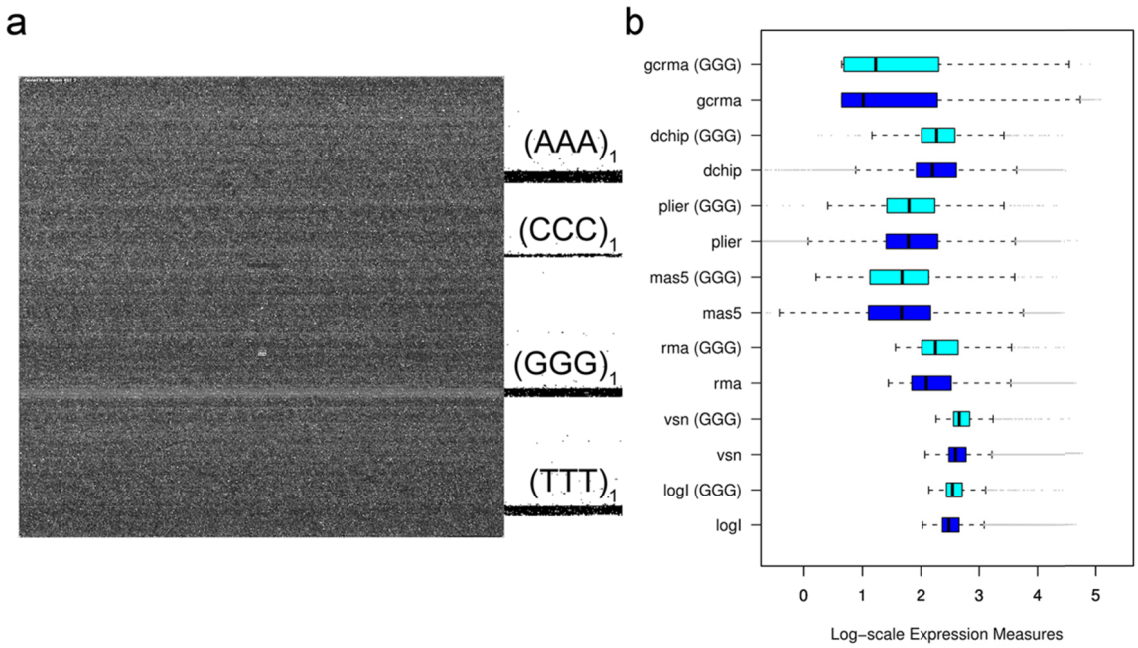

Figure 6.1: (a) Fluorescence image of a hybridized Affymetrix GeneChip Mouse Genome MG430 2.0 array (GEO GSE12545). The chip surface divides into a grid of nearly 10^6 fluorescing probe spots. The bright horizontal stripe matches with probes the 25meric sequence of which starts with triple degenerated guanines ($(GGG)_1$ motif). Triple runs of other nucleotides are not associated with bright srtipes. The position of the respective probes are shown at the rigth border of the figure. (b) Boxplots of expression measures obtained from the intensity data shown in panel a using various preprocessing methods (see text for assignments). The boxplots are computed separately for all probe sets (45,100) and for probe sets with at least two probes containing the $(GGG)_1$-motif (836, i.e. 2% of the total number). `log I' denotes the distributions of raw intensity data. Note that essentially all methods except Plier and partly mas5 are unable to correct expression values for the $(GGG)_1$-bias. The respective distributions of probe sets containing $(GGG)_1$-probes are systematically shifted towards larger expression values compared with the distribution of all probe sets.

The obligate correction and summarization of raw intensities into expression values prior to downstream expression analysis is called calibration or preprocessing. Numerous preprocessing algorithms are presently available to transform raw intensity data into expression measures (for example, vsn [93], RMA [94] and gcRMA [35], dChip [95], MAS5 [96], Plier [97]). Fig. 1b shows the distribution of expression values obtained after calibration of the intensity data shown in panel a using different preprocessing methods. The boxplots are calculated alternately either for all probe sets of the array or for probe sets which contain at minimum two $(GGG)_1$ probes. The results obtained from most of the preprocessing methods clearly reveal a systematic shift of the expression values of this $(GGG)_1$ sub-ensemble to higher levels. These calibrations obviously fail to correct the strong intensity bias properly.

As one option to solve this problem one can simply exclude the 'bad' $(GGG)_1$ probes from further analysis. However, we show below that also other motifs, for example runs of degenerate guanines along the whole sequence, can cause systematic intensity biases. The masking of such 'bad' probes will exclude a significant fraction of the available intensity

data from expression analysis and thus reduce the information potentially available from the microarray experiment. We suggest therefore an alternative strategy which intends to correct the probe-related intensity effects. It aims at extracting the 'hidden' information about target abundance in terms of corrected intensities for further use in downstream analysis.

This section addresses this issue and presents a systematic study of the effect of sequence motifs on the probe intensities. We will here focus on short motifs of up to four adjacent nucleotides at all possible sequence positions. Other approaches focusing on longer motifs require combining data from many microarray hybridizations [98]. We here compare the results obtained for different hybridizations after variation of the sample RNA, chip type and/or the amplification protocol. Our approach aims at identifying the minimum motif length for appropriate intensity prediction using a positional and motif dependent model. We focus on the effect of runs of degenerated guanines which have been found to behave unusually compared with other motifs in different chip assays including Affymetrix expression and SNP arrays [29, 99–103].

6.1.1 Used expression data

We here investigate various data sets dealing with different generations and types of Affymetrix GeneChip arrays which were taken from the public Gene Expression Omnibus (GEO) data repository (www.ncbi.nlm.nih.gov/geo/). The central examples are summarized in Table 6.1: (i) Human Genome HG U133A arrays taken from the 'HG133A_S' dataset were reanalyzed to verify the effect of G-stacks reported recently [103]. (ii) Identical human reference RNA was hybridized to both HG U133A and HG U133 Plus 2.0 arrays in the ' HG133P_Z' and 'HG133A_Z' datasets [104]. The latter arrays offer smaller feature sizes (11 versus 18 microns) and a larger number of probe sets (54.675 versus 22.300). All probes of the HG U133A are replicated on the the HG 133 Plus 2.0 array allowing direct comparison of the signal response of identical probes upon hybridization with the same RNA. (iii) In the 'Mouse' dataset we analyzed Mouse Genome 430 arrays referring to the same generation as the HG U133 Plus 2.0 array. (iv) The 'ENCODE'-dataset comprises human tiling arrays taken from the ENCODE-project [105]. This array-type not only contains a further increased number of probes but also uses different hybridization and labelling chemistries compared with the expression arrays of the other data sets. Particularly, cRNA-targets are replaced with cDNA targets and nucleotide-labelling throughout the sequence is changed into end-labelling. Arrays of the ENCODE type can also be applied in ChipChIP experiments with altered amplification protocols to explore protein/DNA interactions. We included ChipChIP data to study the effect of the amplification protocol.

Table 6.1: Chip characteristics of selected data sets studied.

Data set	HG133A_S	Mouse	ENCODE	HG133P_Z	HG133A_Z
GEO[a]	GSE1133	GSE12545	GSE6292	GSE3061	GSE3061
Chip type	HG U133A	MG 430 2.0	Human Tiling	HG U133plus2	HG U133A
# probes × 10^6 [b]	≈ 0.5	≈ 1.0	≈ 1.5	≈ 1.2	≈ 0.5
# probe sets[b]	22,300	45,101	300,000[c]	54,675	22,300
% absent[d]	61.9%	63.1%	94.8%	54.9%	42.8%
<logN>chip[e]	2.0	2.3	1.1	1.94	2.09
<log M>[e]	4.48	4.71	3.45	4.32	4.45
%$(GGG)_1$ probes[f]	2%	1.9%	2%	2%	2%
%$(GGG)_1$ probe sets[f]	20%	19%	-	20%	20%

a Gene Expression Omnibus (GEO) accession number
b number of probes and of probe sets per array
c pseudo sets are assembled using five consecutive probes
d percentage of absent probes per array
e mean value of the logged non specific background intensity and logged saturation intensity
f percentage of probes containing the (GGG)1-motif and of probe sets containing at minimum one of these probes

6.2 Positional-dependent sensitivity profiles

We apply the positional dependent sensitivity model to the intensity data shown in Figure 6.1a. The model provides sensitivity profiles of rank 1-4, the maximum rank being limited by the available number of data points. Figure 6.2 (left part) shows the profiles which were obtained using the intensities of 'absent' called PM probes. The sensitivity terms can be interpreted as the logged intensity increment due to the respective sequence motif of r consecutive bases starting at position k of the 25meric sequence (see subsection 3.4).

The shapes of the four single base profiles (r = 1) virtually agree with previously published data [36, 50, 53, 106, 107]: The sensitivities of adenines (A) and cytosines (C) are roughly symmetrical with respect to the x-axis and change in a parabola-like fashion, the maximum being near the middle of the probe sequence. The profiles of guanine (G) and thymine (T) indicate a more monotonous dependence. All profiles are asymmetrical with respect to the ends of the probe sequence: They converge towards the surface-attached side at k = 25 but differ significantly near the solution end at k = 1. The sensitivities and thus the base- and positional dependent contribution to the intensities increase according to A < T < G < C for most sequence positions.

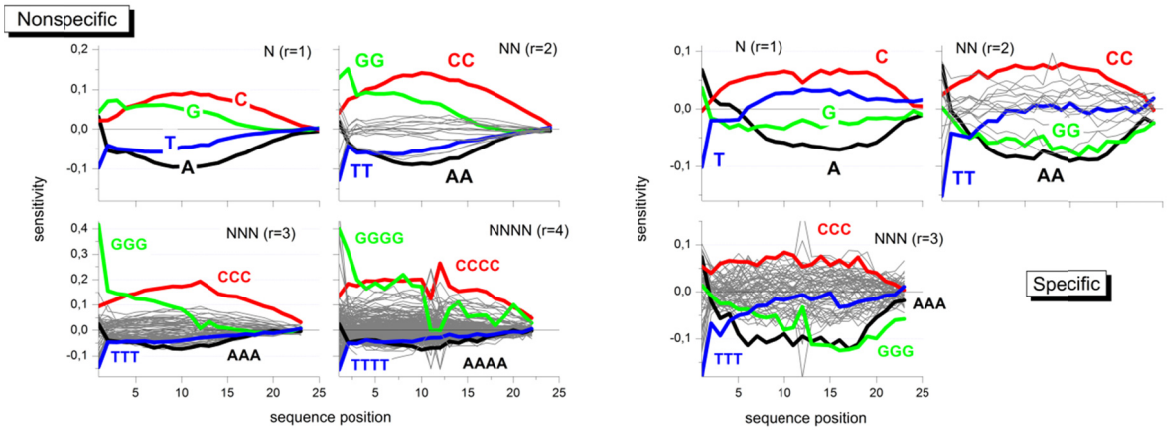

Figure 6.2: Positional-dependent sensitivity profiles of different rank for non-specific (left) and specific (right) hybridization computed from the Mouse-dataset shown in Figure 6.1. Runs of equal bases (e.g. AAA) are emphasized by thick lines in different colors. Note the different ranges of the ordinate scales in both rows of figures.

The nearest neighbor model (r = 2) provides a total of 16 profiles. Most of them relatively tightly group about the x-axis resembling essentially the parabola-like shape of the single base profiles for A and C. The contributions of the CC-profile however markedly inflates for all sequence positions whereas the GG-profile increases especially at small k. This latter trend is partly counterbalanced by negative values of T-containing NN-terms, especially of 'TT'.

Inspection of the 384 NNN-profiles (r = 3) shows that these trends further intensify for stacks of cytosines ('CCC') and guanines ('GGG') where the relative level of the latter profile increases compared with that of 'CCC'.

G-quadruples clearly dominate at small position indices k < 4 among the 1512 quadruple profiles of rank r = 4. Also other motifs indicate a relatively strong contribution at small k as well (e.g. 'GCCC' and 'GGGA'). The contribution of 'GGGG'-quadruples (and of other triple-G containing motifs) markedly drops for k > 13, i.e. for positions closer to the surface end of the probes. Note also, that the parabola-like shape of the profile of runs of adjacent cytosines changes into a broad plateau which decreases only near the ends of the probe sequence.

Hence, the contribution of a few motifs, especially of degenerated runs of C and G but also of selected 'GC'-rich tuples, increases above average with the extension of the model rank from r = 1 to r = 4: more than twofold for CCCC and up to tenfold for GGGG compared with the respective single nucleotide values. Longer homo-motifs obviously adapt to specific intensity effects.

The sequence effect of some of the motifs reaches its maximum in the middle of the sequence. With increasing model rank, these peaks reshape into broad plateaus of virtually

constant sensitivity values which markedly change only near the ends of the probe sequence. In contrast, G-rich subsequences add strong intensity contributions at small position indices especially at the first sequence position. The respective contributions progressively increase for r = 1 to 3 but then remain virtually unchanged for r = 4. Note that also the guanine profiles of lower rank (G, GG and GGG), show exceptional large positive values at sequence positions k < 4. The possible origin of this behavior will be discussed below.

The right part of Figure 6.2 shows the corresponding profiles of the PM probes predominantly with specific hybridization. Only 8% to 20% of all probes on the chip meet this criterion. This relative small number of probes restricts the rank of the model to r = 1-3 and, moreover, gives rise to a relatively large level of noise. The specific profiles possess essentially the same properties as the non-specific ones shown in the left part of Figure 6.2 except that for G- and T-rich motifs. In particular, profiles of homo-runs of guanines shift markedly towards smaller values compared with their non-specific values. Note also that the $(GG)_1$ and especially $(GGG)_1$ motifs at the solution end contribute much less to the specific profiles.

6.3 Guanine effects

6.3.1 Sequence motif assessment

We assume that a model of rank r applies with different quality to different sequence motifs of length s at position k, $(b_s)_k$. Note that the length of the motif s is independent of the rank of the model. For example, triple motifs (s = 3; e.g., GGC) can be analyzed either using the nearest neighbor model (r = 2; i.e., GG+GC) or the next-nearest neighbor model (r = 3; i.e., GGC). To assess the fit quality in a motif specific fashion we collect all probe sequences which contain $(b_s)_k$ into class $p((b_s)_k)$ with $\#p((b_s)_k)$ members per chip and define the motif-specific SSR in analogy with Eq. (3.10)

$$\mathrm{SSR}(r,(b_s)_k) = \frac{1}{\#p((b_s)_k)} \sum_{(b_s)_k} \mathrm{RES}^2 = \langle \mathrm{RES}^2 \rangle_{(b_s)_k} \tag{6.1}$$

One can subsume all motif effects independently of their position by substituting $(b_s)_k \to b_s$ in Eq. (6.1) to get the total SSR of tuple b_s, $\mathrm{SSR}(b_s)$.

Note that the total SSR (Eq. (3.10)) is given as the weighted sum of the motif-specific SSR

$$SSR(r) = \sum_{(b_s)_k} f_{(b_s)_k} \cdot SSR(r,(b_s)_k), \tag{6.2}$$

where $f_{(b_s)_k} = \#p((b_s)_k)/\#p$ denotes the fraction of probes containing the respective motif.

6.3.2 Quality of fit and standard error

The positional and motif-specific SSR (Eq. (6.1)) estimate the contribution of a subensemble of probes containing the motif $(b_s)_k$ to the total sum of squared errors after fitting the positional dependent sensitivity model of rank r to the whole ensemble of considered probes (Eq. (6.2)). Ideally, the residuals scatter with equal variance and center zero for each chosen motif. To detect and to estimate systematic biases of the fits in a motif specific fashion we calculate the squared sum of the respective residuals to judge the quality of the fits for each considered sequence motif,

$$QF(r,(b_s)_k) = \left(\frac{1}{\#p((b_s)_k)} \sum_{(b_s)_k} RES\right)^2 = \langle RES\rangle^2_{(b_s)_k}. \tag{6.3}$$

Ideally one expects $QF(r,(b_s)_k) = 0$ for centered distributions of the residuals. Non-zero values $QF(r,(b_s)_k) \neq 0$ thus indicate systematic deviations of the fits of the model of rank r with respect to motif $(b_s)_k$.

The motif-specific variance of the residuals and the respective standard error are given by $Var((b_s)_k) = \langle RES^2\rangle_{(b_s)_k} - < RES\rangle^2_{(b_s)_k}$ and

$$SE((b_s)_k) = \sqrt{Var((b_s)_k)/\#p((b_s)_k)}. \tag{6.4}$$

The standard error allows to estimate the confidence level of the positional dependent sensitivity terms $\sigma_k(b_s)$.

6.3.3 Triple guanine motif causes large intensities

Part a of Figure 6.3 compares the sensitivity profiles of non-specifically hybridized probes of the mouse data set shown in Figure 6.2 with the respective profiles of the ENCODE and HG133A_S data sets. As a general trend, the sensitivity level of poly-C terms nearly linear increases with increasing rank of the model as indicated by the dotted lines. This trend reflects a constant incremental contribution per additional cytosine in the considered motifs. In contrast, the sensitivity of poly-G motifs starting at k = 1 steeply gains at r = 3 (ENCODE and mouse data sets) or, to a less extend, at r = 4 (HG133A_S data set).

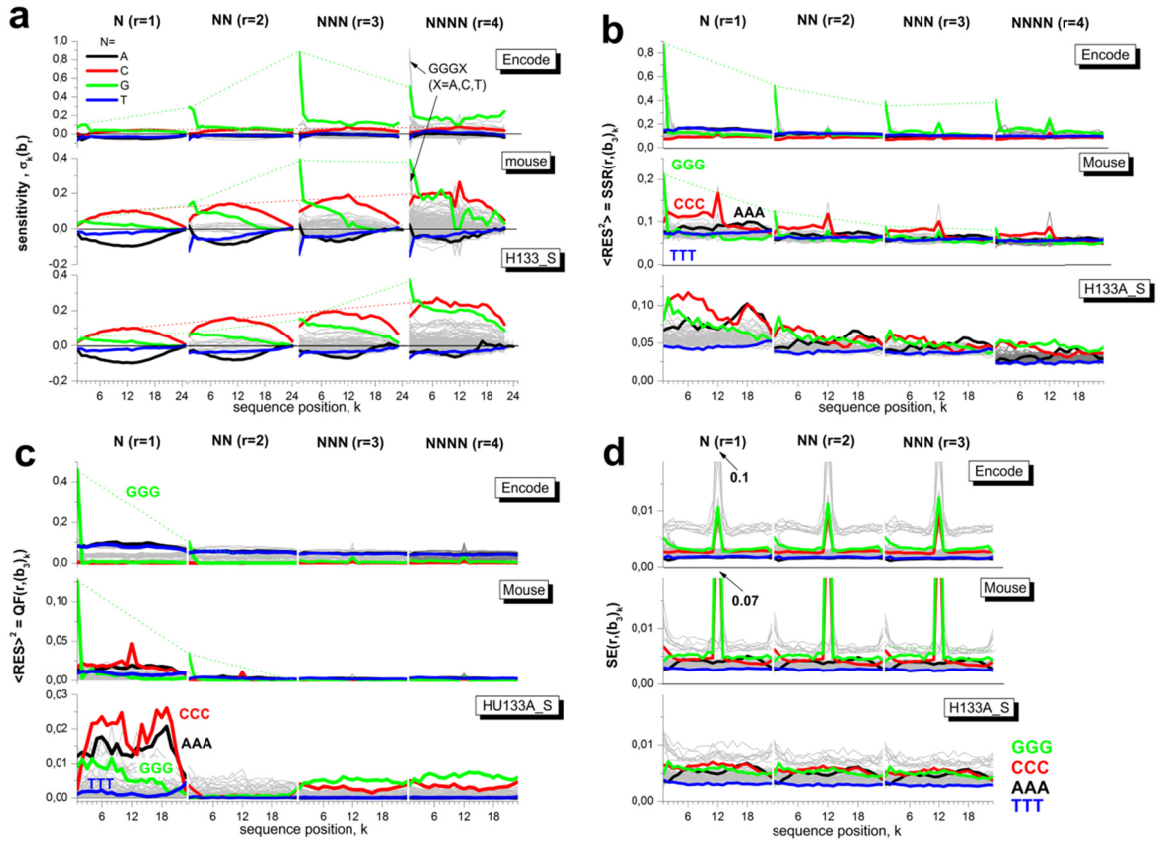

Figure 6.3: Sensitivity profiles of rank r = 1-4 of different data sets (panel a, see Table 6.1) and the respective triple-related fit statistics: sum of squared residuals (panel b, Eq. (6.1)), quality of fit (c, Eq. (6.3)) and standard error (d, Eq. (6.4)). Homo-motifs of consecutive A, C, G and T are shown by colored curves. The thin dotted lines indicate the basic trends of the poly-G and poly-C motifs at position k = 1 and k = 12, respectively. Note the different scaling of the ordinates in panels a and c.

We re-plot the respective sensitivity values in the left part of Figure 6.4. They reflect an extraordinary strong intensity increment due to three consecutive guanines starting at the first sequence position in the former situation; and four consecutive guanines along the probe sequence in the latter situation. We will call these properties shortly $(GGG)_1$- and poly-G-effect, respectively.

The $(GGG)_1$-effect is further supported by similar values of the sensitivity terms for quadruples starting at k = 1 with threefold degenerated guanines GGGB (B = A,T,G,C; see the arrows in Figure 6.3 and that of the respective triple-G, i.e. $\sigma_1(GGG) \approx \sigma_1(GGGB)$ (see Figure 6.4). It shows that the $(GGG)_1$-motif adds the dominating intensity contribution to that of the GGGB-quadruples.

Note that the $(GGG)_1$-effect of the ENCODE-data set largely exceeds that of the mouse data set by nearly one half order of magnitude: An initial run of three G increases the intensity relative to the mean intensity level by the factor of $10^{1.0} = 10$ and $\sim 10^{0.4} = 2.5$ in the former and latter data set, respectively. The intensity increment due to a triple-C motif in the middle of the probe sequence is distinctly smaller and amounts to a factor of

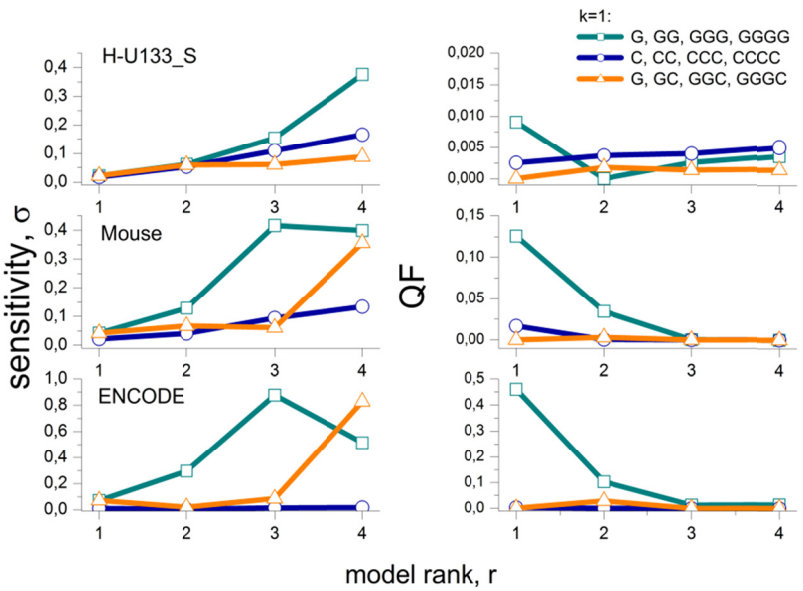

Figure 6.4: Sensitivity terms and quality of fit of selected motifs of rank r at position k = 1 of the probe sequence. The data are replotted from Figure 6.3 part a and c. The fact that the sensitivity of degenerated G levels off for r > 2 whereas that of GGGC steeply increases for the ENCODE and mouse data set indicated the strong $(GGG)_1$-effect. The respective quality of fit reaches acceptable values only for r > 2 which indicates that at least runs of three guanines must be explicitly considered.

about $\sim 10^{0.2} = 1.6$. Subtle differences between the sensitivities due to the different hybridization chemistries (DNA/RNA versus DNA/DNA in the mouse and HG133A_S sets versus ENCODE) will be discussed separately below.

In summary, triple degenerated guanines at the solution end of the probe sequences cause exceptionally large intensities in selected data sets. Longer runs of consecutive G along the probe sequence are also associated with large intensities, however to a smaller extend.

6.4 Quality of motif-specific fits

6.4.1 Model-rank assessment with the F-test

The number of independent parameters of the positional dependent sensitivity model increases with the rank according to

$$\#\sigma(r) = (4^r - 1)\cdot(25 - r + 1) + 1 \qquad (6.5)$$

providing $\#\sigma(r) = 76, 361, 1450$ and 5611 for $r = 1...4$, respectively.

The significance of increasing the rank ((r-1) → r) of such nested models can be tested using the F-statistics

$$F(r) = \frac{SSR(r-1) - SSR(r)}{(df(r-1) - df(r))(SSR(r-1) - SSR(r))}. \qquad (6.6)$$

It follows the F-distribution with the degrees of freedom $df(r) = \#p - \#\sigma(r) + 1$ and allows to estimate the significance of model extension in terms of a p-value. Usually one gets $df \cong \#p$ because the number of probes ($> 10^5$) largely exceeds the number of model parameters ($< 10^3$). One consequence of the large number of probe values is that essentially each improvement of the fit with $F > 1.5$ is judged as significant with $p < 10^{-2}$ for $df > 10^5$.

Eq.(6.6) applies under the assumption of normally distributed, independent residuals. We found that systematic errors partly contribute to the estimated SSR questioning the applicability of the F-test. We therefore use the F-values as a simple empirical measure characterizing the improvement of the fits.

Motif-specific F-values $F(r,b_s)$ and $F(r,(b_s)_k)$ can be calculated for the respective SSR and with the respective substitution for the number of probes ($\#p \rightarrow \#p(b_s)$; $\#p((b_s)_k)$) to judge the improvement of the model with respect to the chosen sequence motif. The number of relevant parameters is given by the number of model tuples b_r required to describe the sequence motif b_s at all positions for the positional independent case. It provides $\#\sigma(b_s) = (s-r+1)\cdot(25-s+1)$ and $\#\sigma((b_s)_k) = s-r+1$ for the positional dependent and indepentent cases, respectively.

6.4.2 Motif-specific differences

The discussed sensitivity profiles are obtained by multiple linear regression fits of Eq.(3.9) to the intensity data of non-specifically hybridized probes of the respective arrays by minimizing the total sum of squared residuals (SSR) (see Eq. (3.10)). The fit of models of increasing rank r improves the goodness of fit in terms of the total SSR(r) (Eq. (3.10)). Table 6.2 lists the total SSR(1) values of the single base model and the respective F-values for models of rank r = 2-4 (Eq.(6.6)). Maximum improvement is observed for the NN model compared to N and smallest improvement for NNNN compared to NNN.

The total SSR was decomposed into motif and positional dependent terms according to Eq.(6.2)) to characterize the model fits of rank r = 1-4 in more detail (Figure 6.3). In general, the mean level of the SSR-terms decreases with increasing rank of the model indicating the improvement of the fits in parallel with the decrease of the total SSR discussed above. The partial SSR values of selected motifs (e.g. degenerated cytosines and

Table 6.2: Sum of squared residuals of the fits of model ranks r = 1 ... 4: SSR of all probes and of probes containing C-triples and the $(GGG)_1$-motif are given for the N-model (r = 1). The respective F-values for the higher ranks r = 2 - 3 evaluate the improvement of the fits with respect to the model of next smaller rank r - 1.

	HG133A_S	**Mouse**	**ENCODE**
SSR(1)	0.048	0.072	0.11
SSR(1, CCC)	0.072	0.088	0.094
SSR(1, $(GGG)_1$)	0.071	0.21	0.85
F: N → NN	147.52	202.08	381.73
F: NN → NNN	11.11	20.6	78.07
F: NNN → NNNN	3.09	6.16	8.89

guanines) are larger than the average level for the N-model. Especially the value of the $(GGG)_1$-motif largely exceed the total SSR value by nearly one order of magnitude (ENCODE) and by the factor of 2 - 3 (mouse data set) indicating inadequate fitting of this motif (see Table 6.2 and Figure 6.4).

The SSR-values estimate the deviation between the fitted and the experimental data. They can be attributed to two potential origins, namely the systematic bias due to the inadequacy of the model and/or the random scattering of the experimental data. We calculate motif- and positional dependent profiles of the qualitiy of fit (QF, Eq. (6.3)) and of the standard error (SE, Eq. (6.4)) as suited measures to estimate the respective contributions. Particularly, one expects vanishing QF-values for adequate fits of the model. The motif and positional data shown in part c of Figure 6.3 reveal that the N-model fails fitting the probe intensities of all considered data sets. The NN-model markedly improves the fit for all motifs except $(GGG)_1$. Clearly this motif gives rise to residual systematic deviance between the fits and the respective intensities of the mouse and ENCODE data sets. It however largely vanishes for r = 3. This result confirms our hypothesis that the observed intensity effect is related to threefold degenerated guanines $(GGG)_1$. The QF-profiles of the HG133A_S data set reveal small systematic deviations of degenerated guanines motifs along the whole sequence for r = 3 and 4 due to the poly-G-effect.

The standard error is relatively invariant for most of the motifs and positions with $SE < 0.01$ as a rule of thumb (part d of Figure 6.3). A notable exception are selected GC-rich motifs in the middle of the probe sequence which show high standard errors up to $SE \approx 0.1$ for the ENCODE data set and up to $SE \approx 0.07$ for the mouse data set, respectively. Figure 6.5 shows that these motifs are very rare on the MG230 2.0 and ENCODE arrays with partly less than 100 probes containing them. These small numbers gives rise to imprecise estimates of the respective sensitivity terms.

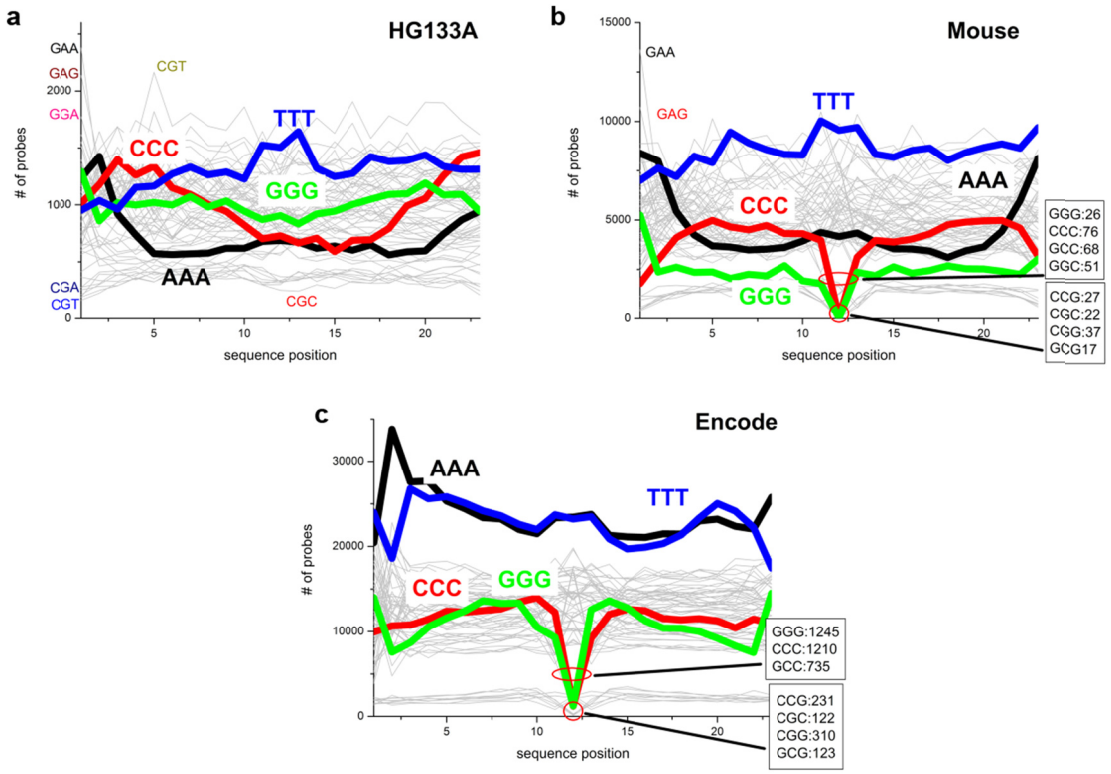

Figure 6.5: Frequency of triple motifs. The figures show the positional dependence of the triple motifs in the probe sequences of three selected array-types. Homo-motifs are highlighted by thick colored curves. Note the partly different scaling of the ordinates. Rare triple motifs at position k = 13 are explicitly given in the boxes together with the respective number of probes containing the motif. For example, only 27 probes on the MG-430 2.0 array contain CCG-triples starting at k = 13. Note that these rare motifs give rise to large spikes of the respective standard errors for the triple terms.

In summary, the decomposition of the total fit statistics into motif- and positional dependent contributions reveals adequate fits of most of the motifs using the NN-model. As a clear exception, the $(GGG)_1$ effect requires explicit consideration of NNN-terms for adequate fitting.

6.5 Chip-type and target effects

The data sets so far address different target samples which are hybridized onto different chip types. Both factors potentially affect the motif and positional dependent sensitivity profiles, and, in particular, the poly-G effects discussed above. To discriminate between effects due to target and chip-type we compare the sensitivity profiles for different hybridizations of the same RNA sample (Universal Human Reference RNA) to two different chip types, namely the newer HG133P_Z and the previous-generation HG133A_Z. The nearly 55.000 probe sets of the former chip integrate more than 22.000 probe sets of the latter one and this way allow direct comparison of the intensity of probes of identical sequences on the two chip types after appropriate masking of the additional

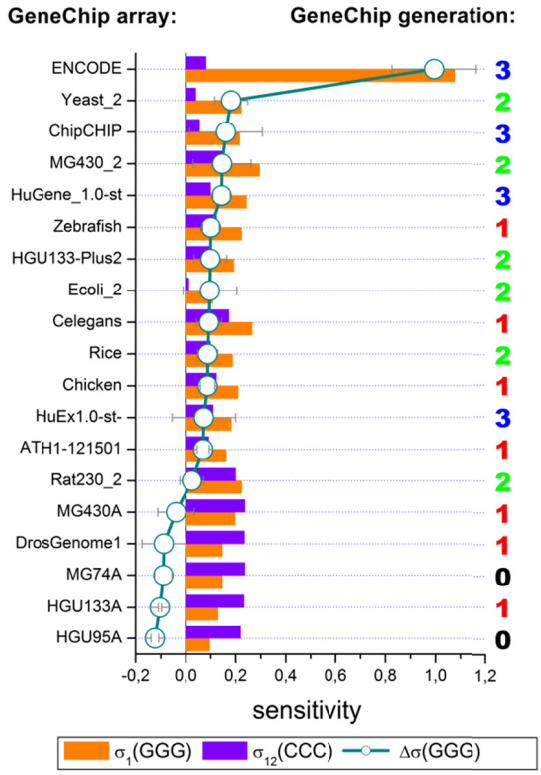

Figure 6.6: The amplitude of the $(GGG)_1$-effect on GeneChips of different type. The bars refer to the sensitivity terms of triple-G and -C at the first and the middle position of the sequence, respectively. The arrays are ranked with respect to the difference Δσ (GGG) which characterizes the amplitude of the $(GGG)_1$-effect (circles, see text). Sensitivity profiles of three independent hybridizations are averaged for each value. The numbers on the right assign the chip generation 0 to 3 (see text). The amplitude of the $(GGG)_1$-effect tends to increase with the chip generation. The GEO-accession numbers of the samples analyzed are given in the Appendix A.

probes in the HG133P_Z data set. The obtained three sets of profiles of rank r = 1 - 4 are very similar and provide no indication that the two considered chip types strongly modify their shape (see Figure 6.6a). For example, the poly-G effect is observed in all three data sets.

In the next step we compare the profiles of different RNA-hybridizations to the same chip type (MG430 2.0, see Figure 6.6b). Also in this case the profiles of most of the motifs look similar for the different hybridizations except the sensitivity terms of homo-G runs at the first sequence position which indicate different amplitudes of the $(GGG)_1$-effect. For direct comparison we normalize the respective triple sensitivity term with respect to the maximum sensitivity value of triple-C motifs in the middle of the sequence and calculate the difference $\Delta\sigma(GGG) = \sigma_1(GGG) - \sigma_{12}(CCC)$ as a relative measure of the amplitude of the $(GGG)_1$-effect (see Figure 6.6b for illustration). Part c of Figure 6.6 shows the distribution of the obtained Δσ(GGG)-values for a series 29 independent hybridizations using MG430 2.0 arrays. The data show that the amplitude of the $(GGG)_1$-effect varies over a wide range for different target hybridizations of the same chip type.

In the next step we estimated the amplitude of the $(GGG)_1$-effect for eigtheen different array types. Figure 6.6 plots the mean sensitivity amplitudes σ_1(GGG), σ_{12}(CCC) and their difference. The considered chip types can be roughly classified into four chip-generations (numbered 0 to 3) which use different probe spot sizes, number of probe spots per chip and partly different hybridization chemistries. The spot sizes decrease from 18-20 µm (generations 0 and 1), 11 µm (generation 2) to 5 µm (generation 3) which results in the marked increase of the number of probes per chip. Generation 3 (Human Gene 1.0 ST and Human Exon 1.0 ST arrays) uses a PM-only design without MM-probes and DNA/DNA instead of DNA/RNA hybridization chemistry. We assign the ENCODE arrays also to generation 3 because it applies DNA/DNA hybridizations as well. However, it still uses MM probes and larger spot sizes (10 µm) compared with Gene 1.0 ST and Exon 1.0 ST arrays. `ChipChIP' assigns arrays of the ENCODE-type which are applied in ChipChIP experiments. On chips of generations 1 – 3 most of the probes containing $(GGG)_1$ motifs are located in a row as shown in Figure 6.1.

It turned out that the $(GGG)_1$-effect can be identified for all arrays of generations 1 to 3. Its amplitude tends to increase for chips of later generations 2 and 3. The differences between the chip generations are however moderate without clear indication that type-specific factors such as the arrangement of probes, their spot size, density and number explicitly explain the $(GGG)_1$-effect.

Interestingly, our data reveal a large difference of the amplitude of the $(GGG)_1$-effect between the ENCODE-expression and ENCODE-ChipChIP-hybridizations (Figure 6.6). Both experiments use the same type of ENCODE tiling arrays but different amplification protocols: The former one amplifies sample mRNA via T7-priming and subsequent reverse transcription to double stranded cDNA whereas the latter one amplifies genomic DNA after immunoprecipitation via random priming without the T7-protocol [105, 108–110]. Note that fragments of the T7-primers used in the amplification step of mRNA-sample preparation partly remain bound to the amplified targets as has been discussed in [111]. The respective common G-rich sequence motif of the primer (5'-GGGCGGAGG…) contaminates a large fraction of the targets at their 5'-end and preferentially bind to probes with complementary, C-rich motifs [111].

In summary, we found systematic differences between the amplitude of the guanine effects between GeneChips of different generations which are rather gradual than fundamental. On the other hand, our data suggest that the amplification protocol for the used targets strongly affects the $(GGG)_1$-effect. Previous studies showed that the targets become contaminated with G-rich primer fragments after T7 amplification. One might hypothesize that these fragments are prone to associate to selected G-rich probe sequences.

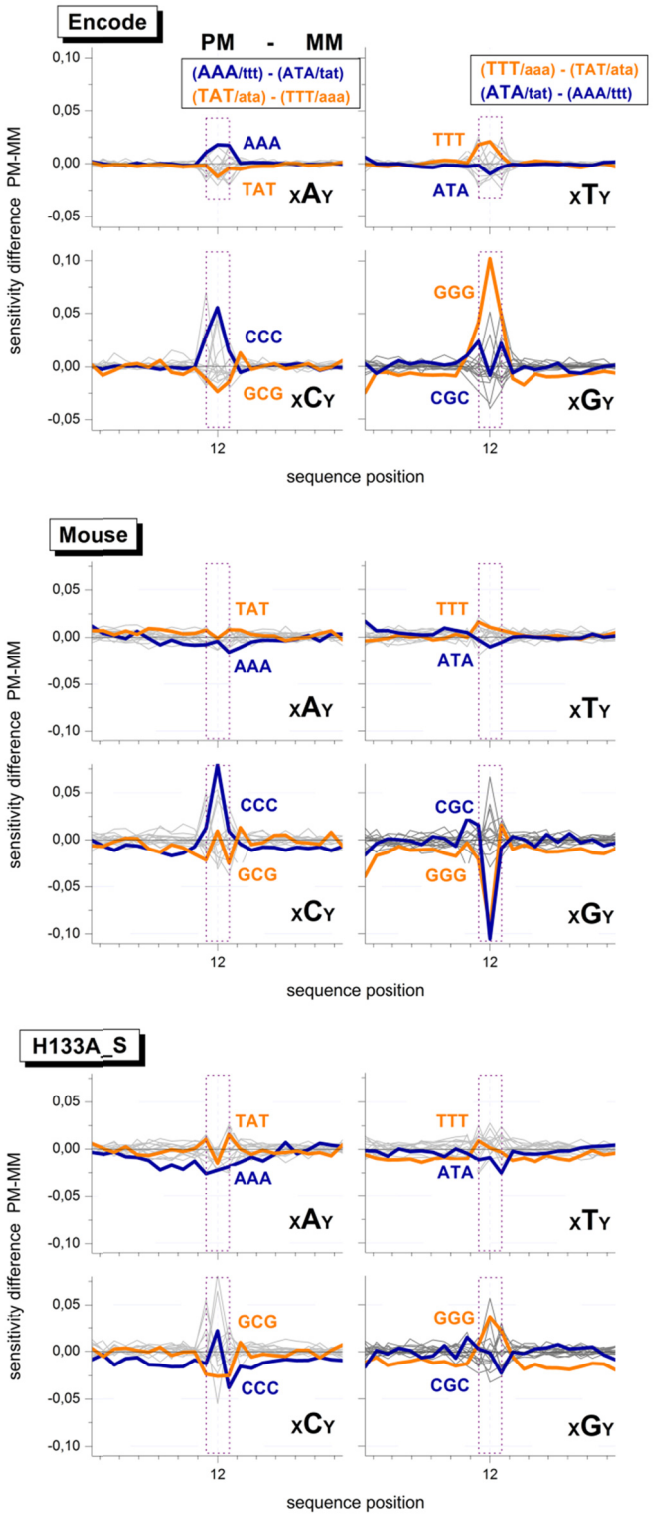

Figure 6.7: Sensitivity difference profiles obtained by fitting the positional dependent NNN model to the logged intensity difference of each probe pair, $\Delta = \log I^{PM} - \log I^{MM}$. The profiles are sorted according to the central base in the NNN-terms of the PM probes, xBy with B,x,y = A,C,G,T. Only profiles of triple degenerated homo-motifs and the respective 'mirror' motifs with complementary center base are highlighted by colored thick lines. The upper panel gives the respective base pairings in the duplexes of the PM and MM (upper and lower case letters refer to the probe and the target respectively). The dotted rectangles refer to the middle tripe (sequence position k = 11 ... 13). The triple terms within this range are different for PM- and MM-probes. Their amplitudes refer to the respective swap of the middle base at k = 13 in the sequence of the MM-probes. Note the symmetries of the obtained profiles for complementary degenerated triples. The PM and MM probe sequences are identical outside the middle range. The absence of large amplitudes indicates that the intensities of both, PM and MM probes, similarly respond to sequence effects.

6.6 Perfect match and mismatch probes

Each perfect match (PM) probe is paired with one mismatch (MM) probe on most of the Affymetrix microarray types. The MM probes use the same 25meric sequence as the respective PM probes except for the middle base, which is substituted by its complement. To extract subtle differences between the sensitivity profiles of both probe types we calculate the logged intensity difference of each probe pair, $\Delta = \log I^{PM} - \log I^{MM}$, and subsequently fit the NNN-sensitivity model of rank r = 3 to the intensity data of the three data sets given in Table 6.1.

The obtained terms characterize subtle intensity differences between both probe types in a motif- and position-dependent way. Their amplitudes virtually vanish for $k < 11$ and $13 < k$ (Figure 6.7). This result seems trivial because the sequences of PM and MM probes are identical at these positions. It clearly indicates, however, that the $(GGG)_1$- and poly-G effects apply to the PM and MM probes as well.

The NNN-sensitivity data markedly deviate from the baseline at positions k = 11…13 at which the triple motifs diverge between the PM- and MM-probes owing to the swapped middle base (in Figure 6.7 this range is indicated by the dotted vertical lines). Here we focus our discussion to triple degenerated homo-motifs in the middle of the PM- (or MM-) sequence which combine with motifs of broken degeneracy in the respective paired MM- (or PM-) sequence. For example, $(GGG)_{11}$ combines with $(GGC)_{11}$, $(GGG)_{12}$ with $(GCG)_{12}$ and $(GGG)_{13}$ with $(CGG)_{13}$. The calculated sensitivity amplitudes consequently characterize the logged intensity difference due to both motifs.

Figure 6.7 sorts the profiles with respect to the central base B of the middle triples in the PM sequence, xBy with B,x,y = A,C,G,T. The complete base pairings in the triple motifs are given in the figure. Base pairings in DNA/DNA duplexes are symmetrical with respect to bond reversal [112]. One expects therefore a central symmetrical pattern for the profiles of degenerated triples and the triples with swapped central base, e.g. AAA versus ATA and TTT versus TAT. The obtained sensitivity-profiles indeed show this symmetrical pattern. One expects also equal amplitudes for complementary homo-motifs, e.g. AAA and TTT. The observed effect however ranks according to AAA ≈ TTT < CCC < GGG. The slightly larger peak of $(GGG)_{12}$ compared with $(CCC)_{12}$ indicates the poly-G effect along the sequence.

The mouse and HG133A_S data sets refer to DNA/RNA hybridizations. The chemical asymmetry of base-pairings between the DNA probes and RNA targets (see, e.g., [113, 114]) explains the slightly modified pattern of the obtained triple motifs compared with that of the ENCODE data set. Particularly, one gets for the mouse data set

GGG ≪ AAA ≪ TTT ≪ CCC which is compatible with solution data (see also below). It therefore provides no indication of the poly-G effect. In contrast, in the HG133A_S data one observes the reversed relation for guanines and cytosines, GGG > CCC, which indicates a slightly larger intensity contribution of degenerated runs of guanines.

In summary, the joint analysis of the PM- and MM-intensities shows that both probe types are affected by the poly-G and $(GGG)_1$-effect to a similar extent. It also reveals a relatively large intensity contribution of poly-G motifs in the middle of the sequence in some cases. The amplitude of this effects is however relatively small compared with the $(GGG)_1$-effect.

6.7 Specific and non-specific hybridization

Our analysis so far mainly uses the positional sensitivity profiles of non-specifically hybridized PM probes and of the logged PM-MM difference. Selected profiles due to specific hybridization revealed a decreased sensitivity level of runs of degenerated guanines and, in particular, of the $(GGG)_1$ motif (see the right part of Figure 6.2 for the mouse data set, the specific profiles of the other data sets analyzed are given in the supplementary material). This result suggests that the $(GGG)_1$-effect is only weakly or even not at all associated with specific hybridization.

It should be taken into account, however, that the specific sensitivity profiles are relatively uncertain owing to incomplete correction for parasitic effects such as saturation of the probe spots and bulk hybridization which deform the shape of the profiles and shift their level against each other [54, 57, 115]. Moreover, the number of probes in the sub-ensembles of probes used for calculating the specific profiles are typically much smaller than that of the non-specific probes. In addition, the specific sub-ensemble of probes is typically contaminated with contributions due to non-specific hybridization. All these factors give rise to relatively noisy profiles which still reflect properties of non-specific hybridization.

We therefore apply a different approach to answer the question whether the $(GGG)_1$-effect extends also to specific hybridization or not. Part a of Figure 6.8 plots the smoothed probe intensities of the mouse data set as a function of the expression degree which was calculated using the hook method. This calibration approach inverts the two-species Langmuir hybridization isotherm and estimates the linearized intensity-equivalent due to specific hybridization $L^S = M \cdot X^S$ (see Section 3.1) using the respective raw intensity values. The graphs in Figure 6.8 thus characterize the mean dependence of the intensity as a function of the specific transcript concentration [S] which is directly related to L^S. These

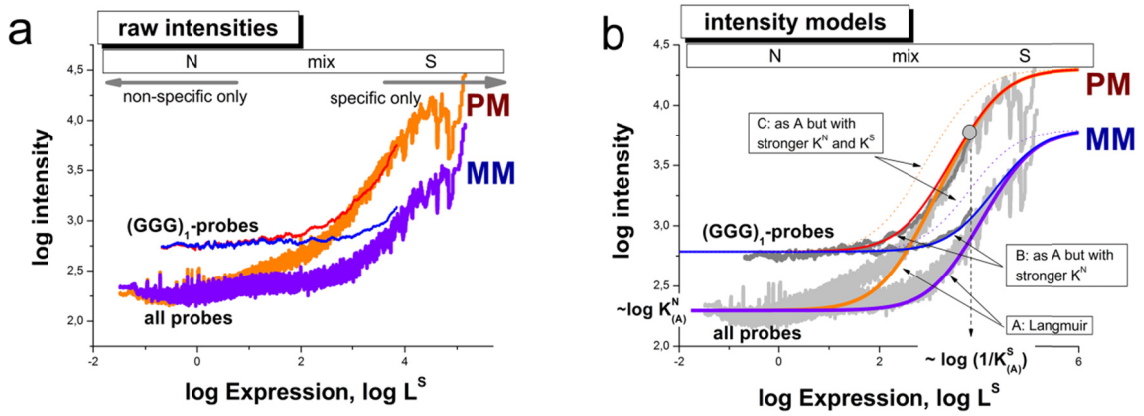

Figure 6.8: Hybridization isotherms of the mouse data set: The isotherms in panel a were calculated by plotting the probe intensities as a function of their expression value which is directly related to the concentration of specific transcripts, $L^S \propto [S]$. The data were subsequently smoothed over a moving window of 1000 probe intensities. The isotherms were calculated using either all PM- or MM-probe data of the chip or, alternatively, the sub-ensemble of probes containing the $(GGG)_1$-motif, i.e. a run of three consecutive guanines starting at the first sequence position. The horizontal bar in the upper part of the figure assigns the hybridization ranges (N, mix and S) which are described in the text. The arrows indicate the regions which are dominated either by specific or non-specific hybridization. Panel b shows theoretical isotherms which were calculated using Eq. (3.2) assuming three scenarios: (A) the reference situation describing the behavior of all probes; (B) stronger non-specific binding compared with A and; (C) stronger non-specific and specific binding compared with A (see also the text for details). Note that the intensity level in the N-range is directly related to KN, the mean binding constant of non-specific hybridization, whereas the position of the inflection point halfway between the N- and asymptotic saturation levels is inversely related to 1/KS, the mean binding constant of specific hybridization as indicated in the figure. The experimental data are compatible with stronger non-specific binding and invariant specific binding to (GGG)1-motifs compared with the respective main level of the binding strength of the array (scenario B). The MM-probes virtually behave like weak-affine PM probes with respect to specific binding.

isotherms roughly divide into the N-range which is dominated by non-specific hybridization at small abscissa values; into the S-range in which the intensity is dominated by specific hybridization at large abscissa values and into the mix-range in-between, in which both, specific and non-specific hybridization significantly contribute to the observed intensity (see also Figure 6.8 for assignment).

Figure 6.8 shows two different isotherms for the PM and MM probes each. One was calculated by averaging over all PM- (or MM) probes of the chip and the other one by selecting the respective sub-ensembles of probes containing the $(GGG)_1$-motif. In the N-range, the intensity level of the $(GGG)_1$-containing probes is clearly larger compared with that of all probes. The respective log-intensity increment of about 0.5 roughly agrees with the sensitivity amplitude of the $(GGG)_1$-motifs $\sigma_1(GGG) \approx 0.4$ (see Figure 6.2). The difference between both types of isotherms, however, progressively decreases with increasing expression degree and virtually vanishes in the S-range.

In panel b of Figure 6.8 we plot theoretical isotherms calculated using Eq. (3.2) with the substitution $L_p^P \rightarrow L^P = M \cdot (K^{P,S}[S] + K^{P,N}[N])$ as a function of the specific transcript concentration [S] for P = PM, MM. Three scenarios, (A) - (C), are considered to interpret the experimental data: (A) The `reference' case with a parameter set which was chosen to

fit the mean isotherms of the array averaged over all PM- or MM-probes; and scenarios (B) and (C) which aim at reproducing the behavior of the $(GGG)_1$-subensemble. Particularly, in scenario (B) only the value of the non-specific binding constant is increased compared with the reference case (A) according to $K^N_{(B)} = 10^{0.5} \cdot K^N_{(A)}$ whereas the value of the specific binding constant remained unchanged $K^S_{(B)} = K^S_{(A)}$. In scenario (C) also the value of the specific binding constant is increased by the same factor as K^N in case (B), i.e. $K^S_{(C)} = 10^{0.5} \cdot K^S_{(A)}$ and $K^N_{(C)} = K^N_{(B)}$.

Comparison of the theoretical and experimental curves clearly reveals that the intensity increment of the $(GGG)_1$-containing subensemble is readily described by the second case (B) which only assumes the stronger non-specific binding of the probes. Case (C) assumes also an increased specific binding. It clearly fails descibing the data: The inflection point of the calculated isotherms shifts to smaller abscissa values whereas that of the experimental isotherms remains roughly at the same position.

Hence, comparison between measured and calculated isotherms provides no indication that specific hybridization contributes to the $(GGG)_1$-effect to a similar extent as non-specific binding. Instead they show that the $(GGG)_1$-effect is mainly associated with non-specific hybridization.

The isotherms of the MM-probes are shown in Figure 6.8 together with the isotherms of the PM-probes. Both probe types are equally affected by non-specific hybridization on the average in both considered probe ensembles. Particularly, the $(GGG)_1$-motif increases the intensity level of the MM-probes in the N-range to the same extent as that of the PM-probes. The slight shift of the mix- and S-ranges of the MM-probes towards larger expression values is caused by the weaker specific binding of the MM due to their swapped middle bases which mismatches the target sequence. Hence, the MM-probes virtually behave like weak-affine PM-probes with respect to specific hybridization. This difference also implies that the mean saturation intensity of the MM-probes is smaller than that of the PM-probes owing to post-hybridization washing [43, 116, 117]. The calculated isotherms of the MM-probes clearly show that specific binding is virtually not affected by the $(GGG)_1$-motif by the same arguments as for the PM-probes.

6.8 Correction of microarray data for sequence effects

6.8.1 The NN+GGG hybrid rank model

Our analysis shows that the quality of fit of sequence models is heterogeneous with respect to the selected motifs and their position along the probe sequence. The positional dependent NN model well describes most sequence-dependent intensity effects due to non-

specific hybridization with the exception of motifs of three or more consecutive guanines. Higher order models of rank r = 3 or 4 are able to successfully remove the associated sequence bias. However they are computationally expensive. Minimization of the linear regression model Eq. (3.9) provides a system of $(4r - 1) \cdot (25r + 1)$ linear equations, the solution of which requires a runtime in the order of $O(\#p \cdot (4r)^2)$. In practice, profiles with rank up to r = 2 can be computed in minutes per array on a standard personal computer whereas models of rank r = 3 and 4 run hours or even days, respectively.

We therefore developed a hybrid-rank model based on the positional dependent nearest neighbor approach plus additional higher order contributions for selected `critical' motifs such as $(GGG)_1$ which applies to the intensity components due to non-specific binding. The algorithm fits the NN-model of rank r = 2 to all probes which do not contain the critical poly-G motifs in their sequence. The intensities of these probes is corrected according to Eq. (3.6). The intensities of probes which contain such motifs are separately fit to a NNN-model of rank r = 3 which only considers triple-G motifs at all possible sequence positions. In general, this approach can be modified to apply to other special motifs.

The algorithm works in detail as follows:
1) The set of predominantly non-specifically hybridized probe sets, the so-called `absent' or N-subset, is identified as described in Section 3.3
2) The N-subset is further split into two sub-ensembles not-containing and containing triple-G motifs, PS_{NN} and PS_{GGG}, respectively. They are subsequently corrected in two steps for sequence effects:
2a) The PS_{NN} sub-ensemble is used to train the NN model by multiple linear regression of the data using (3.8) - (3.10) with r = 2. The fit provides the basal set of NN-terms $\sigma^{NN} \equiv \sigma_k(b_2)$.
2b) Each probe set of the second PS_{GGG} sub-ensemble contains at least one probe with at minimum one motif of three consecutive guanines. Eq. (3.5) rewrites for these probes into

$$K_p^{P,h} = K_0^{P,h} \cdot \exp(\delta A^{NN,P,h}(\xi_p)) \cdot \exp(\delta A^{GGG,P,h}(\xi_p)) \tag{6.7}$$

where $\delta A^{NN,h}(\xi_p)$ is given by Eq. (3.6) with r = 2 and the set of NN-terms estimated in step 2a. The excess correction term $\delta A^{GGG,h}(\xi_p)$ considers the effect of the critical motif in the probe sequences in analogy with Eq.(3.7)

$$\delta A^{GGG,P,h}(\xi) = \sum_{k=1}^{23} \sigma_k^{P,h}(GGG) \cdot \delta(GGG, \xi^{k,k+2}) \tag{6.8}$$

With Eq. (3.8) one gets the theoretical sensitivity

$$Y^{theo} = Y_{NN}^{theo} + \sum_{k=1}^{23} \sigma_k(GGG) \cdot (\delta(GGG, \xi^{k,k+2}) - f_k(GGG)) \qquad (6.9)$$

Y_{NN}^{theo} denotes the basal sensitivity which is calculated using Eq. (3.9) and the basal set of NN-terms estimated in step 2a. After minimizing Eq. (3.10) one gets the profile of excess terms $\sigma_k(GGG)$.

3) The corrected intensities of the probes of the PS_{NN}- and PS_{GGG}-subsets are calculated after rearrangement of Eqs. (3.5) and (6.7), respectively.

4) The present probes not included in the N-sub ensemble are corrected as described previously [42, 45]. In short: A NN-model of rank r = 2 is parameterized using the probe sets which are hybridized to more than 80% with specific transcripts. They are then corrected using this model. Probe sets with a fraction of specific-hybridization of less than 80% are corrected by a weighted combination of the sensitivity profiles referring to specific and non-specific hybridization determined in step 2.

5) The sensitivity-corrected intensity data are exported in the standard *.CEL file format. The corrected signal values can then be feed into standard GeneChip preprocessing programs for further improvement and/or downstream analysis.

The correction algorithm is implemented in the *Larpack* program package which can be downloaded freely from the project website currently available under the URL www.izbi.uni-leipzig.de/downloads_links/programs/hook.php.

6.8.2 Effect of the correction

Figure 6.9 compares the performance of the hybrid rank correction with that of the N and NN models using the same type of representation as in Figure 6.8. It clearly shows that the latter two models only insufficiently correct the $(GGG)_1$-effect as expected. On the other hand, the systematic bias of the $(GGG)_1$-containing probes in the non-specific hybridization range almost completely vanishes after applying the NN+GGG correction to the non-specifically hybridized probes using the algorithm described in the previous subsection.

Residual profiles of the triple-G motifs of four different data sets are shown in Figure 6.10. They clearly reveal the strong intensity excess at position k = 1 due to the $(GGG)_1$-effect (mouse and ENCODE data sets). The mean level of the poly-G effect affecting the remaining sequence positions is about $\sigma_k(GGG) \approx 0.1$ for these chips. This excess sensitivity value refers to an intensity bias of $10^{0.1} \approx 1.25$ compared with the NN-model. Interestingly, hybridizations of ENCODE arrays using the ChipChIP technique indicate a negative GGG-level throughout the sequence for k > 1. This indicates an average intensity bias in the opposite direction of about $10^{-0.07} \approx 0.85$.

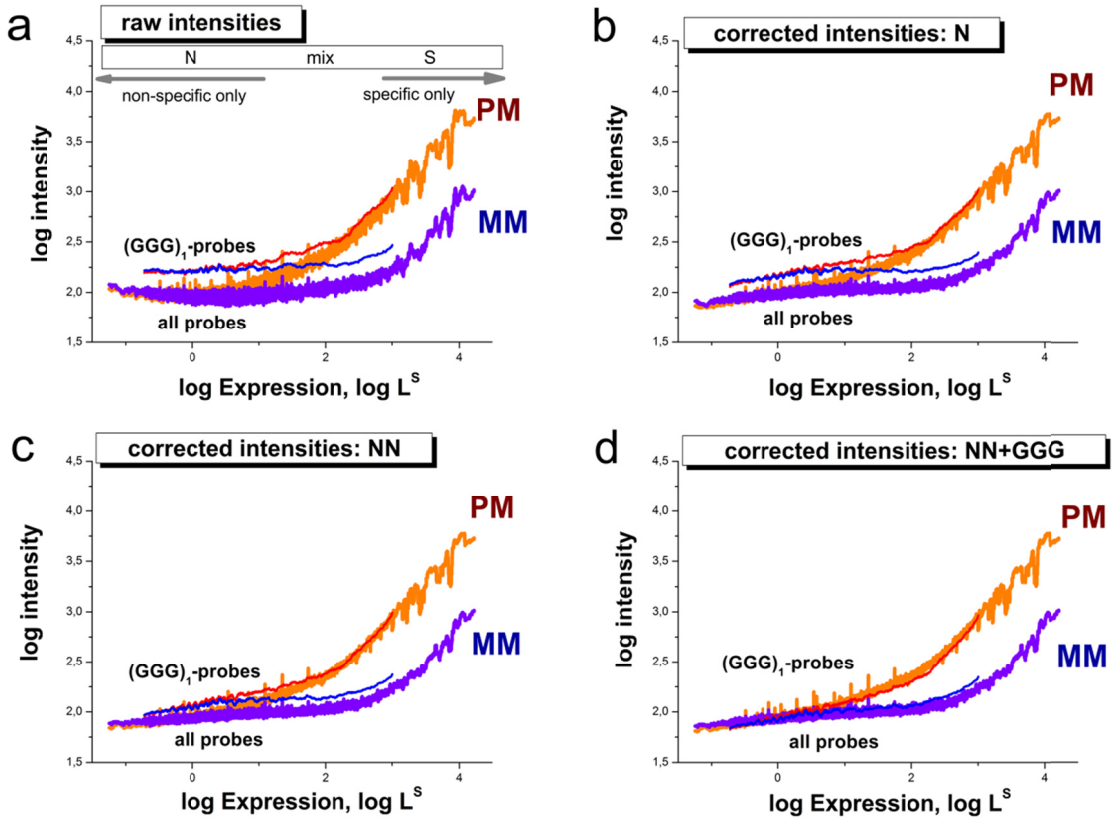

Figure 6.9: Correction of microarray intensity data using models of rank r = 1, 2 and the hybrid rank model NN+GGG for the non-specifically hybridized probes of the mouse data set. Specific hybridization is corrected using the NN-model in all cases. The figure shows the averaged intensity as a function of expression as in Figure 6.8. The systematic bias of probes containing the $(GGG)_1$-motif progressively decreases with increasing rank of the model and it virtually vanishes for the NN+GGG model. Correction using the NNN model provides a plot which is virtually indistinguishable from that of the NN+GGG model (not shown).

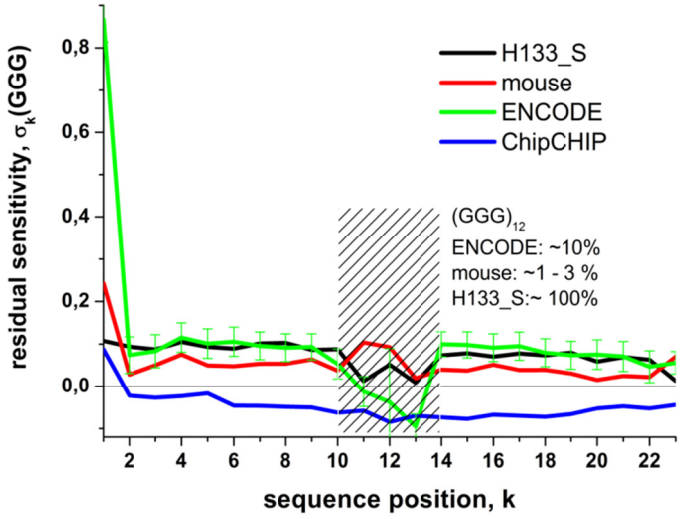

Figure 6.10: Positional dependent residual sensitivity profiles of triple-G motifs. The data clearly reveal the poly-G and the strong $(GGG)_1$-effect of the mouse and ENCODE data sets. The hatched region refers to sequence positions with very small numbers of probes containing the $(GGG)_{12}$-motif printed on the mouse and ENCODE arrays (see Figure 6.5). Interestingly, ChipChIP applications of the ENCODE arrays give rise to negative residual GGG-sensitivity values for most of the sequence positions.

We argued above that the ChipChIP targets lack G-rich primer fragments which otherwise cause the strong intensity bias due their involvement into G-stack formation on expression arrays. Their absence would explain a tiny or even zero but not a negative amplitude-level of the triple-G excess sensitivity. A similar negative sensitivity effect of poly-G motifs has been found for SNP GeneChip arrays [29]. These arrays use genomic DNA for hybridization after amplification via ligation and not via T7 priming [118]. This `dim' effect has been attributed to G-stack formation in agreement with previous assumptions [99, 119]. Such probe quadruplexes reduce the amount of free probe oligomers available for the binding of specific and non-specific targets. This trend then decreases the intensity of the respective probe spots because only targets are labeled with optical markers.

In summary, the NN+GGG hybrid-rank model properly corrects the intensity bias associated with probes which contain poly-G motifs. In addition, the obtained excess GGG-profiles provide further insights into the amplitude of the effects due to degenerated guanines in different hybridizations. It changes sign and switches from positive to negative values for hybridizations which use different amplification protocols.

6.8.3 Preprocessing of microarray intensity data

Calibration of microarray measurements aims at removing systematic biases from the probe-level intensity data to get expression estimates which linearly correlate with the transcript abundance in the studied samples. The performance of different preprocessing algorithms to correct intensity data for the $(GGG)_1$-effect are illustrated in Figure 6.1b by means of boxplots which roughly characterize the distribution of the expression values in terms of their median and interquartile range. The results revealed that the strong intensity effect is not removed from the expression data after standard preprocessing with several popular methods.

To get further insights we plot the density distributions of the preprocessed expression values of all 45,100 probe sets of the mouse data set and of the sub-ensemble of 836 probe sets containing at minimum two probes with a $(GGG)_1$-motif (Figure 6.11). The results indicate the systematic shift of the $(GGG)_1$ sub-ensemble towards larger expression values in decreasing order for the preprocessing methods vsn [93], RMA [120–122] and gcRMA [123]. Note that vsn and RMA use global baseline-corrections for non-specific hybridzation which subtracts one common background value from all probe intensities of a selected microarray. Clearly these approaches fail to describe the probe specifics of the $(GGG)_1$-motif giving rise to a strong bias due to improper background correction.

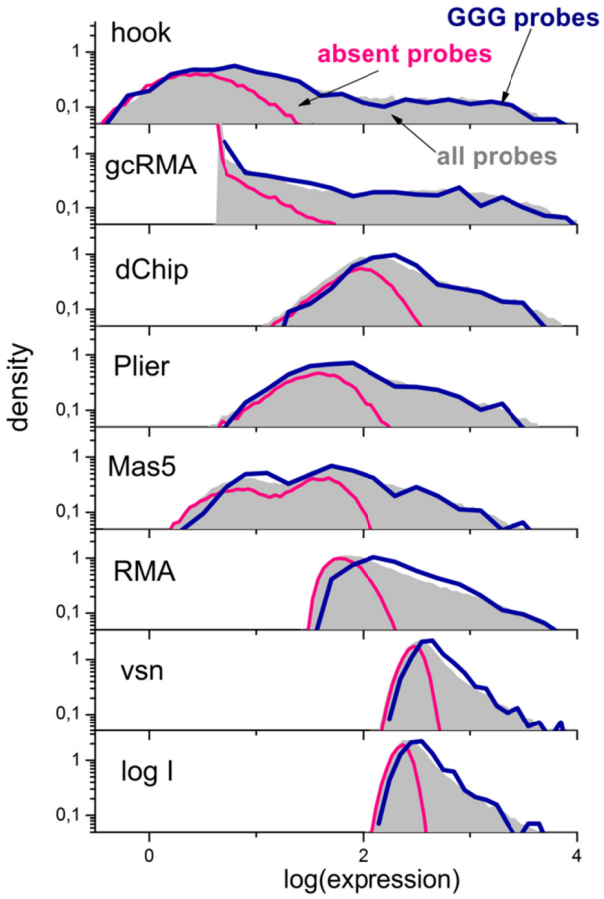

Figure 6.11: The distribution of expression measures obtained from intensity data shown in Figure 6.1 and various preprocessing methods. The whole density distributions reveal subtle differences produced by the different methods. The distributions are computed separately for all probe sets (45,100) and for probe sets with at least two probes containing a $(GGG)_1$-motif (836, i.e. 2% of the total number) and for absent probe sets hybridized exclusively nonspecifically (45% of all probes). The multichip methods (RMA, gcRMA, dChip, vsn, Plier) are applied by computing intensity data of 5 arrays from the respective experimental series. 'log I' denotes the distributions of raw intensity data. The distributions of expression measures of probe sets containing $(GGG)_1$ probes for RMA, gcRMA and to a less degree for MAS5 and dChip are systematically shifted to the right compared with the distribution of all probe sets. These methods are partly unable to correct expression values for the $(GGG)_1$-bias whereas hook and Plier remove the bias.

Figure 6.11 also shows the distribution of the sub-ensemble of `absent' probe sets (49% of all probe sets) which have been identified using the hook method. Comparison with the other distributions reveals that the amplitude of the $(GGG)_1$-bias decreases with increasing expression value. However, it affects not only the range of non-specific background but extends to probe sets with a significant contribution of specific hybridization. These signals are potentially used in downstream expression analysis. The right tail of the distribution is dominated by specific hybridization which has been shown to remain virtually unaffected by the $(GGG)_1$-effect.

The preprocessing methods dChip [124], gcRMA, MAS5 [125], Plier [97] and hook [42] apply probe-specific baseline correction algorithms which estimate an individual background value for each probe. The obtained distributions significantly widen, and

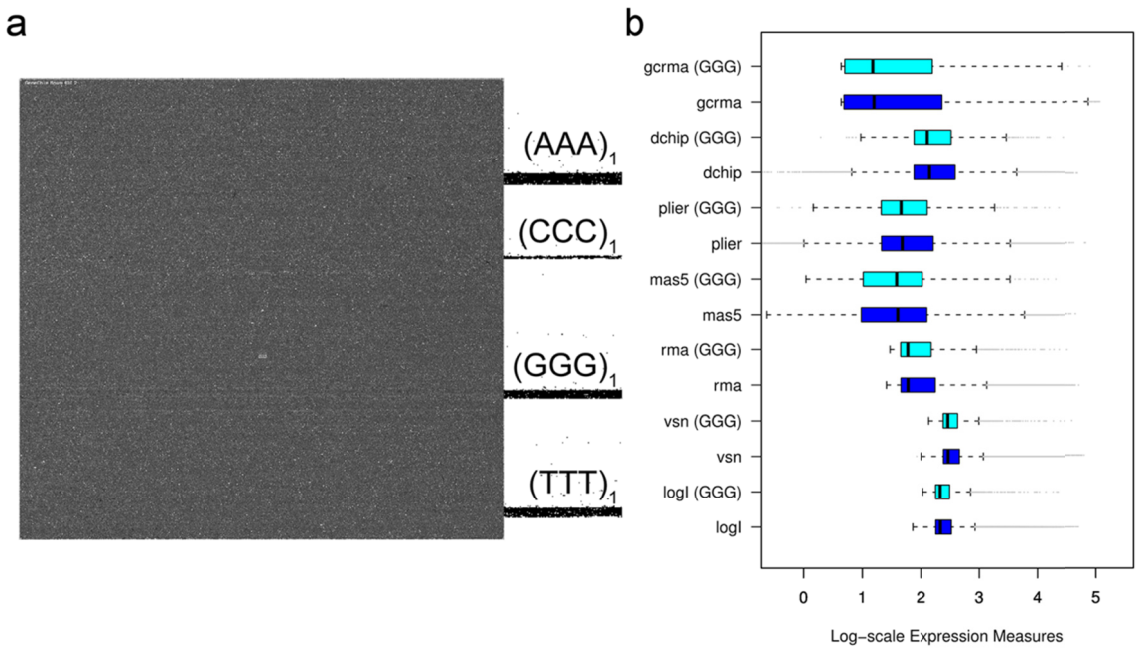

Figure 6.12: The figure shows the same data as in Figure 6.1 after sensitivity correction using the NN+GGG model. (a) Pseudo image of the chip calculated from the CEL-file of corrected intensities. The bright stripes seen in Figure 6.1a disappear. (b) Boxplots of expression measures obtained from the pre-corrected intensities. The GGG-bias essentially vanishes after correction (compare with Figure 6.1).

extend towards smaller expression values offering a larger dynamic range of the obtained expression estimates. The detailed inspection of the density distributions however also reveals a small $(GGG)_1$-bias in the left part of the distributions obtained by MAS5, dChip and gcRMA which is dominated by non-specific hybridization. gcRMA applies a positional dependent sequence correction of rank r = 1 similar to ours (Eq. (3.6)) which is obviously insufficient to account for the $(GGG)_1$-effect. MAS5, dChip and also Plier explicitly use the intensities of the MM probes to estimate the non-specific background of the PM signals. PM- and MM-probes are both affected by the poly-G motifs to a similar extent which enables its effective correction by combining PM- and MM-data. Finally, hook and Plier almost completely remove the $(GGG)_1$-bias from the data over the whole width of the distributions.

Figure 6.12 reproduces Figure 6.1 for corrected intensity values using the NN+GGG model. Panel a shows a pseudo-image of the array using the CEL-file of corrected intensities. The bright stripes due to the $(GGG)_1$ probes evident in Figure 6.1a clearly disappeared. Panel b illustrates the performance of different preprocessing methods with respect to the $(GGG)_1$-bias after applied correction. The boxplots clearly show that our correction effectively removes the $(GGG)_1$-effect from the resulting expression values.

In summary, most of established preprocessing methods only inadequately calibrate raw intensity data for strong sequence effects of the non-specific background contribution. Methods which explicitly process suitable reference probes, such as the MM, perform

better than PMonly methods. Precorrection of the intensity data using the NN+GGG sensitivity model removes the bias due to degenerated guanines from the data.

6.8.4 Comparison of sequence-specific intensity corrections

The correction for sequence-specific intensity effects is a crucial step which largely affects the performance of the preprocessing of microarray data. It applies to specific hybridization (`affinity' correction) as well as to non-specific hybridization (correction for the chemical background). Numerous sequence models have been developed for microarray analysis so far. They can be roughly divided into the following four classes:

(i) `Fully' physical, ΔG based approaches (here ΔG symbolizes the change of the free energy upon probe/target binding) [102, 115, 126–133]: These models explicitly and in-detail consider different processes which potentially affect probe hybridization such as probe/target duplexing including their zippering, bulk dimerization of the targets or folding of target and probe in terms of effective reaction constants or statistical thermodynamics. Elementary interactions are described on the level of base pairings using stacking free energy parameters which have been estimated in independent dimerization experiments of oligonucleotides in solution [112, 134]. Such models helped to improve our basic understanding of the functioning of microarrays and also to judge the relevance of different contributions to the observed probe intensities. These approaches often apply special fitting approaches and/or idealized assumptions to describe intensity data of selected microarray experiments (for example spiked-in data sets). Often, the used tools and algorithms however fail in practical microarray analysis because particular factors significantly affecting the performance of chip measurements are either considered in a simplified fashion or even neglected. For example, the lack of knowledge about the exact length, full sequence and concentration of the targets circumvents the detailed estimation of their folding and duplexing products. On the other hand, these `physical' models clearly showed that microarray hybridization is in agreement with elementary physical rules of interacting probes and targets, which however take place in a complex environment owing to the attachment of probes to the chip surface and the heterogeneous composition of the target solution. The latter conclusion was also supported by the results of reverse top-down studies which extract interaction parameters on the level of base pairings from microarray intensity data. For example, the resulting intensity-based NN parameters in most cases correlate well with the respective stacking free energies of independent solution experiments [29, 54, 102, 129].

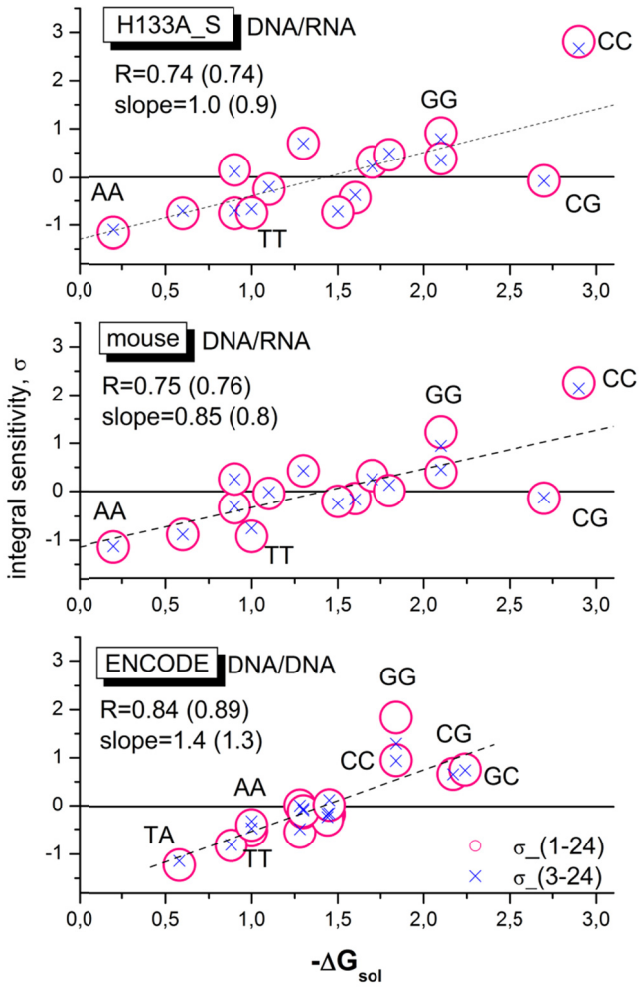

Figure 6.13: Correlation plots between the integral sensitivity of the positional dependent NN-model and solution free energies of DNA/DNA- and DNA/RNA-hybridizations taken from [112, 135] *and* [114]*, respectively. The three panels refer to DNA/DNA (ENCODE) and DNA/RNA (mouse, HG133A_S) hybridizations. The integral sensitivities are calculated using either all sequence positions (circles) or positions 3 - 24 (crosses). The latter data are normalized using the normalization factor 24/21 for direct comparison with the former data. Regression lines are shown for the latter data. The regression coefficients (R) and the slopes are given in the figure. The values in parentheses refer to the reduced sum. Note that the integral sensitivities of GG nearest neighbor motifs clearly decrease if one neglects the first two sequence positions. However, the effect on the regression remains small. Selected NN motifs are assigned in the figure.*

Our results confirm these previous findings (Figure 6.13). In particular, we calculated the sum of all terms of the NN-profiles over all 24 sequence positions for the selected data sets to obtain positional independent mean sensitivity estimates. The obtained integral NN-terms were correlated with the respective nearest-neighbor free energies for DNA/DNA or DNA/RNA duplexes in solution which were taken from [112, 135] and [114], respectively. The microarray sensitivities well correlate with the solution free energies (regression coefficients of $R > 0.7$). To judge the amplitude of the $(GGG)_1$-effect on the integral NN-terms we calculated a second data set which omits the first three sequence positions in each sum for the integral NN-terms (see the crosses in Figure 6.13). Only the values of the GG-

terms reduce notably in the mouse and ENCODE data sets accompanied by a small improvement of the respective fits.

The latter result shows that global parameter estimates can mask special intensity effects associated with selected sequence motifs such as runs of guanines, which results in the poor modeling of the intensities of probes containing these motifs.

(ii) Positional dependent intensity models with freely adjustable parameters in analogy to the approach used in this study: This class of models was independently introduced by Mei *et al.* [101] and Naef and Magnasco [50] which originally use single base terms, rank r = 1. Shortly after the method has been upgraded to NN-terms of rank r = 2 [51] and successfully applied in different calibration algorithms for microarray data using either N- [107, 136] or NN-models [34, 42, 137–139]. The parameters are estimated individually for each array. The model thus accounts for the specifics of each particular hybridization which potentially varies from chip to chip due to different levels of non-specific hybridization, bulk dimerization, washing and/or saturation. All these effect are shown to modify the respective parameter profiles [57, 106]. The obtained parameters are therefore called effective affinities [106] or sensitivities [53, 54] depending on the special experimental setup. Moreover, the model also enables to describe subtle differences between non-specific and specific hybridization on the level of base pairings, for example, due to the presence of defined mismatches in the probe/target duplexes [36, 42, 53]. The approach successfully applies to chips of different generations and types [45, 136] and it can be combined with elements of model class (i), for example, to account for probe and target folding [137, 139] or for special motifs and additional factors [101, 136]. For example, the pioneering approach of Mei *et al.* [101] combines the positional dependent N-model with special correction terms for intramolecular hairpins and G-quadruplexes. The latter effect was separately assigned to runs of at least four guanines at the beginning, the middle and the end of the probes. Here we extended the model to positional dependent triple and quadruplex motifs of rank r = 3 and 4. Our analyses show that the NN-model well accounts for most of the sequence effects except special motifs such as runs of consecutive guanines. We also demonstrated the diagnostic power of this approach to detect subtle sequence effects in terms of position and motif.

(iii) Positional dependent approaches with common `shape functions': This class of models is closely related to the previous class (ii). In contrast, it however factorizes the positional and motif dependent sensitivity profiles into two independent contributions namely into positional independent but motif specific `energy' terms and into a positional dependent but motif independent `shape'-function common for all motifs. This so-called PDNN model was originally introduced by Zhang *et al.* [47]. It is used with modifications in

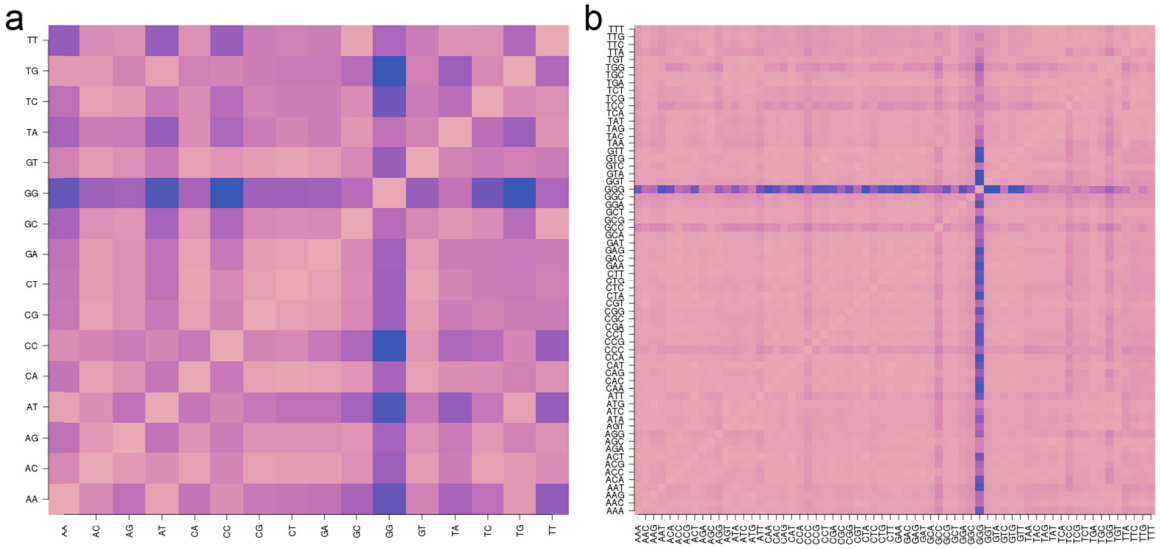

Figure 6.14: Heatmaps of the similarity matrix SI (b2, b2) of the shapes of positional dependent sensitivity profiles of rank r = 2 (panel a) and r = 3 (panel b) of the mouse data set. Pair-wise similarity is color-coded: dark spots indicate small similarity (see text).

different algorithms and applications [102, 140–142]. The common shape function of the PDNN model considerably reduces the number of adjustable sequence parameters by nearly one order of magnitude and consequently also the computational effort compared with the NN-model with motif specific profiles ((16 - 1) + 24 = 39 PDNN-parameters versus 361 NN-parameters, see Eq.(6.5)). It has however to be asked whether the common shape function adequately reflects the positional dependence of the individual NN-profiles or not? The inspection of the plots in Figure 6.1 and Figure 6.2 suggests, for example, that the shape of guanine-rich profiles strongly deviates from the shape of other motifs owing to the $(GGG)_1$-effect. For a systematic evaluation we make use of the NN-model with adjustable positional sensitivities of class (ii) and compare all pairwise combinations of the 16 sensitivity profiles using a simple similarity metrics based on the least squares optimization of a scalable factor a and a shift-term c;

$$\mathrm{SSR}(b_1,b_2)=\sum_{k=1}^{25-r+1}(\sigma_k(b_1)-a\sigma_k(b_2)+c)^2\Rightarrow\min \qquad (6.10)$$

Here b_1 and b_2 denote two selections from the 16 NN terms. The similarity matrix $\mathrm{SI}(b_1,b_2)=0.5\cdot(\mathrm{SSR}(b_1,b_2)+\mathrm{SSR}(b_2,b_1))$ indeed reveals that the profile of GG-sensitivities poorly matches the remaining profiles, except TT (Figure 6.14a). Bad or only moderate agreement is also observed between the profiles of other NN-motifs such as CC, AA and TT. The similarity matrix of the NNN-profiles of rank r = 3 reveals a similar picture with poor matches especially for GGG motifs and partly also for CCC and CCG (see Figure 6.14b). Hence, the assumption of a common shape fails for selected motifs.

iv) Multichip statistical models: These approaches decompose each probe intensity into independent factors due to probe and chip effects. The former factor is assumed to be invariant for each probe in a series of arrays and thus models the respective sequence-specific affinity of the probe. The latter factor is assumed to describe the expression index which usually varies between the chips. The relation between intensity and expression index is either linear (RMA, gcRMA, vsn, dchip, Plier) or hyperbolical assuming a Langmuir isotherm (Nlfit, [143]). The parameters are estimated by fitting the model to the intensities of a series of, at minimum, 5 - 10 arrays. The approach has the potential to correct the intensities for any probe effect because each probe is handled individually without explicitly processing its sequence in terms of a sequence model as in the alternative classes of approaches (i) - (iii). On the other hand, chip and probe effects are not independent in real situations due, e.g., to different levels of bulk dimerization and other effects (see above). More importantly, the probe-related affinity correction of the multichip methods in most cases applies to specific hybridization only whereas the non-specific background is corrected using simpler approaches such as global background (RMA, vsn) or N-profiles (gcRMA). Hence, the performance of the method largely depends on the type of background correction (see also the previous subsection). Note that dChip and Nlfit assume a probe dependent background which partly removes the the $(GGG)_1$-bias from the data (see the results for dChip in Figure 6.11).

We conclude that hybrid models of class (ii) are conceptually best suited to account for special sequence effects in single-chip based calibration algorithms for microarrays which use a high number ($> 10^5$) of short (length < 30 bp) oligonucleotide probes such as GeneChips. Here the large number of intensity values allows successful fitting of hundreds of model parameters. Possibly, the performance of models of this class can be further improved using amendments taken from physical models of type (i), e.g. to consider the folding propensity of the targets and/or their length. The non-linear approach [143] offers an interesting option of models of class (iv) because it allows to apply adequate hybridization laws beyond the linear approximation in combination with sophisticated affinity corrections. Its multichip character, however, adds normalization tasks to consider variations between different hybridizations which might produce biased expression estimates [57]. Models of class (iii) must be complemented with special terms to account for special sequence effects deviating from the mean positional dependence of the array. With this amendment they represent an interesting choice for array-types using long oligonucleotide probes (length > 30 bp) because it requires fitting of a reduced number of positional parameters compared with models of class (ii).

6.9 Summary and conclusions

We analyzed the specifics of probe intensities on the level of short motifs of one to four adjacent nucleotides along the 25meric probe sequence using positional dependent sensitivity models. The decomposition of the fit statistics into motif- and positional dependent contributions reveals that most of the motif-specific terms are adequately described using a nearest-neighbor model. In contrast, runs of degenerated guanines require explicit consideration of next nearest neighbor terms for adequate fitting.

Longer runs of at minimum three consecutive guanines along the probe sequence and especially triple degenerated G at its solution end typically cause exceptionally large probe intensities on expression arrays. This intensity bias affects PM- and MM-probes to a similar extend. Our analysis clearly shows that it is associated with non-specific hybridization. Hence, the interpretation of the extraordinary strong signals of probes containing runs of degenerated guanines in terms of high expression levels of the respective genes seems not justified.

The $(GGG)_1$-effect tends to increase gradually for microarrays of later GeneChip generations. It was detected for hybridizations which use DNA/RNA as well as DNA/DNA probe/target-chemistries. Different amplitudes of the guanine effect were found for hybridizations which apply different amplification protocols. In particular, the T7 amplification step for sample messenger RNA is associated with strong amplitudes of the guanine effect whereas amplification protocols for genomic DNA lacking T7 priming behave differently.

The origin of the very strong $(GGG)_1$ effect is unknown. Its association with the T7-protocol however implies that the T7-amplified targets containing the G-rich primer fragments are prone to form mixed probe/target G-stacks via association with G-rich probe motifs. The large concentration of G-rich targets in the hybridization solution facilitate their strong binding to G-rich probes resulting in their strong intensity. The absence of these G-rich target motifs in the ChipChIP hybridization possibly explains the much smaller intensity of the respective $(GGG)_1$ probes compared with the ENCODE. This hypothesis requires further verification using, e.g., methods developed in [111].

Established preprocessing methods only insufficiently remove the guanine bias from data. Methods which explicitly process the intensities of the MM probes as suitable references perform better than PMonly methods. We propose a positional dependent NN+GGG hybrid-rank model to correct the intensity bias associated with probes containing poly-G motifs. It can be applied prior to established preprocessing methods in a pre-correction step. The positional and motif dependent sensitivity models are conceptually best suited to

account for special sequence effects in single-chip based calibration algorithms for microarrays which use a high number of short oligonucleotide probes such as GeneChips.

The structural rationale behind the guanine effects has been concordantly assigned to the propensity of degenerated G-motifs to arrange into stable stacks of guanine tetrads which bundle four oligonucleotide strands into molecular quadruplexes [29, 99–101, 103]. These structures potentially affect the efficiency of oligonucleotide synthesis and/or the hybridization of the probes to their target sequences accounting for the abnormal performance of G-runs on the array [29]. Upton *et al.* [99] suggested a mechanism which increases the intensity of poly-G containing probes via the local opening of regions in the vicinity of quadruplexes formed by adjacent probes.

Alternatively one can assume that G-rich probes form G-quadruplexes of different stoichiometry which involve either exclusively adjacent probe oligonucleotides or also non-specific targets containing longer runs of guanines. We suggest that T7 amplification contaminates the targets with G-rich primer fragments which drastically increase their propensity to form such mixed probe/target G-quadruplexes. This model predicts that the large concentration of G-rich targets in the hybridization solution gives rise to their strong binding to G-rich probes which finally causes their strong intensity. The absence of these G-rich motifs upon hybridization of genomic DNA then explains the much smaller intensity of the respective probes.

7 Prevalence and impact of technical bias

7.1 Technical artifacts can be observed in batches

Non-biological, systematic variation due to varying experimental conditions constitutes a technical bias that negatively affects the reliability of microarray results. This was impressively shown in the introductory example in Section 1.1 where the results of the study of Spielman *et al.* were found to be spurious because more than 79% of genes were differentially expressed between two groups of samples processed at different times - an unrealistic number that cannot be explained by biological variation. These *batch effects* are a major issue in microarray data analysis and corrupt gene expression measurements via factors clearly unrelated to biology [144]. Correlation of such a factor with the biological variable of interest can prevent identification of the true biological source of variation and render the results of a microarray experiment worthless.

It is therefore of great importance to study the various sources of batch effects, their prevalence and their impact. A possibility to assess whether batch effects are present in a data set is to test for correlations between the potentially confounding factors and the expression measurements. A prerequisite however is that one has data on the factors potentially varying between batches of samples, for example the quality of the RNA, the used hybridization buffers and the employed instruments. In practice however, only a few of those factors are recorded in the course of an experiment - typically experimental date or location. These are frequently used as surrogate variables for the actual sources of variation.

In this section, we employ the methodology developed in the previous sections to the broader issue of common sources for batch effects. We investigate the general prevalence of a number of known technical effects using a large and representative number of microarray samples. For each of the considered effects, we will assess its impact on the experimental results in the form of gene expression estimates, and suggest how to avoid or remove them.

7.1.1 Human expression data

We have downloaded the *HumanExpressionAtlas* data set (E-MTAB-62 on Array Express) compiled by Lukk *et al.* [145] consisting of 5372 ('qc-included') samples hybridized to Affymetrix HG-U133a microarrays. This data set has been collected from 206 public experiments and represents 369 distinct human cell and tissue types, disease states and cell

lines. The resulting *expression space*, the combined and processed gene expression data from this diverse collection of human samples, can also be queried using the dedicated database ArrayExpress Atlas [146].

In [145] the 5372 samples have been selected from a larger data set of 8268 samples after application of strict quality control (qc). We obtained a full list of the 8268 samples from the authors and downloaded the remaining 2896 ('qc-excluded') samples from public databases. From these, 137 samples could however not be retained as they were removed from the databases, leaving in total $8268 - 137 = 8131$ samples. The full set of 8268 unique samples represents virtually all HG-U133a data publicly available in the two major public databases GEO and ArrayExpress in 2006. This *HumanArraysSet* therefore is a representative set of available human microarray samples.

7.1.2 Principal component analysis for gene expression data

The typical result of a gene expression experiment has the form of a huge n×m matrix containing estimates for the expression of n genes in m samples. The size of n ranges between a few hundred genes for spotted microarrays up to many thousands of genes. For example there are about 55.000 probe sets representing over 38,500 genes on a recent GeneChip microarray (see Table 6.1). The number of observations m typically ranges between a handful of samples for screening experiments up to thousands of samples for large cohorts.

A widely applied method for explorative analysis of such high-dimensional, multivariate data is Principal Component Analysis (PCA). It reduces the number of dimensions by transforming the possibly depending input variables into linearly independent variables called principal components [147]. These new variables are selected such that they explain most of the variance in the data (see [148]). Consequently, PCA captures the predominant patterns among the experimental features including both biological and technical variability. One typically focuses on the first couple of principal components ordered by decreasing amount of variability explained.

Consider the HumanExpressionAtlas data set described in the previous section. Lukk *et al.* classified each of the 5372 samples into 369 biological categories representing a particular cell or tissue type, disease state or cell line, and also introduced several 'meta-groups'. Figure 7.1 displays the first two principal components of the HumanExpressionAltas expression data where each point representing a sample is colored according the meta-groups hematopoietic system, solid tissues, incompletely differentiated cell types and connective tissues (left side) as well as the meta-groups cell lines, neoplasms, non-

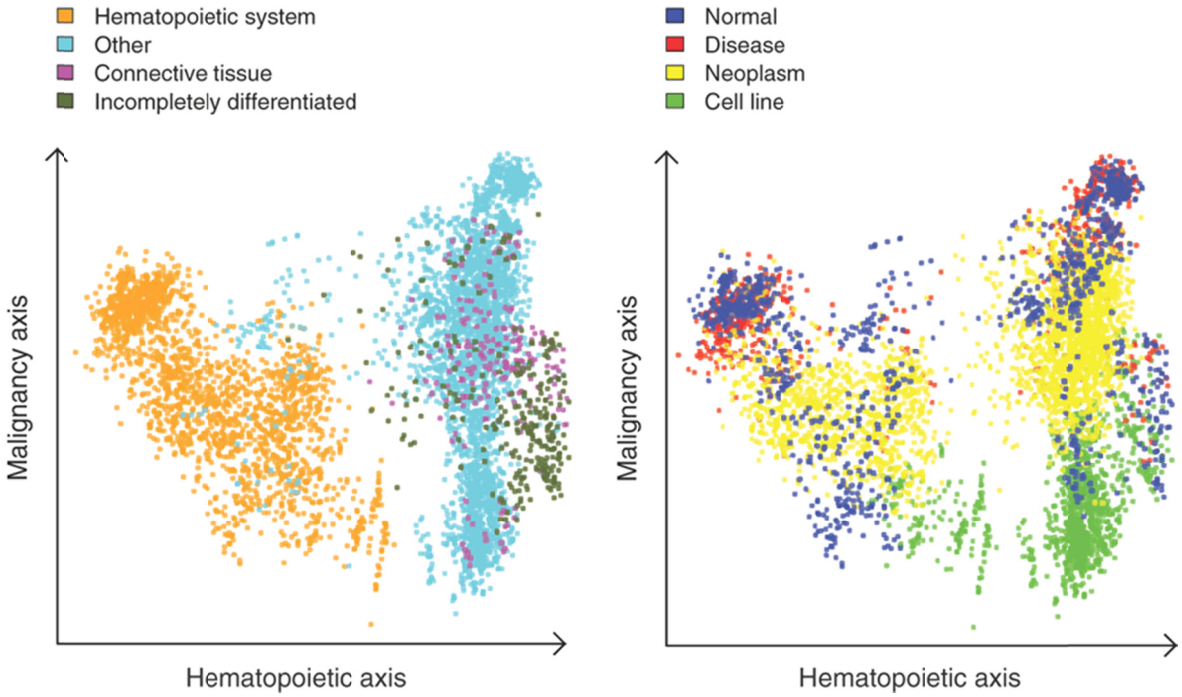

Figure 7.1: The first two principal components of the HumanExpressionAtlas data set. Each dot represents one of the 5372 samples colored according to its biological group. The first principal component (x-axis) separates hematopoietic system–derived samples from the other samples (left side) whereas the second principal component (y-axis) separates 'malignancy types' with cell line samples at the bottom, neoplasm samples in the middle and nonneoplastic disease and normal samples at the top (right side). Image taken from [145].

neoplastic diseases, and normal (right side). These categories separate along the first two principal axes indicating a possible biological interpretation. Lukk *et al.* found that the first three principal components, which explain more than 37% of the variability, have biological interpretations [145].

We here seek to investigate whether these biological factors of the HumanExpressionAtlas are confounded by other, non-biological factors, raising the possibility for alternate interpretations of the principal components. Similar to the approach above, correlation between a known technical variable (e.g. RNA-quality) and a major principal component indicates the presence of an unwanted technical side-effect in the resulting gene expression data.

7.2 RNA quality

Good RNA quality is an important prerequisite for obtaining reliable results from a microarray gene expression experiment (compare Section 5.1). Low RNA quality propagates to the obtained gene expression estimates and consequently to differential expression results. These risks combined with the previous detection of a noticeable degradation effect upon the majority of microarray samples [68] suggest that variation in

RNA quality could constitute a major technical bias. In this section we thus investigate the general variability of RNA quality among a large, representative set of array samples, and assess the prevalence of samples with critically low RNA quality. Lastly, we study the impact of this factor on the gene expression estimates of the HumanExpressionAtlas data set.

The RNA Integrity Number (RIN) provides a measure for RNA quality that is determined for most microarray samples before hybridization [72]. RIN values scale between 1 and 10, and using only samples with $RIN \geq 7$ is recommended for microarray analyses [84]. However, RIN values are unfortunately seldom stored in conjunction with the experimental data. The d^k degradation parameter (Eq. (5.20)) provides a sensitive estimate of RNA-quality that can be computed from raw microarray data and, as we showed in Section 5.3.4, correlates well with RIN. For fresh tissue, the cutoff of $RIN \geq 7$ corresponds to a $d^k \geq 0.45$ cutoff. Note that d^k values are not only sensitive, but also specific for RNA quality since only RNA degradation and amplification have such a systematic effect on the probe intensity decay (see Section 5.1.1).

We have computed the d^k values for all 8331 samples of the HumanArraySet which were either included or excluded from the HumanExpressionAtlas data set as described above. Figure 7.2a shows the resulting density distribution of the d^k values for the qc-included/qc-excluded sample sets. Most samples included after quality control have a degradation index between $0.5 \leq d^k \leq 0.8$ referring to acceptable RNA quality. On the other hand, a large fraction of the qc-excluded samples exhibits values of $d^k < 0.45$ referring to critically low RNA quality. This applies to 25% of the qc-excluded samples and to 10% (868) of all investigated samples.

Furthermore, 3% (162) of the qc-included samples are so severely degraded that they should have been excluded by RIN analysis. Expression estimates of these samples are biased, with negative consequences for the reliability of downstream results. That these samples are however included in the HumanExpressionAtlas suggests that a more rigorous assessment of RNA-quality should be applied in quality control procedures. Note that these results correspond well with a previous estimation of 2% of low RNA-quality samples given by Upton *et al.* [68].

Interestingly, only few qc-included samples have values larger than of $d^k = 0.8$ which obviously represents an upper limit referring to the ‘least possible intensity decay’ (compare Section 5.3.1). The presence of this limit could be attributed either to the insufficiency of the cleanup assays to stop RNAase activity or to the ubiquitous incomplete amplification of aRNA fragments. A fraction of 8.1% of the qc-excluded samples has values of $d^k > 0.8$ which could be due to other signal deficiencies (e.g. surface effects).

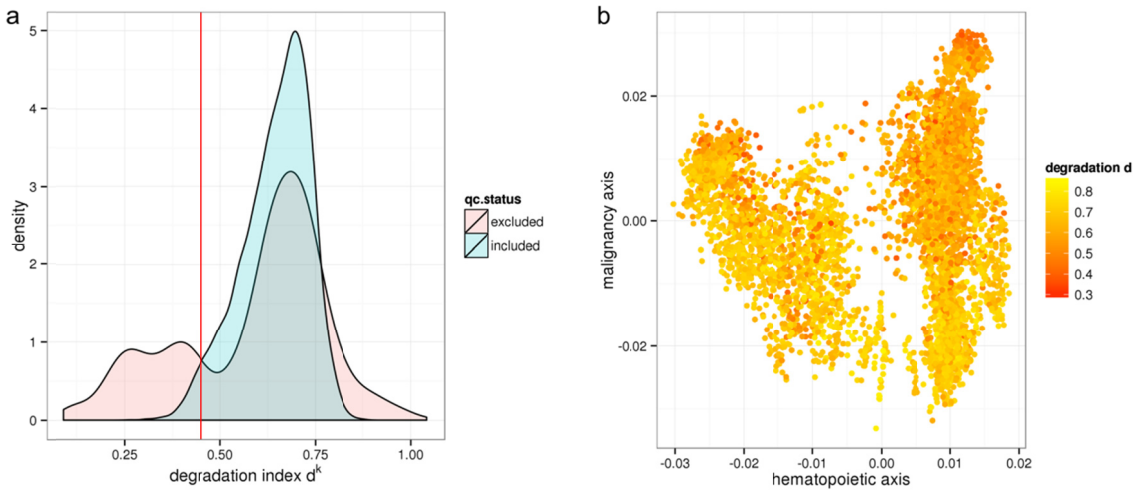

Figure 7.2: Variaton of RNA quality among a large set of microarray samples and the impact on expression results. Panel a shows the density distribution of d^k values measuring the degradation for samples either included or excluded in the HumanExpressionAtlas data set by independent quality control. The red line indicates the low quality threshold corresponding to RIN ≤ 7. Panel b replicates Figure 7.1 showing the first two principal components of the HumanExpressionAtlas data set, but this time the points are colored according to the value of d^k. A correlation between the second principal component and degradation is clearly visible.

To test for confounding of the HumanExpressionAtlas with the effect of variable RNA quality we re-plot its first two principal components similar to Figure 7.1. This time the samples are not colored according to biological groups but instead according to the value of the degradation index d^k. The resulting plot in Figure 7.2b shows a clear color gradient along the second principal axis ('malignancy axis') with, in general, lower RNA quality (red spots) at the top and higher RNA quality (yellow spots) at the bottom. The second principal component, a major pattern in the expression space, visibly relates with the RNA quality. Formally, one can test for correlation between the first couple of principal components and the technical factor or variable of interest as done for example by Leek *el al.* [149]. We found the highest correlation (in absolute terms) to be with the second principal component with a Pearson's correlation coefficient of r = -0.44.

In summary, investigation of the RNA quality of publicly available microarray data suggests that a substantial fraction of samples has substandard quality and should be excluded from further analysis. A correlation between degradation index d^k and the second principal component of the HumanExpressionAtlas data was found. According to Lukk *et al.*, this so-called 'malignency axis' differentiates cell lines, neoplasms and normal/non-neoplastic disease tissues (see Figure 7.1). Our analysis shows that this axis also differentiates RNA quality where normal/non-neoplastic disease tissues are associated with low RNA quality. Given the confounding of biological classification with RNA quality, identification of the true origin of this important source of variation – is it biological or is it rather related to the preparation of the respective cell and tissue types – requires further investigation.

7.3 Amount of hybridized RNA

Ideally sufficiently large amounts of aRNA transcripts at constant levels in the range of 10-100ug should be used for hybridization to the surface-attached microarray probes to obtain good quality data [150]. In practice these ideals are hard to archive due to the considerable variation in the amount of available source RNA. In some types of experiments the amount of source RNA is highly limited, down to nanogram and even picogram ranges, for example in applications where specific cells are selected by laser capture microdissection [151]. Specialized RNA amplification and sample preparation assays have been developed helping to obtain sufficient amounts of aRNA (for a comparison of these methods see e.g. [66, 152]). Too low and too high amounts of aRNA can reduce the dynamic range of the fluorescence signals and increase the signal-to-noise ratio by insufficiently exhausting the measuring range for specific transcripts. Consequently, varying RNA amounts can affect gene expression estimates and reduce data quality, rendering the assessment of the prevalence and impact of the thereby induced technical bias reasonable.

In Section 3.6 we showed that the summary measure $\langle\lambda\rangle$ (Eq. (3.15)) within its limitations is a sensitive parameter for varying amounts of RNA. The density distribution of the $\langle\lambda\rangle$ parameter, as previously separated for the qc-included/qc-excluded sample sets, is displayed in Figure 7.3a. For most good quality samples $\langle\lambda\rangle$ ranges between 1.0 and 1.5 with the peak at $\langle\lambda\rangle = 1.2$. Interestingly, the peak of the $\langle\lambda\rangle$ distribution is significantly shifted to the left to $\langle\lambda\rangle = 1.05$ for samples excluded by quality control, indicating that low quality samples have decreased relative specific transcript levels possibly relating to low RNA amounts (see below).

Virtually none (< 0.1%) of the samples that passed stringent quality control exhibit values smaller than $\langle\lambda\rangle = 0.95$, which we consequently consider a conservative threshold for samples of critically low quality due to decreased RNA amounts. We find that 133 (1.6%) of all samples exhibit $\langle\lambda\rangle$ values below this threshold. This equals a fraction of 4.6% from the qc-excluded samples.

It should be noted that $\langle\lambda\rangle$ describes the average specific transcript level of all genes in units of the non-specific one, and the unexpectedly low expression levels of some genes can have other origins than low RNA amounts, for example local surface deficiencies (e.g. fingerprints). Also note that low RNA amounts can as well be a result of degraded RNA (see Section 5.4), suggesting an overlap between both technical effects.

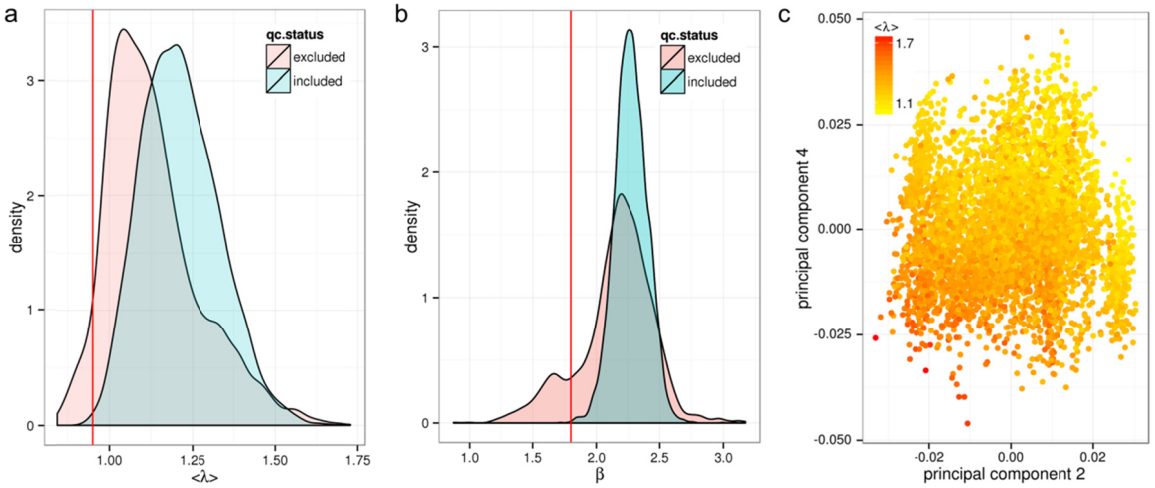

Figure 7.3: Distribution of the <λ> and β parameters characterizing the amount of hybridized RNA for qc-included/qc-excluded samples (panels a and b) and the correlation of <λ> with the principal components 2 and 4 of the HumanExpressionAtlas data (panel c, compare Figure 7.2).

The density distribution of the β parameters describing the measurement range (see Section 3.6) of the microarray hybridization is shown in Figure 7.3b. Low RNA amounts are associated with a larger β values whereas high RNA amounts increase the non-specific background with negative consequences for the measuring range, and thus the signal calibration [57]. For qc-included samples the summary values are distribute closely (±0.3) around the peak at $\beta = 2.25$ whereas for qc-excluded samples the β values spread much broader with a second peak for smaller measuring ranges. Selecting a threshold of $\beta < 1.8$, we find that 344 samples (4.2%) have a low measurement range.

We also assessed the impact of the non-biological variables $\langle\lambda\rangle$ and β by relating them with the first five principal components of the expression space of the HumanExpressionAtlas data set. We obtained a correlation of $r = -0.61$ of $\langle\lambda\rangle$ with the fourth principal component. Furthermore, a correlation of $r = -0.33$ with the second principal component, which we previously showed to relate with RNA quality, was found. The other three components show only low correlations of $-0.16 < r < 0.21$. With coefficients of $-0.04 < r < 0.01$, the β parameter exhibits no correlation with the first five principal components.

In summary, a significant fraction of the human samples is affected by a 'decreased specific transcript level' set of effects that relate to low amounts of hybridized RNA. We find that the predominant patterns of expression variation are significantly affected by the technical variable $\langle\lambda\rangle$ which highly correlates with the fourth principal component of the HumanExpressionAtlas data. This is a different principal component than the one showing high correlation with the RNA quality measure d^k. Analysis of the β parameter shows that about 4% of the samples have low measurement ranges, and that β has no impact on the predominant patterns in the expression space.

7.4 Sequence effects

7.4.1 Maximum sensitivity amplitude

Nucleic acid folding and formation of DNA/DNA or DNA/RNA duplexes on surfaces are fundamental reactions for any microarray assay and largely depend on the conditions under which they occur. For example, the temperature and time given for the reactions affects sensitivity and specificity of nucleic acid binding [14]. Condition changes can thus lead to sequence-dependent variations in the probe intensity signals, which can further propagate to the gene expression estimates and therefore constitute potential technical artifacts.

To this end, we investigate how sequence-dependent binding affects gene expression data. We first define the *maximum sensitivity amplitude* based on most extreme sequence contribution δA (see Eq. (3.6)) in positive and negative direction

$$\Delta_{\max} \log(K_{\text{diff}}^{P,h}) \equiv \max_{\xi}(\delta A^{P,h}(\xi)) - \min_{\xi}(\delta A^{P,h}(\xi)) \tag{7.1}$$

measured in units of log intensity contributions. We here refer to the perfect-match probes (P = PM) of the non-specific hybridization mode (h = NS). Given the estimated sensitivities, it determines how much a probe could shine brighter than another one given that both probes target the same transcript. For example, a value of $\log(K_{\text{diff}}) = 5$ for a particular hybridization means that, on the average, two hypothetical probes (most likely with the sequences AAA…A and CCC…C) would differ in their intensity values by 5 orders of magnitude. It can thus be thought of as the maximum strength, or impact, of the sequence effect.

As previously, we computed $\log(K_{\text{diff}})$ for all 8331 samples of the HumanArraySet and plot the respective density distribution in Figure 7.4a. By trend qc-excluded samples show a lower maximum sensitivity amplitude, rendering it a potential marker for low quality samples. Based on the observation that barely any good quality samples (< 0.1%) exhibit a smaller maximum sensitivity amplitude , $\log(K_{\text{diff}}) = 3$ is chosen as conservative threshold selecting samples with critically low sequence effect size. This applies to a fraction of 4.1% of the samples.

In order to assess the impact of the sequence effect size we computed correlations of the $\log(K_{\text{diff}})$ parameter with the first five principal components of the HumanExpressionAtlas data. The largest correlation in absolute scales is r = -0.17 with the third principal component. Correlations for the remaining principal components are smaller than |r| = 0.11. In conclusion, the sequence effect size is not a technical variable with a large impact on the most common patterns in the expression space.

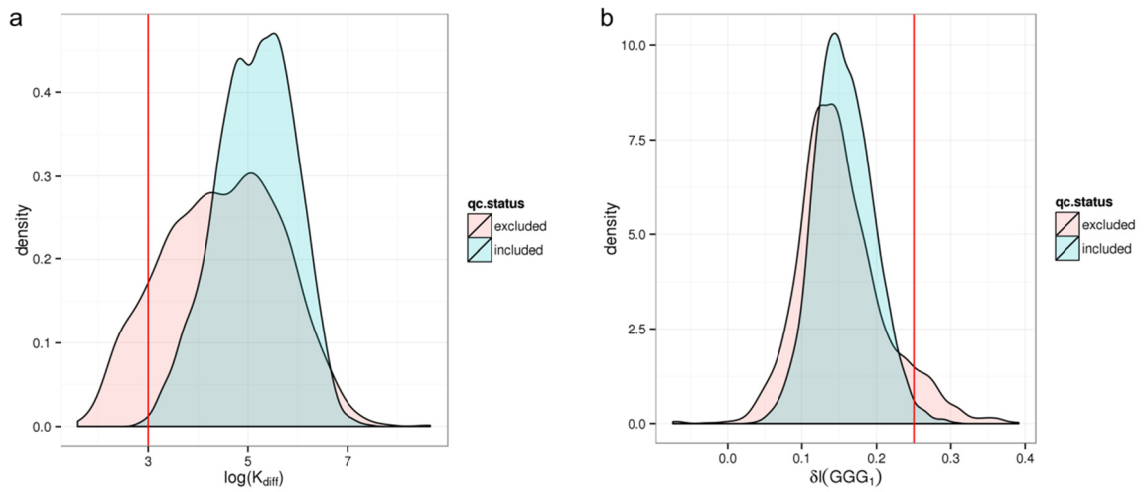

Figure 7.4: Distribution of summary parameters related to sequence effects for groups of samples either included or excluded from the HumanExpressionAtlas data set by independent quality control (compare Figure 7.2a). Panel a shows the parameter log(K_{diff}) as a measure of the total strength of the sequence effect. Panel b shows the intensity increase due to the $(GGG)_1$ motif, $\delta I(GGG)_1$.

7.4.2 Guanine effects

In Section 6.3 we found that runs of guanines within the probe sequence, particularly runs of guanines as long or longer than 3, significantly affect the obtained signal intensities. We showed later that this $(GGG)_1$ effect propagates through the various preprocessing methods of microarray analysis and can lead to biased expression estimates. The origin of the $(GGG)_1$ effect lies in the formation G-quadruplex structures. The formation of duplexes between negatively charged nucleic acids in general, and between G-quadruplexes in particular, depends on the ionic strength and thus on the employed solution buffer [153]. This dependency on the ionic strength also applies to hybridization reactions on solid surfaces [154]. Given both its dependence on changing conditions and its potential effect on the expression results, it is reasonable to study the overall prevalence and impact of guanine effects in microarray expression experiments.

We here define the strength of the guanine effect in terms of the *intensity increase due to the $(GGG)_1$ motif* as follows

$$\delta I(GGG_1) = \left\langle \log I_p^{PM} \right\rangle_{\xi_p^{1,3}=(GGG)} - \left\langle \log_p^{PM} \right\rangle_{(TTT)\in\xi_p} . \qquad (7.2)$$

A value of $\delta I(GGG)_1 = 0.3$ thus reflects an on the average $10^{0.3} \approx 2$ times as large intensity of probes containing the $(GGG)_1$ motif compared to probes containing (TTT) anywhere in their sequence. The average intensity of (TTT) containing probes here serves as appropriate baseline normalization.

Figure 7.4b shows the density distribution of the $\delta I(GGG)_1$ parameter which varies between $0 < \delta I(GGG)_1 < 0.3$ for qc-included samples. We consider samples with a threshold of $\delta I(GGG)_1 > 0.25$ to be significantly affected by GGG_1 bias – this reflects an intensity increase of +75% of the respective probes. Accordingly, 254 (3.1%) of the samples have a GGG_1 related intensity bias. Many of them are removed by strict quality control, only 63 (1.1%) of qc-included samples are above the threshold.

Assessing the correlation of GGG_1 bias with the first five principal components of the HumanExpressionAtlas expression data we observe correlation coefficients between $0.14 < |r| < 0.19$. Consequently, guanine effects have only minor impact on the common patterns in the expression space. Note that the RMA method was used for calibration, and we showed in Section 6.8.3 that expression estimates from this preprocessing approach are in general susceptible to GGG1 effects.

In summary, guanine effects are an important technical artifact that however only affects the expression estimates of *some* genes in a significant fraction of microarray samples. It does not affect the majority of features and is consequently not a major determinant for the predominant patterns in the expression space.

7.5 Summary and conclusions

In this section we have studied the general prevalence and impact of a RNA quality, RNA quantity and sequence effects using a large and representative set of microarray samples from the Affymetrix HG-U133a platform. To this end, we defined novel parameters, or used previously defined ones, that quantify each technical artifact based on systematic changes in the intensity signals. We determined appropriate thresholds indicating low-quality samples potentially leading to biased expression estimates due to the respective artifact. Their impact on the expression estimates was analyzed by computing correlations between the technical variables and the first five principal components of the expression space of the HumanExpressionAtlas.

We found that a large fraction of 10% of the 8131 samples are so severely degraded, that they should be excluded from further analysis. While most of these samples where indeed excluded from the HumanExpressionAtlas, still about 3% of the low-quality RNA samples passed quality control highlighting the need for a more rigorous assessment of RNA quality in microarray data analysis.

Unexpectedly high impact on the gene expression data was found for RNA quality and RNA abundance variation. Both affect the most common patterns in the expression space. The RNA quality measure d^k and the relative specific transcript level $\langle\lambda\rangle$ highly correlate

with different principal components. Together with the observed high prevalence of these artifacts, they constitute major sources of technical bias and should be monitored carefully in every experiment.

We found that sequence effects are highly variable in the investigated Affymetrix HG-U133a platform. The total sequence effect size is particularly low among low-quality samples where 4% of the samples are affected. 3% of the samples have a strong GGG_1 effect. While the GGG_1 effect can have a critical impact on the expression estimates of some of the genes, we found that the overall impact of the studied sequence effects on the expression space is relatively low.

8 Summary and discussion

In this thesis, we reviewed a number of established microarray technologies with a wide range of genomics applications together with the challenges that arise when their technical limitations meet the high standards required in research and clinical environments. Particularly, we showed how changes in the experimental conditions can have a large impact on the obtained data and can thereby lead to unreliable results. To better understand and control the experimental system we employed a model of microarray hybridization and demonstrated how it can be applied to different types of microarrays.

Using appropriate modifications of that model we studied the effect of selected hybridization biases using publicly available data from Affymetrix GeneChip expression arrays. We showed that varying amounts of hybridized RNA result in changes of the raw intensity signals and of the summary parameters $\langle\lambda\rangle$ and β computed from these. We also found that varying RNA quality strongly affects intensity signals of probes which are located at the 3' end of transcripts. New theoretical approaches and visualization methods were introduced that help assessing the RNA quality of a particular microarray sample. We developed a new metric for determining RNA quality based on the 3'/5' intensity bias of specific probes and showed that it outperforms other microarray-based quality metrics. We proposed a method for the correction of the 3' intensity bias, which, together with the other functionalities, has been implemented in the Bioconductor package *AffyRNADegradation*.

We further found that probe signals are affected by sequence effects which were studied systematically using positional-dependent nearest-neighbor models. Analysis of the resulting sensitivity profiles revealed that particular sequence patterns such as the GGG_1 motif have a strong impact on the probe signals. We showed that sequence effects differ for different chip- and target-types, probe types and hybridization modes. These and other factors introduce a strong sequence bias in the intensities that should be corrected in order to obtain reliable results. We showed that the NN+GGG PDNN model provides a good trade-off between correction efficiency and speed, and provide a software implementation for the sequence correction of raw intensity data of Affymetrix expression arrays in the *Larpack* program package.

In the final chapter, we used the previously developed methodology for the assessment of technical artifacts to study their general prevalence and impact on available microarray data. Using a representative ensemble of over 8000 human microarray samples, we found that in particular RNA quality and quantity have a strong impact on the obtained expression values. We also showed that about 10% of microarray samples have such low RNA quality that they should be discarded from further analysis.

Despite great advances in the efficiency of biological high throughput technologies and data analysis methodology, we still fail at explaining a significant fraction of the observed variability in the data. For example, probe intensities of tiling arrays exhibit a within-gene variability of several orders of magnitude and it is largely unknown whether there is a yet to be found biological explanation, or if it is due to technical artifacts. Hence there is either a lack of understanding of the complex cellular mechanisms and biochemical reactions leading to the production of the measured biomolecules, or a lack of understanding of the technical steps of sample preparation and the measurement process in these widely-used technologies.

The aim of this thesis is to increase the understanding and the control over systematic technical variation in microarray data. More understanding of the mechanisms of surface hybridization can help to improve on existing and potential future technologies. More control about the sources of technical variation increases the amount of reliable information about true biological variation, and thus the amount of knowledge that can be gained from high-throughput experiments.

In this thesis, we pointed out several problems in current microarray data generation and analysis methods, and proposed new approaches helping to solve them. Undoubtedly, further efforts are necessary to increase the validity and utility of the obtained results. First of all, awareness should be raised about existing technical limitations and possible biases in the data. For example, we showed here that biased expression estimates can be a result of sequence effects like the GGG_1 effect, which in turn are highly dependent on the conditions of the hybridization reaction. By these means differences in the experimental conditions, like the use of different buffers in two collaborating laboratories, can propagate to expression measure differences between two batches of samples. Researchers unaware of these effects can easily draw false conclusions.

Further, high standards in data quality control and documentation are immensely valuable, and should be further enforced. A first important step has been made by the establishment of standardized descriptions as MIAME (Minimum Information About a Microarray Experiment, [155]) which are now mandatory on common platforms hosting public microarray data. The required information includes descriptions of the experimental design, the array design and the used biological material and its treatments. While these important community standards help to reproduce and to validate the results of microarray experiments, we believe that the mandatory recording and storage of additional information on the experimental conditions and intermediate measurements would be a large benefit. We showed that more factors than previously thought have a significant impact on the microarray results in the form of expression data. The specifics of the design and protocols of the Affymetrix GeneChip platform allowed us to infer some of the

missing parameters from systematic changes in microarray expression data. However, this is not easily possible for other effects and other platforms. Only storage of intermediate results like RNA integrity or pH measurements along with the primary data enables further analysis of these technical effects and their origins.

Finally and importantly, model-based analysis helped to improve our understanding of microarray technologies. A basic hybridization model based on fundamental physical principles of surface binding applied well to gene expression data, as well as to the data of other microarray technologies with different applications. The model-based analysis of sequence and degradation effects allowed us to understand the introduced biases and to develop appropriate extensions to the basic hybridization model. With continuous refinement of our understanding and of our modeling, we hope to once reach a sufficiently comprehensive model so we can explain most of the technical variation, and can concentrate on understanding biology.

A List of data sets used

Table A.1: Microarray data sets used in this thesis. GSExxxx and GSMxxxx are the accession numbers of datasets downloaded from the GEO repository (http://www.ncbi.nlm.nih.gov/geo/). E-TABM-xxx and E-MEXP-xxx are accession numbers of data sets downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/).

Chip type	Employed publicly available dataset
HG-U95A	Genelogic dilution series (http://www.genelogic.com/support/scientific-studies)
Mapping50k Xba240	Mapping 100k HapMap Trio Dataset (http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx)
HG-U133 plus 2.0	GSE7307 (Human tissue)
Rat230A	E-MEXP-1069 (RatQC)
HG-U133A	GSE1133
HG-U133A and HG-U133 plus 2.0	GSE3061
MG430 2.0	GSE12545
ENCODE	GSE2800, GSE6292
Yeast_2	GSE9302
MG430_2	GSE12545
Zebrafish	GSE5048
EColi_2	GSE6893
CElagans	GSE6547
Rice	GSE6893
Chicken	GSE12268
ATH-121501	GSE7432
Rat230_2	E-TABM-536
MG430A	GSM154799, GSM355022, GSM366810
DrosGenome1	Fruitfly time series (http://camda.bioinfo.cipf.es/camda08/contest_dataset)
MG74A	GSM104601, GSM34328, GSM4310
HG-U133A	Affymetrix Latin Square HG-U133A (http://www.affymetrix.com/support/technical/sample_data/datasets.affx)
HG-U95A	Affymetrix Latin Square HG-U95A (http://www.affymetrix.com/support/technical/sample_data/datasets.affx)
30 random MG430_a arrays	GSM172403, GSM176889, GSM177368, GSM178084, GSM187846, GSM211338, GSM211425, GSM237785, GSM238367, GSM250880, GSM252214, GSM264815, GSM280709, GSM282803, GSM311514, GSM313208, GSM315604, GSM318915, GSM325421, GSM326978, GSM326998, GSM337788, GSM337834, GSM432906, GSM443776, GSM455430, GSM53318, GSM94768
HG-U133A	E-MTAB-62 (HumanExpressionAtlas)

List of figures

List of tables

Bibliography

1. Southern E: **Tools for genomics.** *Nature Medicine* 2005, **11**:1029-34.

2. **Affymetrix reports second quarter 2012 results** [http://www.reuters.com/article/2012/07/31/idUS246048+31-Jul-2012+BW20120731].

3. MacArthur D: **Methods: Face up to false positives**. *Nature* 2012, **487**:427-28.

4. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG: **Common genetic variants account for differences in gene expression among ethnic groups.** *Nature Genetics* 2007, **39**:226-31.

5. Jorde LB, Wooding SP: **Genetic variation, classification and "race".** *Nature Genetics* 2004, **36**:S28-33.

6. Akey JM, Biswas S, Leek JT, Storey JD: **On the design and analysis of gene expression studies in human populations.** *Nature Genetics* 2007, **39**:807-8.

7. Affymetrix: **Statistical Algorithms Description Document**. *Technical Note* 2002:28.

8. Robinson MD, Speed TP: **A comparison of Affymetrix gene expression arrays.** *BMC Bioinformatics* 2007, **8**:449.

9. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG: **The Affymetrix GeneChip platform: an overview.** *Methods in Enzymology* 2006, **410**:3-28.

10. Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays**. *Nature Genetics* 1999, **21**:20-24.

11. **3' IVT Express Kit** [http://www.affymetrix.com/estore/browse/products.jsp?categoryIdClicked=&productId=131415#1_1].

12. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH: **Amplified RNA synthesized from limited quantities of heterogeneous cDNA.** *Proceedings of the National Academy of Sciences of the United States of America* 1990, **87**:1663-7.

13. **Affymetrix Image Library** [http://www.affymetrix.com/about_affymetrix/media/image-library.affx].

14. Relogio A: **Optimization of oligonucleotide-based DNA microarrays**. *Nucleic Acids Research* 2002, **30**:51e-51.

15. **Affymetrix - Help - FAQ - What does the "_s_at" extension represent in the HG-U133 probe set name?** [http://www.affymetrix.com/support/help/faqs/hgu133/faq_4.jsp].

16. Affymetrix: *GeneChip® arrays provide optimal sensitivity and specificity for microarray expression analysis*.

17. Kim H, Klein R, Majewski J, Ott J: **Estimating rates of alternative splicing in mammals and invertebrates.** *Nature Genetics* 2004, **36**:915-6.

18. **User Guide: Ambion WT Expression Kit**
[http://tools.invitrogen.com/content/sfs/manuals/cms_064619.pdf].

19. Eklund AC, Turner LR, Chen P, Jensen RV, deFeo G, Kopf-Sill AR, Szallasi Z: **Replacing cRNA targets with cDNA reduces microarray cross-hybridization.** *Nature Biotechnology* 2006, **24**:1071-3.

20. Okoniewski MJ, Hey Y, Pepper SD, Miller CJ: **High correspondence between Affymetrix exon and standard expression arrays.** *Biotechniques* 2007, **42**:181-5.

21. Pradervand S, Paillusson A, Thomas J, Weber J, Wirapati P, Hagenbüchle O, Harshman K: **Affymetrix Whole-Transcript Human Gene 1.0 ST array is highly concordant with standard 3' expression arrays.** *Biotechniques* 2008, **44**:759-62.

22. McCall MN, Almudevar A: **Affymetrix GeneChip microarray preprocessing for multivariate analyses.** *Briefings in Bioinformatics* 2012, **13**:536-46.

23. Brookes AJ: **The essence of SNPs**. *Gene* 1999, **234**:177-86.

24. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome.** *Nature Reviews Genetics* 2006, **7**:85-97.

25. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-54.

26. **Genome-Wide Human SNP Nsp/Sty Assay Kit 6.0**
[http://www.affymetrix.com/estore/browse/products.jsp?productId=131534#1_1].

27. Affymetrix: *Data Sheet for Genome-Wide Human SNP Array 6.0*. 2009.

28. Affymetrix: *Package Insert for Mapping 100K Array Set*.

29. Binder H, Fasold M, Glomb T: **Mismatch and G-stack modulated probe signals on SNP microarrays.** *PLoS One* 2009, **4**.

30. Oliphant A, Barker D: **BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping**. *Biotechniques* 2002, **32**:56-8.

31. **Agilent SurePrint Technology**
[http://www.chem.agilent.com/Library/technicaloverviews/Public/5988-8171en.pdf].

32. **Performance comparison of Agilent's 60-mer and 25-mer in situ synthesized oligonucleotide microarrays**
[http://www.blossombio.com/pdf/services/Agilent60merand25merMicroarray.pdf].

33. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH: **Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies.** *American Journal of Human Genetics* 2010, **86**:749-64.

34. Binder H, Preibisch S, Berger H: **Calibration of microarray gene-expression data**. *Methods in Molecular Medicine* 2008, **576**.

35. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F: **A model based background adjustment for oligonucleotide expression arrays**. *John Hopkins University, Department of Biostatistics Working Paper* 2003, **1**.

36. Binder H, Preibisch S: **Specific and nonspecific hybridization of oligonucleotide probes on microarrays.** *Biophysical Journal* 2005, **89**:337-52.

37. Binder H: **Thermodynamics of competitive surface adsorption on DNA microarrays - theoretical aspects**. *Journal of Physics Condensed Matter* 2006, **18**:S491-S523.

38. Held GA, Grinstein G, Tu Y: **Modeling of DNA microarray data by using physical properties of hybridization**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:7575-80.

39. Hekstra D, Taussig AR, Magnasco M, Naef F: **Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays**. *Nucleic Acids Research* 2003, **31**:1962-68.

40. Burden CJ, Pittelkow YE, Wilson SR: **Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays**. *Statistical Applications in Genetics and Molecular Biology* 2004, **3**:35.

41. Binder H, Preibisch S: **GeneChip microarrays—signal intensities, RNA concentrations and probe sequences**. *Journal of Physics: Condensed Matter* 2006, **18**:S537-66.

42. Binder H, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: theory and algorithm.** *Algorithms for Molecular Biology* 2008, **3**:12.

43. Binder H, Krohn K, Burden CJ: **Washing scaling of GeneChip microarray expression.** *BMC Bioinformatics* 2010, **11**:291.

44. Burden CJ, Binder H: **Physico-chemical modelling of target depletion during hybridization on oligonulceotide microarrays.** *Physical Biology* 2010, **7**:016004.

45. Binder H, Krohn K, Preibisch S: **"Hook"-calibration of GeneChip-microarrays: chip characteristics and expression measures.** *Algorithms for Molecular Biology* 2008, **3**:11.

46. Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping.** *Trends in Genetics* 2005, **21**:466-75.

47. Zhang L, Miles M, Aldape K: **A model of molecular interactions on short oligonucleotide microarrays**. *Nature Biotechnology* 2003, **21**:818-28.

48. Preibisch S: **Sequenzspezifische Signalanalyse von Genexpressionsdaten**. *Diploma Thesis* 2006.

49. Fasold M: **Methods for genomic tiling array data analysis using thermodynamic models of RNA-RNA interactions**. *Diploma Thesis* 2008.

50. Naef F, Magnasco M: **Solving the riddle of the bright mismatches: hybridization in oligonucleotide arrays**. *Physical Review E* 2002, **68**:11906-10.

51. Binder H, Kirsten T, Loeffler M, Stadler P: **Sequence specific sensitivity of oligonucleotide probes**. *Proceedings of the German Bioinformatics Conference* 2003, **2**:145-47.

52. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proceedings of the National Academy of Sciences* 2004, **101**:6062-67.

53. Binder H, Preibisch S, Kirsten T: **Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays**. *Langmuir* 2005, **21**:9287-302.

54. Binder H, Kirsten T, Loeffler M, Stadler P: **The sensitivity of microarray oligonucleotide probes - variability and the effect of base composition**. *Journal of Physical Chemistry B* 2004, **108**:18003-14.

55. Binder H, Ulbricht C, Fasold M, Brücker J: **Intrinsic metrics for hybridization control and global expression profiling – the fruit fly developmental time series**. 2009.

56. **Gene Logic Scientific Studies** [http://www.genelogic.com/support/scientific-studies].

57. Binder H, Brücker J, Burden CJ: **Nonspecific hybridization scaling of microarray expression estimates: a physicochemical approach for chip-to-chip normalization.** *The Journal of Physical Chemistry* 2009, **113**:2874-95.

58. **GeneChip® Human Mapping 100K Set** [http://media.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf].

59. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo

X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Warrington JA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R: **The External RNA Controls Consortium: a progress report.** *Nature Methods* 2005, **2**:731-4.

60. Lee J, Hever A, Willhite D, Zlotnik A, Hevezi P: **Effects of RNA degradation on gene expression analysis of human postmortem tissues.** *FASEB Journal* 2005, **19**:1356-8.

61. Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, Bareil C, Candeil L, Fraslon C, Conseiller E, Granci V, Mazière P, Kramar A, Ychou M, Pau B, Martineau P, Molina F, Del Rio M: **Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality.** *Journal of Biotechnology* 2007, **127**:549-59.

62. Dumur CI, Nasim S, Best AM, Archer KJ, Ladd AC, Mas VR, Wilkinson DS, Garrett CT, Ferreira-Gonzalez A: **Evaluation of quality-control criteria for microarray gene expression analysis.** *Clinical Chemistry* 2004, **50**:1994-2002.

63. Popova T, Mennerich D, Weith A, Quast K: **Effect of RNA quality on transcript intensity levels in microarray analysis of human post-mortem brain tissues.** *BMC Genomics* 2008, **9**:91.

64. Fleige S, Pfaffl MW: **RNA integrity and the effect on the real-time qRT-PCR performance.** *Molecular Aspects of Medicine* 2006, **27**:126-39.

65. Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, Marcucci U, Kornacker K: **Chipping away at the chip bias: RNA degradation in microarray analysis.** *Nature Genetics* 2003, **35**:292-3.

66. Viale A, Li J, Tiesman J, Hester S, Massimi A, Griffin C, Grills G, Khitrov G, Lilley K, Knudtson K, Ward B, Kornacker K, Chu C-Y, Auer H, Brooks AI: **Big results from small samples: evaluation of amplification protocols for gene expression profiling.** *Journal of Biomolecular Techniques* 2007, **18**:150-61.

67. Ma C, Lyons-Weiler M, Liang W, LaFramboise W, Gilbertson JR, Becich MJ, Monzon FA: **In vitro transcription amplification and labeling methods contribute to the variability of gene expression profiling with DNA microarrays.** *The Journal of Molecular Diagnostics* 2006, **8**:183-92.

68. Upton GJG, Sanchez-Graillet O, Rowsell J, Arteaga-Salas JM, Graham NS, Stalteri MA, Memon FN, May ST, Harrison AP: **On the causes of outliers in Affymetrix GeneChip data.** *Briefings in Functional Genomics and Proteomics* 2009, **8**:199-212.

69. Thompson KL, Pine PS, Rosenzweig BA, Turpaz Y, Retief J: **Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver RNA.** *BMC Biotechnology* 2007, **7**:57.

70. Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary PV, Li J, Overman KM, Atz ME, Myers RM, Jones EG, Watson SJ, Akil H, Bunney WE: **Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain.** *Biological psychiatry* 2004, **55**:346-52.

71. Weis S, Llenos IC, Dulay JR, Elashoff M, Martínez-Murillo F, Miller CL: **Quality control for microarray analysis of human brain samples: The impact of postmortem factors, RNA characteristics, and histopathology.** *Journal of Neuroscience Methods* 2007, **165**:198-209.

72. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T: **The RIN: an RNA integrity number for assigning integrity values to RNA measurements.** *BMC Molecular Biology* 2006, **7**:3.

73. Spiess A-N, Mueller N, Ivell R: **Amplified RNA degradation in T7-amplification methods results in biased microarray hybridizations**. *BMC Genomics* 2003, **4**:44.

74. Wilson CL, Miller CJ: **Simpleaffy: a Bioconductor package for Affymetrix quality control and data analysis.** *Bioinformatics* 2005, **21**:3683-5.

75. Gautier L, Cope L, Bolstad B, Irizarry R: **affy—analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**:307-15.

76. Lee Y-S, Chen C-H, Tsai C-N, Tsai C-L, Chao A, Wang T-H: **Microarray labeling extension values: laboratory signatures for Affymetrix GeneChips.** *Nucleic acids research* 2009, **37**:e61.

77. Garneau NL, Wilusz J, Wilusz CJ: **The highways and byways of mRNA decay.** *Nature Reviews Molecular Cell Biology* 2007, **8**:113-26.

78. **NetAffx® IVT Glossary** [http://www.affymetrix.com/support/help/IVT_glossary/index.affx].

79. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Research* 2002, **30**:207.

80. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A: **ArrayExpress - a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Research* 2007, **35**:D747-50.

81. Archer KJ, Dumur CI, Joel SE, Ramakrishnan V: **Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models.** *Biostatistics* 2006, **7**:198-212.

82. Archer KJ, Guennel T: **An application for assessing quality of RNA hybridized to Affymetrix GeneChips.** *Bioinformatics* 2006, **22**:2699-701.

83. The Tumor Analysis Best Practices Working Group: **Expression profiling--best practices for data generation and interpretation in clinical trials.** *Nature reviews. Genetics* 2004, **5**:229-37.

84. Raman T, O'Connor TP, Hackett NR, Wang W, Harvey B-G, Attiyeh MA, Dang DT, Teater M, Crystal RG: **Quality control in microarray assessment of gene expression in human airway epithelium.** *BMC Genomics* 2009, **10**:493.

85. Salisbury J, Hutchison KW, Wigglesworth K, Eppig JJ, Graber JH: **Probe-level analysis of expression microarrays characterizes isoform-specific degradation during mouse oocyte maturation.** *PLoS One* 2009, **4**:e7479.

86. Stalteri MA, Harrison AP: **Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips.** *BMC Bioinformatics* 2007, **8**:13.

87. Fasold M, Binder H: **Estimating RNA-quality using GeneChip microarrays.** *BMC Genomics* 2012, **13**:186.

88. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beissbarth T, Gaedcke J: **Impact of RNA degradation on gene expression profiling.** *BMC Medical Genomics* 2010, **3**:36.

89. Southern E, Mir K, Shchepinov M: **Molecular interactions on microarrays.** *Nature Genetics* 1999, **21**:5-9.

90. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH: **Predicting oligonucleotide affinity to nucleic acid targets**. *RNA* 1999, **5**:1458-69.

91. Carlon E, Heim T: **Thermodynamics of RNA/DNA hybridization in high-density oligonucleotide microarrays**. *Physica A: Statistical Mechanics and its Applications* 2006, **362**:433-49.

92. Shchepinov MS, Case-Green SC, Southern EM: **Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays.** *Nucleic Acids Research* 1997, **25**:1155-61.

93. Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18**:96-104.

94. Cope L, Irizarry R, Jafee HW, Speed TP: **A benchmark for Affymetrix GeneChip expression measures**. *Bioinformatics* 2003, **1**:1-13.

95. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome Biology* 2001, **2**:1-11.

96. Affymetrix: *Affymetrix Microarray Suite 5.0*. Santa Clara, CA: 2001.

97. Affymetrix: **Guide to probe logarithmic intensity error (PLIER) estimation**. *Technical Note* 2005.

98. Upton GJG, Harrison AP: **Motif effects in Affymetrix GeneChips seriously affect probe intensities.** *Nucleic Acids Research* 2012, **40**:9705-16.

99. Upton GJ, Langdon WB, Harrison AP: **G-spots cause incorrect expression measurement in Affymetrix microarrays.** *BMC Genomics* 2008, **9**:613.

100. Sharp AJ, Itsara A, Cheng Z, Alkan C, Schwartz S, Eichler EE: **Optimal design of oligonucleotide microarrays for measurement of DNA copy-number.** *Human Molecular Genetics* 2007, **16**:2770-9.

101. Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen M-M, Lu G, Fang J, Liu W-M, Ryder T, Kaplan P, Kulp D, Webster TA: **Probe selection for high-density oligonucleotide arrays**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:11237-42.

102. Zhang L, Wu C, Carta R, Zhao H: **Free energy of DNA duplex formation on short oligonucleotide microarrays.** *Nucleic Acids Research* 2007, **35**:e18.

103. Wu C, Zhao H, Baggerly K, Carta R, Zhang L: **Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays.** *Bioinformatics* 2007, **23**:2566-72.

104. Zhang L, Yoder S, Enkemann S: **Identical probes on different high-density oligonucleotide microarrays can produce different measurements of gene expression**. *BMC Genomics* 2006, **7**:153.

105. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**:799-816.

106. Heim T, Wolterink JK, Carlon E, Barkema GT: **Effective affinities in microarray data**. *Journal of Physics: Condensed Matter* 2006, **18**:S525--36.

107. Wu Z, Irizarry RA: **Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Microarrays**. In *RECOMB'04*. SanDiego, California: 2004.

108. Affymetrix: *Affymetrix Chromatin Immunoprecipitation Assay Protocol*. 2005.

109. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308**:1149-54.

110. Emanuelsson O, Nagalakshmi U, Zheng D, Rozowsky JS, Urban AE, Du J, Lian Z, Stolc V, Weissman S, Snyder M, Gerstein MB: **Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome**. *Genome Research* 2007, **17**:886 - 97.

111. Kerkhoven RM, Sie D, Nieuwland M, Heimerikx M, De Ronde J, Brugman W, Velds A: **The T7-primer is a source of experimental bias and introduces variability between microarray platforms.** *PLoS One* 2008, **3**:e1980.

112. SantaLucia J, Hicks D: **The thermodynamics of DNA structural motifs**. *Annual Review of Biophysics and Biomolecular Structure* 2004, **33**:415-40.

113. Binder H, Kirsten T, Hofacker I, Stadler P, Loeffler M: **Interactions in oligonucleotide duplexes upon hybridisation of microarrays**. *Journal of Physical Chemistry B* 2004, **108**:18015-25.

114. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M, Sasaki M: **Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes.** *Biochemistry* 1995, **34**:11211-16.

115. Heim T, Tranchevent L-C, Carlon E, Barkema ET: **Physical-Chemistry-Based Analysis of Affymetrix Microarray Data**. *Journal of Physical Chemistry B* 2006, **110**:22786-95.

116. Burden CJ, Pittelkow YE, Wilson SR: **Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays**. *Journal of Physics Condensed Matter* 2006, **18**:5545-65.

117. Skvortsov D, Abdueva D, Curtis C, Schaub B, Tavare S: **Explaining differences in saturation levels for Affymetrix GeneChip arrays**. *Nucleic Acids Research* 2007, **35**:4154-63.

118. Kennedy GC, Matsuzaki H, Dong S, Liu W-min, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SPA, Jones KW: **Large-scale genotyping of complex DNA.** *Nature Biotechnology* 2003, **21**:1233-7.

119. Langdon W: **Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips**. *Briefings in Bioinformatics* 2009, **10**:259-77.

120. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Research* 2003, **31**:e15.

121. Bolstad B, Irizarry R, Åstrand M, Speed T: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**:185-93.

122. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249-64.

123. Wu Z, LeBlanc R, Irizarry R: **Stochastic models based on molecular hybridization theory for short oligonucleotide microarrays**. *John Hopkins University, Deptartment of Biostatistics Working Paper* 2003, **4**.

124. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:31-36.

125. Affymetrix: **Affymetrix Microarray Suite 5.0**. In *User Guide*. Santa Clara, CA: Affymetrix, Inc.; 2001.

126. Burden CJ: **Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed**. *Physical Biology* 2008, **5**:16004.

127. Deutsch J, Liang S, Narayan O: **Modeling of microarray data with zippering**. *arXiv preprint q-bio/0406039* 2004.

128. Ferrantini A, Allemeersch J, Van Hummelen P, Carlon E: **Thermodynamic scaling behavior in genechips**. *BMC Bioinformatics* 2009, **10**.

129. Kroll K, Barkema G, Carlon E: **Modeling background intensity in DNA microarrays**. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics* 2008, **77**:61915.

130. Matveeva OV, Shabalina SA, Nemtsov VA, Tsodikov AD, Gesteland RF, Atkins JF: **Thermodynamic calculations and statistical correlations for oligo-probes design.** *Nucleic Acids Research* 2003, **31**:4211-7.

131. Mulders G, Barkema G, Carlon E: **Inverse Langmuir method for oligonucleotide microarray analysis**. *BMC Bioinformatics* 2009, **10**:64.

132. Naiser T, Kayser J, Mai T, Michel W, Ott A: **Stability of a surface-bound oligonucleotide duplex inferred from molecular dynamics: a study of single nucleotide defects using DNA microarrays**. *Physical Review Letters* 2009, **102**:218301-218304.

133. Naiser T, Kayser J, Mai T, Michel W, Ott A: **Position dependent mismatch discrimination on DNA microarrays - experiments and model.** *BMC Bioinformatics* 2008, **9**:509.

134. Sugimoto N, Nakano M, Nakano S: **Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes**. *Biochemistry* 2000, **39**:11270-11281.

135. SantaLucia J: **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:1460-5.

136. Johnson W, Li W: **Model-based analysis of tiling-arrays for ChIP-chip**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:12457-12462.

137. Bruun GM, Wernersson R, Juncker AS, Willenbrock H, Nielsen HB: **Improving comparability between microarray probe signals by thermodynamic intensity correction**. *Nucleic Acids Research* 2007, **35**.

138. Gharaibeh R, Fodor A, Gibas C: **Background correction using dinucleotide affinities improves the performance of GCRMA**. *BMC Bioinformatics* 2008, **9**:452.

139. Ono N, Suzuki S, Furusawa C, Agata T, Kashiwagi A, Shimizu H, Yomo T: **An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays**. *Bioinformatics* 2008, **24**:1278-1285.

140. Deng Y, He Z, Van Nostrand J, Zhou J: **Design and analysis of mismatch probes for long oligonucleotide microarrays**. *BMC Genomics* 2008, **9**:491.

141. Furusawa C, Ono N, Suzuki S, Agata T, Shimizu H, Yomo T: **Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays.** *Bioinformatics* 2009, **25**:36-41.

142. Held GA, Grinstein G, Tu Y: **Relationship between gene expression and observed intensities in DNA microarrays - a modeling study**. *Nucleic Acids Research* 2006, **34**:e70.

143. Abdueva D, Skvortsov D, Tavaré S: **Non-linear analysis of GeneChip arrays.** *Nucleic Acids Research* 2006, **34**:e105.

144. Scherer A: *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley; 2009.

145. Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A: **A global map of human gene expression.** *Nature Biotechnology* 2010, **28**:322-4.

146. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone S-A, Sklyar N, Zhao M, Sarkans U, Brazma A: **ArrayExpress update - from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Research* 2009, **37**:D868-72.

147. Ringnér M: **What is principal component analysis?** *Nature Biotechnology* 2008, **26**:303-4.

148. Wold S, Esbensen K, Geladi P: **Principal component analysis**. *Chemometrics and Intelligent Laboratory Systems* 1987, **2**:37-52.

149. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nature Reviews Genetics* 2010, **11**:733-739.

150. **User Manual: GeneChip® 3' IVT Express Kit** [http://media.affymetrix.com/support/downloads/manuals/3_ivt_express_kit_manual.pdf].

151. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection.** *Science* 1996, **274**:998-1001.

152. Clément-Ziza M, Gentien D, Lyonnet S, Thiery J-P, Besmond C, Decraene C: **Evaluation of methods for amplification of picogram amounts of total RNA for whole genome expression profiling.** *BMC Genomics* 2009, **10**:246.

153. Lane AN, Chaires JB, Gray RD, Trent JO: **Stability and kinetics of G-quadruplex structures.** *Nucleic Acids Research* 2008, **36**:5482-515.

154. Gong P, Levicky R: **DNA surface hybridization regimes.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:5301-6.

155. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nature Genetics* 2001, **29**:365-71.

Curriculum vitae

Name:	Mario Fasold
Date of birth:	09.07.1981
Place of birth:	Dresden, Germany
Nationality:	German

EDUCATION

10/2006 – 02/2008	Diploma Thesis in Bioinformatics Chair of Bioinformatics, Universität Leipzig
08/2004 – 04/2005	Study abroad Umeå Universitet, Umeå, Sweden
09/2002 – 03/2010	Master of Arts in Mathematics Friedrich-Schiller-Universität, Jena
09/2001 – 02/2008	Diploma in Bioinformatics Friedrich-Schiller-Universität, Jena

WORKING EXPERIENCE

07/2005 – 08/2005	Student trainee Dept. of Scientific Computing, Sanofi-Aventis, Frankfurt
02/2004 – 03/2004	Internship Qualitype AG, Dresden
07/2003 – 03/2004	Student assistant Bio-Systems Analysis Group, Universität Jena
07/2002 – 08/2004	Internship Institute for Molecular Biotechnology (IMB), Jena
08/1997 – 07/2000	Course instructor Student Computing Center, Dresden

LIST OF PUBLICATIONS

2013	Fasold M, Binder H: **AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data**. *Bioinformatics* 2013, **29**:129-131.
2012	Fasold M, Binder H: **Estimating RNA-quality using GeneChip microarrays.** *BMC Genomics* 2012, **13**:186.

2011	Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S: **DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments.** *Nucleic Acids Research* 2011, **39**:W112-7.
	Binder H, Fasold M, Hopp L, Cakir V, von Bergen M, Wirth H: **Portraying high-dimensional OMICs data with individual resolution.** *CAMDA Conference 2011 Proceedings.*
	Binder H, Fasold M, Hopp L, Cakir V, von Bergen M, Wirth H: **Molecular phenotypic portraits - exploring the 'OMEs' with individual resolution.** *HIBIT Conference 2011 Proceedings.*
2010	Fasold M, Stadler PF, Binder H: **G-stack modulated probe intensities on expression arrays - sequence corrections and signal calibration.** *BMC Bioinformatics* 2010, **11**:207.
2009	Binder H, Fasold M, Glomb T: **Mismatch and G-stack modulated probe signals on SNP microarrays.** *PloS One* 2009, **4**.

LIST OF SCIENTIFIC TALKS

2012	**Technical artifacts in GeneChip microarrays.** *LIFE Science Day.* November 2012, Leipzig, Germany.
	Detecting single-nucleotide variations in small ncRNAs using next-generation sequencing. *Bio-IT World Europe Conference & Expo.* October 2012, Vienna, Austria.
	SNP detection in deep sequencing data. *10. Herbstseminar der Bioinformatik*. October 2012, Doubice, Czech Republic.
	Expression quantification. *Transcriptomics Journal Club.* April 2012, Leipzig, Germany.
2011	**Learning about RNA degradation from microarray data.** *ESF Doktorandenseminar.* June 2011, Leipzig, Germany.
	Learning about RNA degradation from microarray data. *26th TBI Winterseminar*. February 2011, Bled, Slovenia.
2010	**Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *Transcriptomics Journal Club.* March 2010, Leipzig, Germany.
	Chipologie-Basics: Signalverarbeitung und Qualitätskontrolle. *IMISE Workshop*. February 2010, Leipzig, Germany.
2009	**Models for Microarray Analysis: Sequence Effects and RNA Degradation.** *24th TBI Winterseminar*. February 2009, Bled, Slovenia.

Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

_____ _____

Datum, Ort Unterschrift