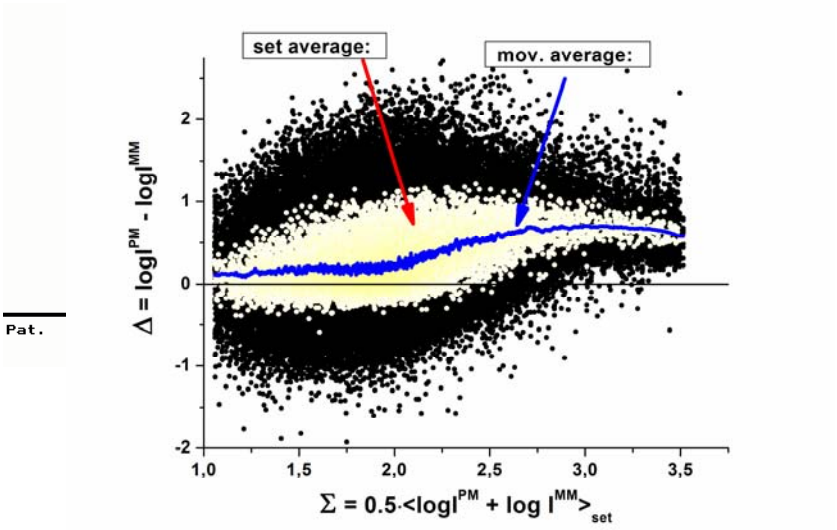
UNIVERSITÄT LEIPZIG

Leipzig Bioinformatics Working Paper
No. 16
November 2007

# Calibration of microarray gene-expression data

Hans Binder, Stephan Preibisch and Hilmar Berger




Pat.

published by the
Interdisciplinary Centre for
Bioinformatics

www.izbi.de/working_papers.html

ISSN 1860-2746

# Calibration of microarray gene-expression data

**Hans Binder[1]*, Stephan Preibisch[1,2] and Hilmar Berger[3]**

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig University, D-04107 Leipzig, Haertelstr. 16-18
[2] Max-Planck-Institute for Molecular Cell Biology and Genetics, D-01307 Dresden, Pfotenhauerstr. 108
[3] Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, D-04107 Leipzig, Haertelstr. 16-18

*corresponding author
E-mail: binder@izbi.uni-leipzig.de, fax: ++49-341-9716679

## Abstract

Calibration of microarray measurements aims at removing systematic biases from the probe-level data to get expression estimates which linearly correlate with the transcript abundance in the studied samples. The improvement of calibration methods is an essential prerequisite for estimating absolute expression levels which in turn are required for quantitative analyses of transcriptional regulation, for example, in the context of gene profiling of diseases. We address hybridization on microarrays as a reaction process in a complex environment and express the measured intensities as a function of the input quantities of the experiment. Popular calibration methods such as MAS5, dChip, RMA, gcRMA, vsn and PLIER are briefly reviewed and assessed in the light of the hybridization model and of previous benchmark studies. We present our hook-method, a new calibration approach which is based on a graphical summary of the actual hybridization characteristics of a particular microarray. Although single chip related, hook performs as well as the multi-chip related gcRMA, presently one of the best state-of-the-art methods for estimating expression values. The hook method in addition provides a set of chip summary characteristics which evaluate the performance of a given hybridization. The algorithm of the method is briefly described and its performance is exemplified.

**Key words:** gene expression, microarray calibration, preprocessing methods, transcript concentration, hook curve, hybridization, Langmuir isotherm

## 1. Introduction

In this article we emphasize on GeneChip microarray data analysis after the chips have been hybridized, scanned and the images have been summarized into hundred-thousands of probe-intensity values. With this enormous amount of data, we need standardized systems and tools for data management in order to analyze the results in a proper and sound way, as well as to be able to benefit from other publicly available gene expression data sets.

The basic principle of microarray experiments relies on the fluorescence intensity measurement for an individual probe to infer the transcript abundance specific for a selected gene. This relationship raises several difficult issues to properly extract the expression degree from the measured intensity. Calibration of microarray measurements aims at removing consistent and systematic sources of variations to allow mutual comparison of measurements acquired from different probes, arrays and experimental settings. Calibration is also called preprocessing because it usually constitutes the first step in the microarray analysis pipeline. It potentially influences the results of all subsequent steps of "higher-level" analyses as well as the biological interpretation of these results, and is therefore a crucial step in the processing of microarray data.

The chapter is organized in three parts: (1) As the essential premise for evaluating existing and developing new calibration methods we acknowledge hybridization on microarrays as a reaction process in a complex environment and express the measured intensities as a function of input quantities of the experiment. (2) Over the past years, microarray preprocessing has adapted a few generally accepted methodologies. In the second part we briefly review these options in regard to the underlying hybridization process and judge advantages and disadvantages in the light of previous benchmark studies. As we focus on Affymetrix GeneChip arrays, special attention is dedicated to the question whether a mismatch-based chip design provides benefits for intensity calibration. (3) Finally, we present our hook-method, a new calibration approach which is based on a graphical summary of the hybridization characteristics of each microarray. It uses a sort of natural metrics for intensity calibration with the potential to estimate expression values on an absolute scale. We briefly describe the algorithm and exemplify its performance.

## 2. Calibration of microarrays

The microarray experiment aims at estimating the "expression degree" of thousands specific target sequences using the integral intensity response of the respective probe spots on the chip surface. The detected intensity is affected by parasitic effects owing to the "technical" variability of repeated measurements and systematic biases which disturb the one-to-one relationship between the input and the output quantity of the measurement (6).

The task of making estimates of the input quantity of a measurement from observations of its output is called calibration. Firstly, it requires the determination of the model describing the basic relationship between the probe intensity and the specific transcript concentration under consideration of all relevant parasitic effects which should be straightened out. Secondly, the magnitude of these effects should be estimated using the intensity information of a given chip or of a series of chips, and, thirdly, one needs practicable algorithms which estimate the expression degree from the intensity values.

### 2.1. The Langmuir-hybridization model

The hybridization of a microarrays probe (P) can be described by the following reversible second-order reactions referring to specific (S) and non-specific (N) target binding, respectively:

$$P^f + S^f \rightleftharpoons PS \quad and \quad P^f + N^f \rightleftharpoons PN \quad . \tag{1}$$

Accordingly, free cRNA (or cDNA) fragments with completely complementary ($S^f$) and partly-complementary ($N^f$) sequences in solution compete for duplex formation (PS and PN) with free DNA-probe-oligonucleotides attached to the chip surface ($P^f$).

The equilibrium constants for specific and non-specific hybridization characterize the "affinity" of the respective targets for duplexing with the probe,

$$K^S \equiv \frac{[PS]}{[P^f]\cdot[S^f]} \quad and \quad K^N \equiv \frac{[PN]}{[P^f]\cdot[N^f]} \quad . \tag{2}$$

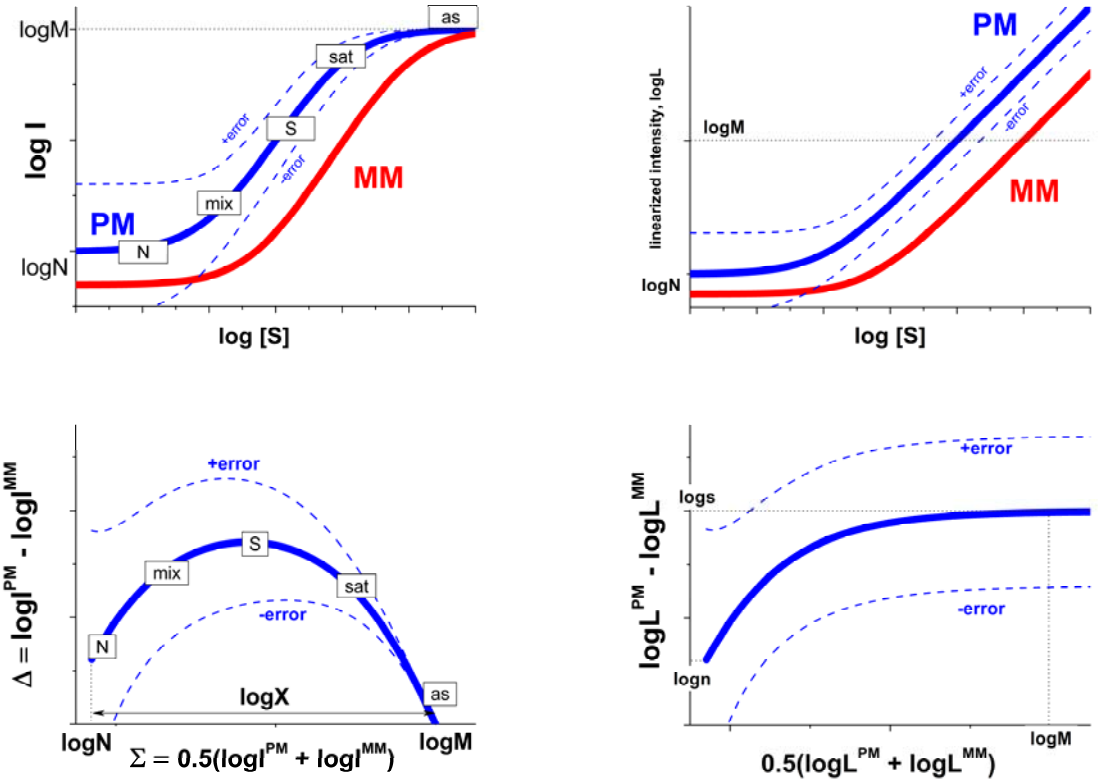



**Figure. 2.1:** Langmuir hybridization isotherm (left part, Eq. (5)) and linearized isotherm (right part, Eq. (7)) of PM and MM probes. The row of figures below shows the "hook"-plot in Δ-vs-Σ coordinates. Error limits are shown by dashed lines (Eq. (8)). The hybridization regimes are indicated in the left part of the figure (see text). The optical background is omitted for sake of clarity.

The brackets denote the respective concentrations. Making use of the condition of material-balance for the probe oligonucleotides, $[P]=[PS]+[PN]+[P^f]$, and assuming excess of free species, $[N]\approx[N^f]>>[PN]$ and $[S]\approx[S^f]>>[PS]$, one obtains the fraction of occupied probe oligonucleotides after insertion into Eq. (2) and rearrangement,

$$\Theta \equiv \frac{[PS]+[PN]}{[P]} = \frac{\left(K^S\cdot[S]+K^N\cdot[N]\right)}{1+\left(K^S\cdot[S]+K^N\cdot[N]\right)} \quad . \qquad (3)$$

Typically, the hybridization solution contains a very large number of non-specific fragments of different lengths and sequences. For sake of simplicity we subsume this diversity by the term $K^N\cdot[N]\equiv\sum_i K_i^N\cdot[N]_i$ referring to a single, effective species.

Our reaction scheme Eq. (1) considers only the bimolecular duplexing between probes and targets for sake of simplicity. Note that the available concentration of free probes and targets are however reduced by parasitic reactions such as bulk dimerization between different targets and intramolecular folding of probes and targets. These effects can be taken into account by substituting the equilibrium constants for the bimolecular binding reactions (Eq. (2)) by effective reaction constants depending on the reaction constants of the additional processes (see, e.g., (6) for details). Typically the effective binding rates are decreased compared with their values in the absence of parasitic reactions.

After the hybridization step free targets are removed by washing, bound targets are labeled with fluorescent markers which attach to biotinylated nucleotides inserted into the target sequences prior to hybridization. Finally, the fluorescence emission of the probe spot is scanned and processed into one intensity value. Assuming good gridding, it directly relates to the fraction of occupied probe duplexes, Θ, i.e.,

$$I = M\cdot\Theta + O \quad . \qquad (4)$$

Here M denotes the proportionality constant in intensity units and O the "optical" background referring to the residual intensity measured in the absence of bound transcripts owing to, e.g., adsorbed free labels or the dark current of the detector.

### 2.2. Probe and chip design

On GeneChip-expression arrays each gene is interrogated by a set of $N_{set}$=11 to 20 probe pairs. Each of them consists of a perfect match (PM) and a mismatch (MM) version. The PM sequence perfectly matches a segment of the target gene with a length of 25 nucleotides. The MM sequence is identical to that of the corresponding PM probe except the middle (13th) base which is changed to its Watson-Crick complement. The MM probes intend to estimate the background of the respective PM. The probe set forms a series of pseudo-replicates probing the same target with different probe-sequences to increase the certainty of the expression estimate.

GeneChip microarrays can be viewed as a sort of multi-photometer chips each of which assembles about $10^5 - 10^6$ virtually independent dual-channel micro-photometers on an area of about 1 $cm^2$. This analogy implies that each PM-probe spot constitutes the "sample"-channel for detecting RNA fragments of a given sequence whereas the MM-spot serves as the "reference" channel for non-specific background correction. The appareteous function given by Eq. (4) applies to each of these "micro-photometers" however with different, sequence- and transcript-specific parameters-values. With (3) one obtains

$$I^P_{p,c} = \frac{L^P_{p,c}}{1 + M_c^{-1} \cdot L^P_{p,c}} + O_c \quad with \quad L^P_{p,c} = S^P_{p,c} + N^P_{p,c} \quad . \tag{5}$$

The concentration dependence of the intensity of a PM/MM probe pair is illustrated in Fig. 2.1. In Eq. (5), probe-related properties are indexed by the superscripts P=PM, MM to account for the probe type, and by the subscripts p=1,…$N_{probe}$ and c=1,…$N_{chip}$ for the probe- and chip-effects in terms of the probe-number on the chip and the chip-number in a series of microarray hybridizations, respectively. Each gene/transcript is subsumed in the chip effect because its expression degree is a sample- and thus chip-related property. $L^P_{p,c}$ is the linear approximation of the amount of target binding in intensity units. It additively decomposes into contributions due to non-specific and specific hybridization, $N^P_{p,c}$ and $S^P_{p,c}$, respectively. The latter term can be further split into factors characterizing the affinity ($A_p^P$) and the expression degree ($E_c$) according to Eq. (3),

$$S^P_{p,c} = A^P_p \cdot E_c = M_c \cdot K^{P,S}_p \cdot [S]_c \quad . \tag{6}$$

The right hand side of Eq. (6) refers to an absolute scale where the expression degree is given in concentration units (material per volume, e.g., mole per litre). Note that the binding constant defines the concentration of "half-occupancy" at which 50% of the probe-oligonucleotides become occupied in the absence of non-specific hybridization (see Eq. (3) with $K_p^S \cdot [S] \cdot = 1$ and [N]=0). Contrarily, the middle part of Eq. (6) defines the expression degree and affinity in arbitrary units with an uncertainty of a constant factor.

Microarray calibration experiments using sets of spiked-in transcripts at different concentrations confirmed the predicted non-linear intensity response to a good aproximation (15-19). This hyperbolic function levels off into an intensity asymptote of $I^P_{p,c} \rightarrow M_c + O_c$ upon saturation of the probe spots with bound transcripts at high transcript concentrations (see Fig. 2.1). It can be "linearized" provided the asymptotic and optical background values are known,

$$L^P_{p,c} = \frac{I^P_{p,c} - O_c}{1 - M_c^{-1} \cdot \left( I^P_{p,c} - O_c \right)} \quad . \tag{7}$$

This transformations is illustrated in the rigth part of Fig. 2.1.

### 2.3. Calibration error: Linear or logarithmic scale

The raw intensity data are highly "noisy". Application of simple error propagation formalisms to Eq. (5) provides the intensity error in the linear and logarithmic scales

$$\begin{aligned} e &\equiv \delta(I-O) \approx \pm\sqrt{(\delta b)^2 + ((I-O)\cdot \delta g)^2} \\ \log e &\approx \delta \log(I-O) = \pm\sqrt{\left(\frac{\delta b}{(I-O)}\right)^2 + (\delta g)^2} \end{aligned} \quad . \tag{8}$$

It splits into an additive contribution due to fluctuations of the transcript concentrations and the optical background, $\delta b \propto \delta[S] \propto \delta[N] \propto \delta O \in N(0,\sigma_b)$; and into a multiplicative term caused by variations of the binding affinity, $\delta g \propto \delta \log K \in N(0,\sigma_g)$. The former term dominates at small intensities whereas the multiplicative contribution is the most significant source of variation at higher intensities. Most of the available data analysis algorithms assume a homoskedastic, intensity-independent Gaussian error. The linear scale meets this assumption at small intensities but progressively underestimates the error with increasing signal. In turn, log-transformed data underestimate the error at low intensities. Mostly, relevant expression values refer to medium and higher intensity levels. Therefore for most purposes the data analysis is more adequately performed in log-scale than in the linear scale.

An apparently better alternative makes use of the so-called generalized logarithm, $\mathrm{g\,log(x)} \equiv \log\left(\frac{1}{2}\left(\mathrm{x}+\sqrt{\mathrm{x}^2+\mathrm{c}}\right)\right)$. It behaves linearly at small and logarithmically at high arguments ensuring a virtually constant error width, $g \log e \approx \delta g$. However, its proper use requires scaling of the argument and of the parameter c (20, 21).

Note that the standard deviations of the considered distributions are only constants in the absence of saturation. Otherwise the error width decreases with progressive saturation at high intensities according to:

$$\sigma_g \approx \left(1 - M^{-1} \cdot (I - O)\right) \cdot \sigma_g^0 \quad and \quad \sigma_b \approx \sqrt{\left(\left(1 - M^{-1} \cdot (I - O)\right) \cdot \sigma_c^0\right)^2 + \left(\sigma_O^0\right)^2} \quad . \tag{9}$$

The error limits of the hybridization isotherm are illustrated in Fig. 2.1 by dashed lines.

### 2.4. Reference probes: MM or half-price solution

The use of MM probes as background-reference for the PM probes, as originally intended, brings up two practical problems: (i) for a considerable fraction of probe-pairs the MM fluoresce brighter than the PM. This observation appears "unphysical" because MM probes are assumed to bind transcripts at maximum in equal but never in higher amounts than the PM. (ii) The MM probe intensities on the average scatter stronger than that of the PM.

As a consequence, calibration algorithms either empirically attenuate the MM intensity values to ensure strictly positive PM-MM intensity differences or they deal completely without MM data (see below). Half-price solutions for chips without MM are proposed to replace the "superfluous" MM by additional PM-probes (22). New GeneChip generations such as the Exon 1.0 arrays are designed as PMonly chips without MM-probes. However, intensity calibration of microarray data is still a challenging tasks and the question whether the use of internal reference probes such as the MM can bring some real benefit into chip analysis is not answered yet.

For example, the problem of bright MM can be rationalized in terms of the "reversed" base pairings that forms the complementary middle bases of the PM and MM probe sequences upon non-specific hybridization and of the purine-pyrimidine asymmetry of binding strengths of RNA/DNA interactions (23, 24). Also the variability problem of the MM probes can be, at least partially, explained on the level of base-pairings of the middle base: In the MM it changes from a complementary Watson-Crick-pairing in the non-specific duplexes into a mismatched pairing in the specific duplexes whereas the respective pairing of the PM remains virtually unchanged (23, 24).

Below we present a new calibration algorithm which explicitly accounts for these effects. Moreover, this "hook"-method uses the MM probes not only as background reference but it interprets them as a sort of "weak" PM which also respond to specific hybridization according to Eq. (5). In this approach the MM operate as a hybridization reference over the full range of transcript concentrations. This way they enable the scaling of the intensities in a natural metrics system. We suggest that this idea opens a new view on the potential design and use of mismatched reference probes.

### 2.5. The calibration tasks

The intensity contribution due to specific hybridization, S, measures the expression degree on a relative scale. Consequently, the inversion of Eq. (5) with respect to S and the solution of Eq. (6) with respect to E (or [S]) furnishes a starting point to discuss the essential tasks for calibrating probe level data. It implies the need for estimating:

(i) the background contributions, N and O;

(ii) the sequence-specific affinities $K^S/A$ and $K^N$ affecting N and S, respectively; and

(iii) the degree of saturation in terms of the saturation parameter M
for correcting the intensity of each probe.
Microarray intensity data are noisy with non-gaussian frequency distributions. Proper calibration requires therefore also the consideration of
(iv) appropriate error models based on the frequency distribution of the intensities and of their specific and nonspecific contributions (Eq. (5)).
The special design of GeneChip arrays raises two additional tasks for probe intensity calibration, namely
(v) the aggregation of the individual probe-level expression values of one probe set into one transcript-related expression value; and
(vi) the proper use of the MM probes to adjust the PM data.
Usually, the expression measure E is given in arbitrary units which are related to the special conditions of a particular hybridization. For comparison with other chips calibration therefore requires finally
(vii) adjustment of the chip-related expression measures into one common scale which is, ideally, the absolute scale of transcript abundance in concentration units.

## 3. Preprocessing: state of the art

Microarray data calibration is usually called preprocessing because it is performed prior to higher level statistical analysis, such as differentially expressed genes selection. A preprocessing method for GenChips typically consists of three basic "ingredients": background correction, normalization and summarization. The background correction step is typically done in an attempt to remove non specific binding and the optical background; the normalization step reduces systematic variation between chips and the summarization step generates an expression value for each gene/probe set. Background correction typically uses information only from one array, normalisation makes a series of arrays comparable and summarization can be performed on the basis of single-chip and multi-chip data.
Numerous algorithms exist for the steps dealing with one or several of the calibration tasks specified in the previous section. Many of them can be applied in different combinations and order providing numerous potential preprocessing methods with apparently little consensus as to which is the most suitable. In the next sections we give a short overview over some of the most popular methods and review their performance on the basis of the results of different benchmark studies.

### 3.1. Linear approximations

The linear approximation of Eq. (5), $I^P_{p,c} \approx L^P_{p,c} + O_c = S^P_{p,c} + N^P_{p,c} + O_c$, neglects saturation at high transcript concentrations. It is used in basically all popular preprocessing methods: Microarray Suite 5 (MAS5, (25)), robust multi-array analysis (RMA, (26, 27)), gcRMA (28), dChip (29), probe logarithmus intensity error (PLIER, (30)) and variance stabilization normalization (vsn, (20)).
The kernel of these methods except vsn essentially deals with the baseline correction and summarization steps which, in principle, can be independently combined with stand-alone normalization algorithms such as quantile (31), global mean (32), loess or invariant probe set normalizations (29) (see below). In contrast, vsn provides baseline-corrected and normalized probe-

**Table 3.1:** Comparison of preprocessing methods with respect to background correction, scaling of the expression values and chip-processing. The grey areas highlight adequate and useful approaches with respect to probe-specific effects, error propagation and single-chip analysis.

| | vsn | RMA | gcRMA | PLIER | dChip | MAS5 | hook |
|---|---|---|---|---|---|---|---|
| **background** | global | | specific | | | | |
| **scale** | glog | log | | glog | lin | log | glog |
| **# of chips** | multi | | | | | single | |

level expression values which can be further processed with any stand-alone summarization algorithm such as median polish (see below). To straighten the discussion we understand by "method" the complete processing pipeline starting from raw intensity data and ending up with transcript related expression values.

Available algorithms can be roughly divided into global and probe-specific baseline-correction algorithms (see Table 3.1 for an overview). RMA and vsn, referring to the former group, correct all probe intensities of a selected microarray by one common background whereas the other algorithms estimate a specific background value for each probe, partly, using the MM probe intensities. For summarization all methods, except MAS5, process a series of chips in parallel. The obtained expression values are consequently context-sensitive and require reprocessing upon elimination, substitution or addition of arrays in the respective series. The methods can also differ with respect to the used error model which fits the data either in linear, log- or glog-scale.

In the following we outline the algorithmic backbone of the selected methods:

*Microarray Analysis suite 5 (MAS5):* MAS5 is a single-chip background and summarization method. It performs background correction in two steps: Firstly, the optical background is estimated by dividing the chip surface into a 4x4 grid, taking the average over the 2%-weakest intensities within each zone and subtracting an interpolated background depending on the x-y-position of each probe to account for spatial inhomogenities. Secondly, the MM-intensities serve as estimates for the N-contribution, $S^{MAS5}= I^{PM} - I^{MM}*$, where however "bright" MM are substituted by "representative" values $I^{MM}*$ which transform negative differences ($I^{PM} – I^{MM}$) into small positive ones ($I^{PM} – I^{MM}*)\geq 0$ to obtain strictly positive specific signals for each probe, $S^{MAS5}\geq 0$. Finally, the $S^{MAS5}$-values were transformed into log-scale and summarized for each probe set using One-Step Tukey's Biweight median which effectively removes signals with large median absolute deviations. Besides the expression measure MAS5 calculates the detection call, a useful qualitative value, which indicates whether a transcript is reliably detected (present) or not detected (absent). MAS5 uses global normalization as standard, which simply rescales the log-intensities of each probe by a chip-specific factor that ensures agreement between all chip-averages in the considered series.

*dChip*: Two alternative versions of this method provide either PMonly- or PM-MM-estimates of the expression degree using the equations $I_{pc}^{PM}=A_p^{PM}\cdot E_c +B_p+e$ or $I^{PM}- I^{MM}=A_p^{PM-MM}\cdot E_c +e$ to fit the respective intensities by non-linear least squares (e is the additive error term). The model assumes equal background contributions of the PM and MM on all chips of a series including also the optical contribution, $B_p=N_p+O$ with $N_p=N_p^{PM}=N_p^{MM}$. The method constrains the squared set average of the affinity to unity, $\langle A_p^2\rangle_{p\in set} =1$, with the consequence that the expression degree is obtained as the affinity-weighted average of the specific signal over the probeset, $E_c=\langle S_{c,p}\cdot A_p\rangle_p$ with larger weights given to high-affine probes. dChip uses invariant-set normalization as standard: This method selects a subset of PM-probes with small rank-differences of their intensities in a series of arrays, calculates an intensity-dependent correction curve from this subset which is then applied to all probes.

*Robust multiarray analysis (RMA):* To get strictly positive expression estimates ($S\geq 0$) RMA decomposes the frequency distribution of the intensities into an exponential signal ($P^S(S)\sim\exp(-\alpha\cdot S)$) and a gaussian background ($P^B(B)\sim N(B, \mu_c,\sigma_c)$) distribution: $P^I(I_p) = P^B(B)\cdot P^S(S_p)$. The distribution parameters $\alpha$, $\mu_c$ and $\sigma_c$ are estimated from the chip data. The background-corrected signal referring to a given intensity is then obtained as the weighted average over the background and signal distributions with the constraint $S_p^{RMA} = I_p - B_c^{RMA} \geq 0$: $B_c^{RMA} = \mu_c + \sigma_c \cdot \left(\sigma_c \cdot \alpha - \Delta\phi\right)$ ($\Delta\phi$ is the difference of normalized error functions). Summarization is performed by the fit of the log-transformed specific data of each probe set in a series of chips to the additive model, $\log(S_{pc}^{RMA})=\log E_c^{RMA}+\log A_p^{RMA}+\log e$, using median polish to minimize the residual log-error. The used constraint $\text{Median}(\log A_p^{RMA})_{p\in set} = 0$ results in expression measures which are roughly related to the median of the log-signal, i.e. $\log E_c\sim\text{median}(\log S_{p,c})_{p\in set}$. RMA uses quantile normalization as standard. This algorithm transforms the different intensity distributions of a chip series into one "average" one.

*gcRMA*: This method is essentially identical with RMA except for the background correction step. Here gcRMA accounts for the sequence specificity of non-specific hybridization using the intensity of pseudo-MM as "representatives" taken from a subset of the MM possessing the same GC-content as the PM probe of interest. Then the logarithm of the specific signal, $\log S^{gcRMA}$, is calculated as weighted average over the gaussian background distribution and a signal distribution following a power law. As in RMA the center of the background distribution is shrunken with respect to that of the

pseudo-MM due to correlations with the PM, i.e., $B_{p,c}^{gcRMA} = \exp\left(\rho \ln I_p^{MM} + (1-\rho)\cdot \mu_c\right)$ ($\rho$ is the coefficient of correlation between the PM and the MM data and $\mu_c$ the center of the MM-distribution).

*Variance stabilization normalization (vsn):* The vsn-approach shifts and rescales the intensity of a series of chips to transform their intensity-dependent heteroskedastic error-distribution into an intensity-independent homoskedastic one. Instead of the logarithm it uses the arcsinh-function as a special case of the glog-transformation arcsinh(x)=glog(x) with c=4 to get the background corrected signal, $\text{arcsinh}(S_{pc}^{vsn})=\text{arcsinh}((I_{pc} - B_c^{vsn})/F_0^{vsn})$. The chip-specific parameters $B_c^{vsn}$ and $F_c^{vsn}$ are obtained via maximum likelihood optimization for a subset of virtually invariant genes in a series of chips. The arcsinh-transformed probe-level expression values can then be summarized using, e.g., median polish, according to $\text{arcsinh}(S_{pc}^{vsn}) \approx \log A_p^{vsn} + \log E_c^{vsn} + \log e$ (for $S_{pc}^{vsn} > 1$).

*Probe logarithmic intensity error (PLIER)*: This method uses the MM probes for background correction and the glog-transformation for appropriate error handling. It fits the equation $S_{p,c}^{PLIER} = A_p^{PLIER} \cdot E_c^{PLIER} = e \cdot I_{p,c}^{PM} - e^{-1} \cdot I_{p,c}^{MM}$ using an outlier-resistant non-linear least squares algorithm for minimizing the error term $\log(e) = \text{glog}(S^{PLIER}) - \text{glog}(I^{PM} - I^{MM})$ with $c = 4I^{PM} \cdot I^{MM}$. The fit returns strictly positive signals $S^{PLIER} \geq 0$ for all non-negative intensities independently of the relation between the PM and MM values, i.e., including also bright MM, $I^{MM} > I^{PM}$.

For sake of completeness we will notice the existence of alternative and partly interesting approaches such as PDNN, the positional-dependent nearest neighbour method which uses a non-linear, sequence-specific model (33); TM, which is based in a very simple but effective fashion on the trimmed mean of PM-MM differences (34); FARMS, factor analysis for robust microarray summarization, a probe-specific RMA-like, multivariate approach (35); and a method based on strict signal deconvolution based expression-detection (36).

### 3.2. Benchmark criteria and calibration data

In the preceding section we briefly outlined some of the most popular preprocessing methods. The diversity of competing algorithmic approaches implies profound effects on the derived expression measures with consequences for subsequent higher-level statistical analysis. The correct choice of a method might depend on the scientific question being asked and on the particular experimental design and microarray data structure. Here, benchmark studies might permit users to judge each method using scientifically meaningfully summaries.

Two basic benchmark criteria, precision and accuracy, are essential for judging calibration methods. The precision specifies the systematic bias of the method in terms of the deviation of the expression estimates from its true (usually unknown) value. In turn, the accuracy characterizes the resolution (or "uncertainty") of the expression estimates. It is inversely related to their variability in replicate measurements.

Different test-scenarios are used for calibration/benchmark studies to model different experimental situations:

In the *Latin-Square spiked-in experiment* the concentrations of a small set of ~15 - 40 transcripts are varied in definite concentration steps in a hybridization solution containing a cell extract as a constant background (4). These calibration data are suited to assess the concentration dependence of the intensity and the performance of the background-correction- and summarization-steps. The small number of variable transcripts affecting less than 1% of the available probe sets and the Latin-square design of the experiment which cyclically permutes the spikes among the chips give rise to a rather small inter-chip variability. It makes the data not optimal for judging normalization algorithms.

Contrarily, in the *golden spike experiment* a relatively high number of transcripts referring to ~ 25% of all probe sets are hybridized on the chips without special background addition (37). The concentration of about one half of these spikes is varied in a "treatment-versus-control" design. Experiments of the golden-spike-type might help to develop new, improved normalization algorithms because the basic assumptions of global normalization methods are violated in many expression studies. Particularly, normalization methods such as quantile and global mean normalizations presume that only a small fraction of genes is differentially expressed, and that there are roughly equal numbers of up and down regulated genes. These assumptions are rather restrictive and prevent the exploration of global changes of the expression level (see below).

In *dilution experiments*, the total amount of RNA in the hybridization solution is changed in definite steps (4). In the closely related *mixing experiments* two RNA-extracts are mixed in different

proportions leaving the total amount of RNA constant (3). These types of experiments provide a good basis for studying the effect of the mutual interference between different transcript fractions in the hybridization solution on the performance of preprocessing methods.
Another approach uses *quantitative real-time PCR* (38) as the gold standard method of measuring gene expression in tissue samples for the evaluation of microarray calibration. Alternative studies analyze statistical characteristics such as the false discovery rate (34), correlations between genes (39, 40) or sources of variation between samples (41) to validate preprocessing methods on "real" data sets collected in a biomedical context. The practical relevance and consistency of the used criteria must be checked as the case arises: For example, correlation-based criteria favour methods that produce, on the average, zero correlations between randomly selected genes (39). Here methods are preferred which remove biases but unfortunately also the "valuable" expression signals.
Also *computer simulations* are an interesting option to compare preprocessing methods. However, there is the problem to avoid inherent circularities, e.g., if the data model relies on assumptions used in the analysis algorithm. For example, it is not surprising that methods ignoring probe-specific background levels perform well on data synthesized without probe-specific background contributions (42). Therefore results from simulation studies must be critically reflected in the context of the actual simulation design.

### 3.3. Which method is the best?

Numerous studies have assessed preprocessing methods in a wide range of conditions to benchmark their performance. In a general sense there is apparently no "best" method which clearly outperforms the others under all circumstances. Moreover, all these methods have been proven in numerous applications to provide reasonable results.
For example, in patient-cohort studies researchers typically select sets of genes that are differentially expressed between certain known conditions (supervised approach) or they aim at detecting biological relations between samples or genes by grouping them according to their expression profiles (unsupervised approach). Often the goal is to obtain predictors for prognostically relevant categories. It has been argued that the choice of the pre-processing method has less influence on the final outcome especially in studies based on large numbers of arrays, whereas it can have important effects on the results of smaller studies (40). The existence of a certain minimum number of differentially expressed genes is obviously sufficient for predictor-selection without the need of exact quantification of the observed changes. Clearly, the reliability of such analyses will improve with the number of samples and/or with the significance level for detecting differential expression.
On the other hand, genomic regulation is governed by the specificity of molecular interactions between genomic, transcriptomic and proteomic factors, their mutual relations and levels. Particularly, the estimation of transcript levels on an absolute scale using microarrays is a challenging task which becomes necessary for exploring mechanisms of gene regulation. For these issues exact calibration and the choice of appropriate methods is an essential prerequisite.
Calibration data reproducing the basic concentration dependence of the intensity without complex inter-chip variations of the hybridizations clearly show that the non-specific hybridization-background correction is the main factor that explains differences between the methods (37-39, 43). Global background correction algorithms such as vsn and RMA obviously underestimate the level of non-specific hybridization leading to attenuated estimates of differential expression with strong negative biases especially a low expression levels. Methods with MM-corrections such as MAS5, PLIER and dChip outperform methods discarding MM data at medium and higher expression levels providing much better accuracies. On the other hand, MM-corrections give rise to highly variable expression estimates at low intensity levels with partly high false-positive detections. The right balance between accuracy and precision depends on the signal intensity with the problem that the gain in precision at low intensities must be paid by a penalty in accuracy and vice versa.
It seems that the much lower variability of RMA and vsn estimates expression of low-abundance genes in a biased, but very precise manner. Minimizing variability for biased estimates however produces a dangerous sense of confidence in potentially wrong data. Contrarily, a higher variability at low intensities at least circumvents such wrong conclusions as long the variability exceeds the bias. Here, the sequence-based background adjustment of gcRMA emerges as a method that may be the most optimum one across the whole intensity range (37, 38, 43).

Generally, one has to keep in mind that the precision of expression measures can be improved by replicate measurements and also by further developing statistical concepts, e.g., by explicit consideration of the measurement error derived from the hybridization mechanism in a probe-dependent fashion (see above). Contrarily, the accuracy of calibration methods cannot be improved by replicates. It requires the understanding of the essential factors that govern microrarray hybridization and their implementation into feasible algorithms.
All considered methods systematically underestimate the expression level at high RNA concentrations because they neglect saturation. Here, non-linear hybridization models such as the two-species Langmuir isotherm provide a more adequate concept to account for this effect. Other important challenges for the amelioration of calibration methods are the need for better probe-specific background corrections, for normalization algorithms which conserve differential expressions between the samples on an absolute scale and also for better affinity corrections for more precise data. Note that most expressed genes are not nessecarily the key players in genomic regulation. Hence, better background and affinity corrections should increase the resolution of the method to detect also relatively small changes of the expression level.

## 4. Hook-calibration: Towards absolute expression measures

Our hook-calibration method analyzes the intensity data of a given GeneChip microarray in terms of the two-species Langmuir-isotherm (Eq. (5)). The method uses the MM probe intensities as reference for the PM over the whole concentration range to discern typical hybridization regimes, namely that of predominant non-specific binding (N), mixed hybridization (mix), predominant specific binding (S), saturation (sat) and asymptotic binding (as) as illustrated in Fig. 4.1. The intensity data are aggregated into one mean hybridization characteristics called hook curve because of its characteristic shape which is predicted by the Langmuir model (see Fig. 2.1 and Fig. 4.1). The method uses the position-dependent nearest neighbour model to account for the probe-specific binding affinity upon specific and non-specific hybridization. It corrects the probe intensities for probe-specific background, affinity and saturation limit. Note that our model differs from that of Zhang et al. (33) who restricts the positional dependence by a common weight-function for the nearest-neighbour free energy terms. Our positional dependent terms are freely adjusted (see Eq. (10) below). The hook-method is a single-chip approach which provides essential hybridization summaries such as the fraction of not-expressed probe-sets (%N), the mean background intensity ($N_c^{PM}$), and the PM/MM-sensitivity gain upon specific binding ($s_c$).

### 4.1. Algorithm

The algorithm consists of the following basic steps (see also Fig. 4.1):

1) The intensity data are corrected for the optical background using the Affymetrix zone-algorithm (32).
2) The PM and MM probe intensity data are plotted into a special type of M-A-plot, where the ordinate value is the log-difference, $\Delta=\log I^{PM}-\log I^{MM}$, and the abscissa-value the set-averaged log-sum, $\Sigma=0.5<(\log I^{PM}+\log I^{MM})>_{set}$.
3) The data are smoothed using a sliding-window over ~100 probe sets along the abscissa. The obtained $\Delta$-vs-$\Sigma$ relationship is called raw *hook curve* because of its characteristic shape. It divides into four characteristic parts: the N-range referring to the relatively flat starting region, the subsequent mix-range of positive slope, the S-range near the maximum and the sat-range with negative slope beyond the maximum.
4) The intensities of the probes from the N- and S-ranges are used to fit the positional-dependent nearest neighbour model. It decomposes the log-intensity variation about its set-average into a sum of additive sensitivity-terms, $\delta\varepsilon_k^{P,h}(BB')_p$, where BB' is the couple of adjacent bases at position k and k+1 of the probe sequence (k=1,…24; BB'=AA, AT,…, CC). The model is parameterized separately for non-specific and specific binding (h=N,S) of the PM and MM (P=PM,MM), respectively (23, 24), providing thus four sets of 16 BB'-sensitivity profiles which in turn are used to calculate the affinity correction in a sequence-specific fashion,

$$\log A_{p,c}^{P,h} = \sum_{k=1}^{24} \delta\varepsilon_{k,c}^{P,h}(BB')_p \quad . \qquad (10)$$

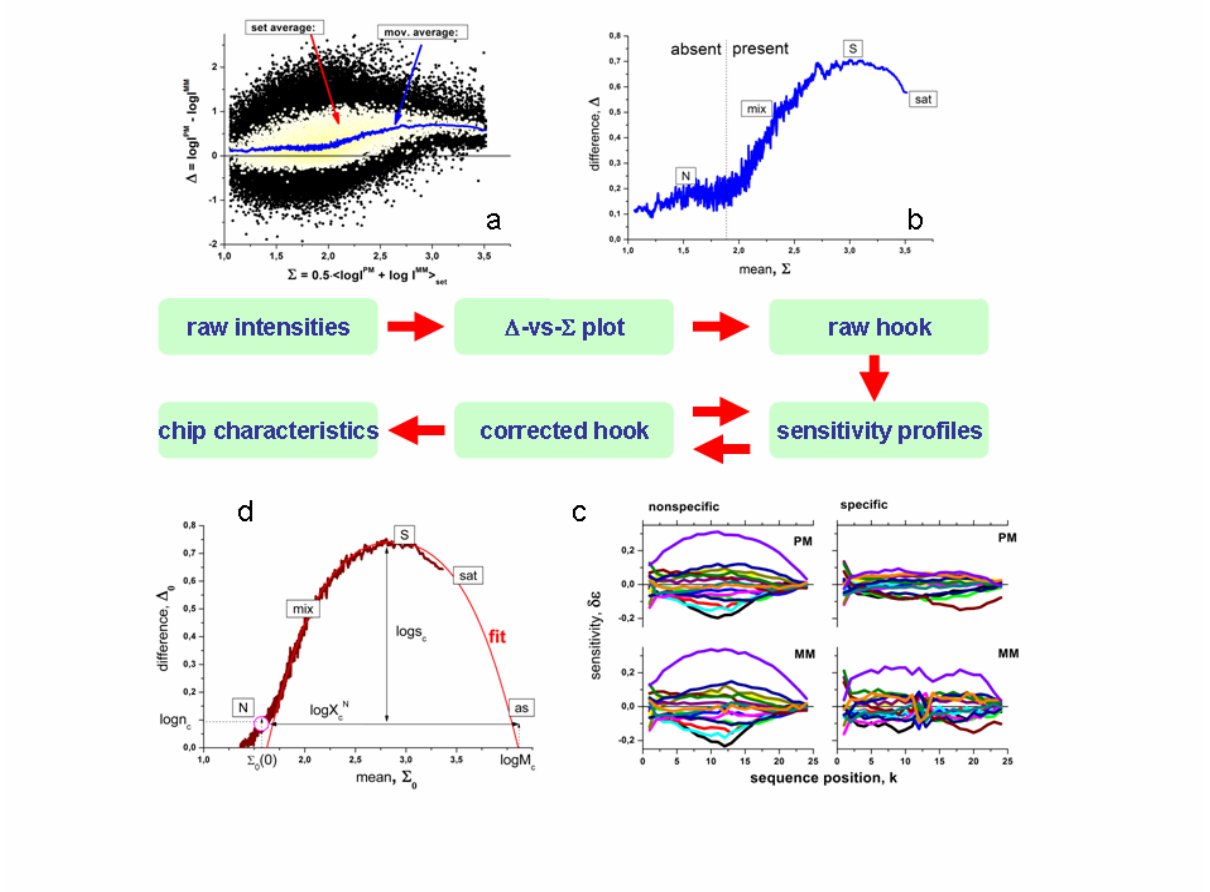



**Figure 4.1:** The hook-method: The raw intensity data of one GeneChip microarray are plotted into the Δ=log(PM/MM)-vs-Σ=1/2·log(PM·MM) coordinate system and smoothed to get the raw hook-curve. Then, probes from the N- and S-hybridization regimes are used to calculate four sets of 16 position-dependent nearest-neighbour-sensitivity profiles of the affinity model (non-specific and specific for the PM and MM each). After affinity correction of the intensities one obtains the corrected hook-curve. It is used to get improved sensitivity profiles in a second iteration step. The mix-, S- and sat-ranges of the corrected hook are well fitted using the two-species Langmuir hybridization model. The dimensions of the hook, its width and height, provide hybridization characteristics of the chip such as the binding strength of non-specific hybridization and the mean PM/MM gain of the binding affinity, respectively.

5) Then the probe intensities are corrected for sequence-specific affinities using the model adjusted in the previous step. In the mix-range we use a weighted superposition of the N- and S-contributions, $I_{p,c}^{P,corr} = I_{p,c}^{P} \cdot 10^{\left(x^S \cdot \log A_{p,c}^{P,S} + (1-x^S) \cdot \log A_{p,c}^{P,N}\right)}$, where $x^S$ is the fraction of specific hybridization contributing to the intensity.
6) The affinity-corrected intensities are used to get the corrected version of the hook-curve with the coordinates $\Sigma^{hook}$ and $\Delta^{hook}$ and an improved set of sensitivity profiles by re-iteration of steps 2.-5. Note the significant differences between the raw and the corrected hooks: Affinity correction clearly reduces the width of the N-range and also the scattering of the data in the remaining hybridization regimes.
7) The mix-, S- and sat-ranges of the corrected hook curve are fitted using the two-species Langmuir isotherm (see next section). The fit and the separate analysis of the N-range provide chip characteristics such as the mean background level ($N_c^{PM}$), the saturation intensity ($M_c$), the width and correlation coefficient of the background distribution ($\sigma$ and $\rho$) and the mean PM/MM-sensitivity gains ($n_c$ and $s_c$) which were used for calibration of the probe-level intensity data in the next step.

8) The probe-intensities are linearized using Eq. (7). Then, the probe-level expression degree is estimated as the weighted glog-average of the total signal minus the respective non-specific background contribution according to Eq. (5):

$$g\log(S_{p,c}^{PM}) = \int N(N_c^{PM}, \sigma^N) \cdot g\log\left(L_{p,c}^{PM} - 10^x \cdot N_c^{PM} \cdot A_{p,c}^{PM,N}\right) \cdot dx \qquad (11)$$

$N(N_c^{PM},\sigma^N)$ is the Gaussian distribution of the non-specific, affinity corrected PM-signal. Alternatively, we also calculate a PM-MM version by substituting the glog-term in the integral of Eq. (11) for $g\log\left(\left(L_{p,c}^{PM} - L_{p,c}^{MM}\right) - 10^x \cdot N_c^{PM}\left(A_{p,c}^{PM,N} - n_c^{-1} \cdot \left(N_c^{PM}\right)^{-\rho} \cdot A_{p,c}^{MM,N} \cdot 10^{(\rho-1)\cdot x}\right)\right)$. This approach uses the bivariate marginal distribution of the PM-MM background, where ρ denotes the coefficient of correlation between the PM and MM background intensity values.

9) The probe-level specific signals are affinity corrected according to $E_{p,c}^{PM} = S_{p,c}^{PM} \cdot \left(A_{p,c}^{PM,S}\right)^{-1}$ for the PMonly and $E_{p,c}^{PM-MM} = S_{p,c}^{PM-MM} \cdot \left(A_{p,c}^{PM,S} - s_c^{-1} \cdot A_{p,c}^{MM,S}\right)^{-1}$ for the PM-MM estimates and then summarized by means of the Turkey-biweight median to get robust transcript-level expression estimates.

### 4.2. Natural metrics of expression values

The hook-like shape of the Δ-vs-Σ dependence can be reproduced using the two-species Langmuir isotherm (see Fig. 2.1). First we applied Eq. (5) separately to the intensities of the PM and MM and then transformed the predicted intensities into Δ-vs-Σ coordinates. The obtained theoretical function fits the experimental data to a good approximation (Fig. 4.1). The hook curve considers all probes of a given chip. It consequently summarizes the properties of a particular hybridization into a sort of mean binding isotherm.

The hook curve is divided into five characteristic ranges which can be assigned to different hybridization regimes (see step 3 in the previous section and also Fig. 4.1): In the N-regime the probes hybridize almost exclusively non-specifically owing to the absence or low concentrations of specific transcripts. In the subsequent mix-regime, both, specific and non-specific transcripts significantly contribute to the observed intensity of the probes. In the S-regime the probes predominantly hybridize with specific transcripts. In the sat-regime the probes become progressively saturated with bound transcripts. This effect first and foremost affects the PM due to their higher specific binding constant. As a consequence, the concentration dependence of the intensity progressively becomes non-linear and Δ starts to decrease. In the as-range the intensities of the PM and MM reach their asymptotic values owing to complete saturation. In typical hybridizations this region is usually not reached.

Note that the Δ-vs-Σ coordinates are simply linear-combinations of the PM and MM intensities. Hence, the hook-curve can be interpreted as a special representation of the binding isotherm where the explicit dependence of the probe intensities on the (usually unknown) transcript concentrations is replaced by the (experimentally available) relation between the PM- and the MM-probe intensities. Here, the MM probes serve as an internal reference subjected essentially to the same hybridization law as the PM, however with modified characteristics. Particularly, one expects to find different binding constants for specific and, possibly, also non-specific binding. Let us denote the respective PM/MM-ratios with $s_c \equiv K_c^{PM,S} / K_c^{MM,S}$ and $n_c \equiv K_c^{PM,N} / K_c^{MM,N}$, respectively. Other hybridization characteristics are the mean background intensity of the PM due to non-specific binding, $N_c^{PM}$, and the maximum intensity, $M_c$, referring to completely saturated probe-spots.

The coordinates of the start- and the end-points of the hook-curve, and to a good approximation also its maximum, can be directly related to basic hybridization characteristics. For example, the Σ-coordinates of the start- and end-points, $\Sigma(0)\approx\log(N_c^{PM})-1/2\log(n_c)$ and $\Sigma(\infty)\approx\log(M_c)$, estimate the mean non specific background and the saturation intensity, respectively. The Δ-coordinates of the start point and of the maximum, $\Delta_s(0)\approx\log(n_c)$ and $\Delta_{max}\approx\log(s_c)+\log(n_c)$, are measures of the mean log-difference between binding constants of the PM and MM for non-specific and specific binding, respectively. Making use of these data one obtains the "width" and the "height" of the hook-curve which estimate the mean binding strength of non-specific hybridization, $\Sigma_{as}(\infty)-\Sigma_s(0)\approx \log(X_c^{PM,N})= -\log(K_c^{PM,N}\cdot[N])$, and the mean affinity gain for specific binding of the PM relatively to the MM, $\Delta_{max}-\Delta_s(0)\approx \log(s_c)$, respectively. The binding strength, $X_c^{PM,N}$ is a dimensionless measure of the concentration in units of the respective binding constant. A value of unity refers to a surface coverage

of Θ=0.5 in the absence of specific transcripts. The mean affinity gain is directly related to the free energy difference due to the replacement of the complementary Watson-Crick- with a mismatched base pairing in the respective probe/transcript duplexes (24).
In summary, the hook curve spans a sort of natural metrics system for the expression estimates. It reflects essential hybridization characteristics in terms of its geometric dimensions: width, height and "start"-coordinates.

### 4.3. Examples: Chip characteristics

Fig. 4.2 shows a collection of representative hook-curves taken from six hybridizations of human-genome chips of different generations, a plant chip (Arabidopsis Thaliana chip ATH-12501) and alternative hybridizations with cRNA and cDNA. Along the chip generations the spot-size of the probes decreases from 20 μm (U95), over 18 μm (U133A and U133Av2) to 11 μm (U133-plus2). The reduction of spot-size has enabled to increase the number of probe sets per chip from 16.000 over 22.000 to 54.000, respectively (44, 45). In addition, this development is accompanied by modifications of the reagent-kits and the scanning technique. Importantly, also probe selection has been improved by applying more sophisticated genomic and thermodynamic criteria especially after the U95-generation. The different shapes of the uncorrected hook curves of the U95 and U133 chips, particularly the broader N-range of the former one, can be explained by the partially suboptimal probe quality of the U95-generation containing a relatively high number of weak-affinity probes. For the U133 series the N-range considerably narrows essentially due to better quality of the probes. It is important to note that our affinity correction levels out this difference to a large extent providing corrected hook curves of very similar shape for the U95 and U133 chips.
The width of the fitted hook-curves estimates the binding strength of the non-specific background in "intrinsic" units of the respective binding constant (see above). A wider hook curve is equivalent with a lower level of non-specific background and thus with an increased dynamic measurement range of the probe spots. The widths of the fits shown in Fig. 4.2 indicate that this range slightly increases with the chip generations (see also Table 4.1).
In general, microarray technology takes advantage of either of two types of chemical entities as the labelled target, cRNA or cDNA, considered to be virtually equivalent for the purpose of expression analysis. Here we compare both options for illustrating the effect of the two binding "chemistries" on the chip characteristics. The substitution of cRNA by cDNA gives rise to essentially two effects (see Fig. 4.2): Firstly, it increases the dynamic range by reducing the background-level, and secondly, it reduces the variability of the uncorrected background intensity. Among the two options, affinity correction to a much less extent improves the hook-curve of the DNA-hybridization. The higher non-specific background level and variability of the RNA-hybridization were attributed to relatively-stable mismatched "G•u-wobble" base pairings in the RNA/DNA duplexes which give rise to less specific binding compared with DNA/DNA hybridizations without such stable mismatches pairings (4).
To generalize the discussed single-chip related results we collect mean values of these characteristics over experimental series taken from different studies dealing with calibration issues, biological samples, cancer specimen, different chip generations and species (Table 4.1). In essence, most of the chip characteristics provide relatively similar values for the different series despite the very heterogeneous origin of the data. The maximum intensity and the optical and non-specific background levels vary roughly over three orders of magnitude.

The PM-affinity gain parameter for specific hybridization shows that the central mismatch of the MM causes an, on the average, tenfold (s ~ 7 -11) increased affinity of the PM compared with that of the MM. Contrary, for non-specific binding one expects on the average the same affinity for the PM and MM. The respective PM/MM-gain parameter however indicates a small but significantly increased PM-affinity, n ~ 1.05 – 1.25. We tentatively attribute this effect to false positive detections in the N-range, i.e. to a certain amount of specific hybridization among the absent probes (see below). The relatively narrow distributions of hybridization characteristics reflect the common physical-chemical basics of the method, for example the oligonucleotide density and size of the probe spots, the common MM probe-design and hybridization conditions.

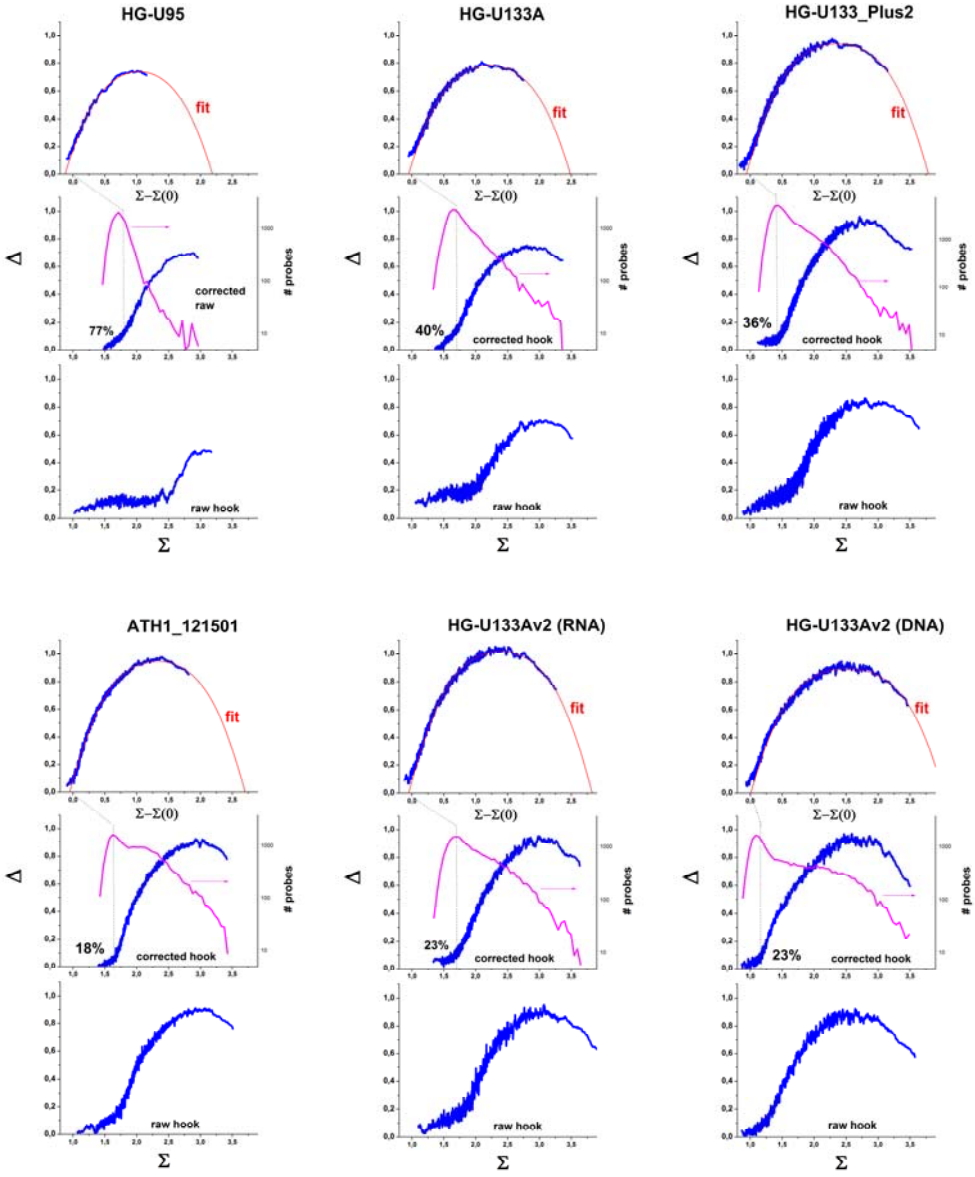



**Figure 4.2:** Hook-curves of six different microarray hybridizations: raw hook (lower panel), affinity-corrected hook and number distribution (middle) and the fit of the specific part of the hook (panel above) for human genome GeneChips of different generations (row of figures above, Affymetrix HG-U95, HG-U133 and HG_U133_plus2 taken from the spiked-in data sets (1) and mixing series (3), of a plant genome (row below; Arabidopsis Thaliana, ATH1_121501 array) and of hybridization with cRNA and cDNA (4). The vertical dotted line indicates the "break" of the hook-curve which was used to estimate the number of "absent" probe-sets given in percent for each hybridization in the figures. See also Tab. 3.1 for the mean hybridization characteristic of the respective experimental series.

**Table 4.1:** Mean hybridization characteristics of GeneChips estimated from different experimental series. The values are given as MED ± MAD where MED is the median and MAD the median absolute deviation [a)] calculated from the respective values over the experimental series in logarithmic scale ($\log_{10}$).

| Data set | Ref. | Affy Chip (# of chips) | Optical BG | Non-specific BG | Sat-intensity | N-binding strength | PM/MM-gain (S) | PM/MM-gain (N) | SD of N-BG |
|---|---|---|---|---|---|---|---|---|---|
| | | | logO | logN | logM | logX | logs | logn | $\sigma^N$ |
| **Calibration data sets:** | | | | | | | | | |
| GeneLogic Dilution | (2) | HG-U95A (74) | 1.74 ±0.13 | 1.54 ±0.22 | 4.27 ±0.15 | 2.75 ±0.20 | 1.00 ±0.05 | 0.10 ±0.015 | 0.28 ±0.008 |
| Affy spiked-in | (1) | HG-U95A (59) | 1.93 ±0.06 | 1.70 ±0.05 | 4.14 ±0.09 | 2.44 ±0.10 | 0.89 ±0.04 | 0.07 ±0.006 | 0.30 ±0.008 |
| Affy spiked-in | (1) | HG-U133A (42) | 1.47 ±0.02 | 1.54 ±0.05 | 4.20 ±0.04 | 2.66 ±0.05 | 0.85 ±0.04 | 0.08 ±0.005 | 0.29 ±0.003 |
| Barnes-Dilution | (3) | HG-U133_plus2 (12) | 1.62 ±0.01 | 1.47 ±0.08 | 4.48 ±0.03 | 3.01 ±0.10 | 1.02 ±0.03 | 0.08 ±0.005 | 0.32 ±0.0013 |
| Eklund sp-in (cRNA) | (4) | HG-U133Av2 (6) | 1.63 ±0.01 | 1.55 ±0.14 | 4.51 ±0.13 | 2.96 ±0.15 | 1.08 ±0.04 | 0.07 ±0.01 | 0.36 ±0.04 |
| Eklund sp-in (cDNA) | (4) | HG-U133Av2 (6) | 1.58 ±0.02 | 1.06 ±0.01 | 4.22 ±0.03 | 3.16 ±0.02 | 0.93 ±0.03 | 0.04 ±0.002 | 0.29 ±0.003 |
| **Patient cohort and cell line studies:** | | | | | | | | | |
| Frontal Brain | (5) | HG-U95Av2 (6) | 1.81 ±0.13 | 1.87 ±0.18 | 4.80 ±0.28 | 2.93 ±0.25 | 0.91 ±0.02 | 0.10 ±0.01 | 0.30 ±0.006 |
| Malignant Lymphomas | (7) | HG-U133A (221) | 1.84 ±0.06 | 1.96 ±0.15 | 4.49 ±0.10 | 2.43 ±0.15 | 0.85 ±0.06 | 0.09 ±0.01 | 0.35 ±0.03 |
| Colon Cancer | (8) | HG-U133Av2 (20) | 1.88 ±0.13 | 1.62 ±0.12 | 4.63 ±0.04 | 3.01 ±0.12 | 0.96 ±0.05 | 0.06 ±0.01 | 0.31 ±0.01 |
| Lymphocytic Leukemia | (9) | HG-U133_plus2 (20) | 1.60 ±0.05 | 1.29 ±0.08 | 4.32 ±0.14 | 3.03 ±0.15 | 0.87 ±0.04 | 0.06 ±0.006 | 0.30 ±0.009 |
| Renal Carcinoma | (10) | HG-U133_plus2 (47) | 1.80 ±0.11 | 1.99 ±0.09 | 4.73 ±0.09 | 2.72 ±0.10 | 0.82 ±0.03 | 0.10 ±0.006 | 0.38 ±0.02 |
| Mouse | (11) | MOE430A (33) | 1.86 ±0.05 | 1.55 ±0.11 | 4.42 ±0.12 | 2.87 ±0.11 | 0.98 ±0.03 | 0.06 ±0.01 | 0.29 ±0.008 |
| Arabidopsis | (12) | ATH1-121501 (16) | 1.84 ±0.09 | 1.41 ±0.15 | 4.46 ±0.06 | 3.01 ±0.15 | 0.99 ±0.06 | 0.03 ±0.007 | 0.26 ±0.004 |
| Yeast | (13) | Yeast-2 (41) | 1.85 ±0.06 | 1.44 ±0.07 | 4.60 ±0.07 | 3.16 ±0.10 | 1.05 ±0.04 | 0.002 ±0.03 | 0.31 ±0.03 |
| Rice | (14) | Rice (25) | 1.62 ±0.03 | 1.31 ±0.06 | 4.51 ±0.09 | 3.20 ±0.10 | 1.0 ±0.03 | 0.03 ±0.008 | 0.30 ±0.01 |

[a)] median (med(x)) and median absolute deviation: MAD=1.4·med(|x-med(x)|)) (the factor accounts for asymptotic normal consistency)

The positional-dependent sensitivity terms, δε (Eq. (10)), represent another type of chip characteristics because they are used to adjust the intensities of each microarray. Fig. 4.3 shows the sensitivity profiles of the MM probes for three of the chips taken from Fig. 4.2. Note the similar profiles of the two selected RNA-hybridizations: Generally one observes C>G>T>A for most of the sequence positions. Contrarily, for the DNA-hybridization this order changes into G>C>A≈T. The positional-dependent sensitivity terms, δε, are directly related to the binding strength of base-pairings in the probe/target-duplexes (23, 24, 46), which are basically independent of a particular hybridization but change with the chemical entity. In Fig. 4.3 we aggregated the 16 nearest-neighbour profiles into four single-base profiles for sake of clarity. In addition the maximum and minimum NN-profiles are shown. For the RNA-hybridizations, for example, adjacent CC provide the strongest intensity

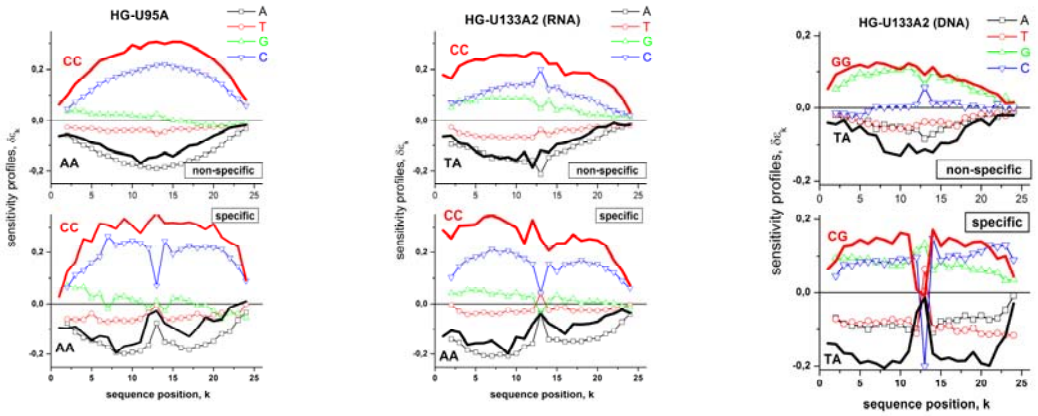



**Figure 4.3:** Sensitivity profiles of three chips shown in Fig. 4.2: Only the MM profiles for non-specific (above) and specific (below) hybridization are shown. The PM-profiles look similar to those of the non-specific MM-profiles. The 16 nearest neighbour-terms (NN) profiles are aggregated into four single-base profiles for sake of clarity (symbols). In addition each figure shows the two NN-profiles with the largest positive and negative values. The profiles of the RNA-hybridizations differ from that of the DNA-hybridization due to the different binding chemistry.

increment whereas for DNA-hybridization one gets GG and CG. Note also the "dents" in the middle of the specific MM-profiles. They reflect the effect of the mismatches on the binding strength with "molecular resolution".

## 4.4. Examples: Expression values

For further validation of the method we analyzed the Affymetrix Latin-square spiked-in and the GeneLogic dilution data sets (4). The corrected hook curves of selected chips of these series are shown in Fig. 4.4 and Fig. 4.6, respectively. The hook-curves of the spiked-in series mainly reflect the hybridization of the cell extract which was added in equal amounts to all hybridizations (Fig. 4.4). In addition, each chip contains a set of "spiked-in" probes covering the whole concentration range of the spikes (0 – 512 pM). The Δ-vs-Σ-coordinates of these spikes spread over the full range of the hook curve (see circles in Fig. 4.4). Their positions shift along the hook to the right with increasing transcript concentration. Probes without specific transcripts and probes with only tiny spiked-in concentrations accumulate mainly within the N-range of the hook curve. In a simple approximation we classify these probes as "absent" in analogy with the absent-calls calculated by MAS5 (32). The insertion in Fig. 4.4 shows that both methods, hook and MAS5, provide very similar absent-rates for the spikes. Note that the vertical shift between the MAS5 and hook data is due to the somewhat arbitrary choice of the threshold-parameters used in both methods. It can be simply reduced by appropriate adjustment.

Fig. 4.5 shows the expression measures obtained from selected preprocessing methods as a function of the spike-in concentration. Perfect calibration refers consequently to a diagonal line of slope unity in this double-logarithmic plot. The hook and gcRMA methods clearly outperform MAS5 and RMA with respect to this criterion. Note that the reduced slope of the RMA-curve indicates a systematic bias which underestimates differential expression roughly by the square root of the true change, $FC^{RMA} \approx (FC^{true})^{0.5}$. Fig. 4.5 also reveals that saturation gives rise to the flattening of all curves at high concentrations except that of the hook method which corrects the data for this effect.

Dilution of the hybridization solution in the dilution series gives rise to the progressive shift of the N-range of the hook curve towards smaller abscissa values leaving the position of the asymptotic as-range unchanged (Fig. 4.6). The associated "widening" the curve is compatible with the global decrease of the transcript-concentration in this experiment (see above). This trend is also paralleled by the disappearance of the "sat"-range, i.e., dilution globally decreases the occupancy of the probes.

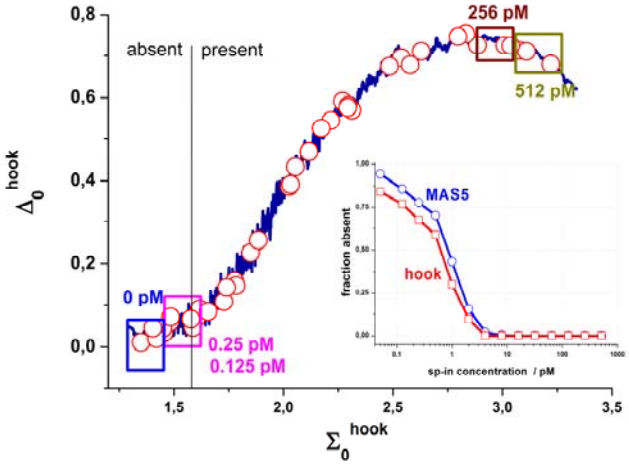



**Figure 4.4:** Hook curve of one spiked-in hybridization (HGU-133A). The open circles refer to the spiked-in probes. Their positions move along the hook to the right with increasing spiked-in concentration of the respective specific transcripts (see figure). The vertical line indicates the break-point between the N- and mix-regimes which classifies the probes into absent and present ones. The insertion shows the fraction of absent probes as a function of the spiked-in concentration obtained from the hook- and the MAS5-methods.

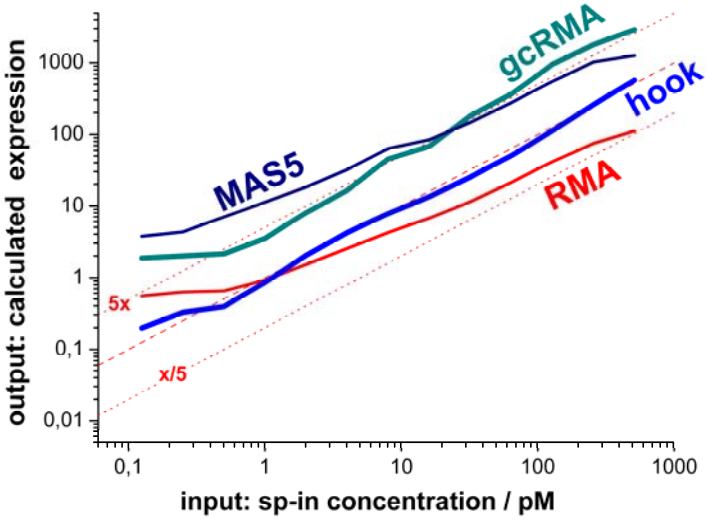



**Figure 4.5:** Mean expression degree of all spike-in probe sets as a function of the spiked-in concentration: The comparison of different preprocessing methods (see figure) shows that the single-chip hook method performs roughly as well as the multi-chip method gcRMA. The diagonal lines of slope one refer to optimum calibration. The dotted diagonals indicate fivefold changes with respect to the dashed diagonal line. The smaller slope of MAS5 and especially of RMA compared with that of hook and gcRMA indicate the accuracy-penalty of these methods. Note that the MAS5 and gcRMA curves are vertically shifted for sake of clarity.

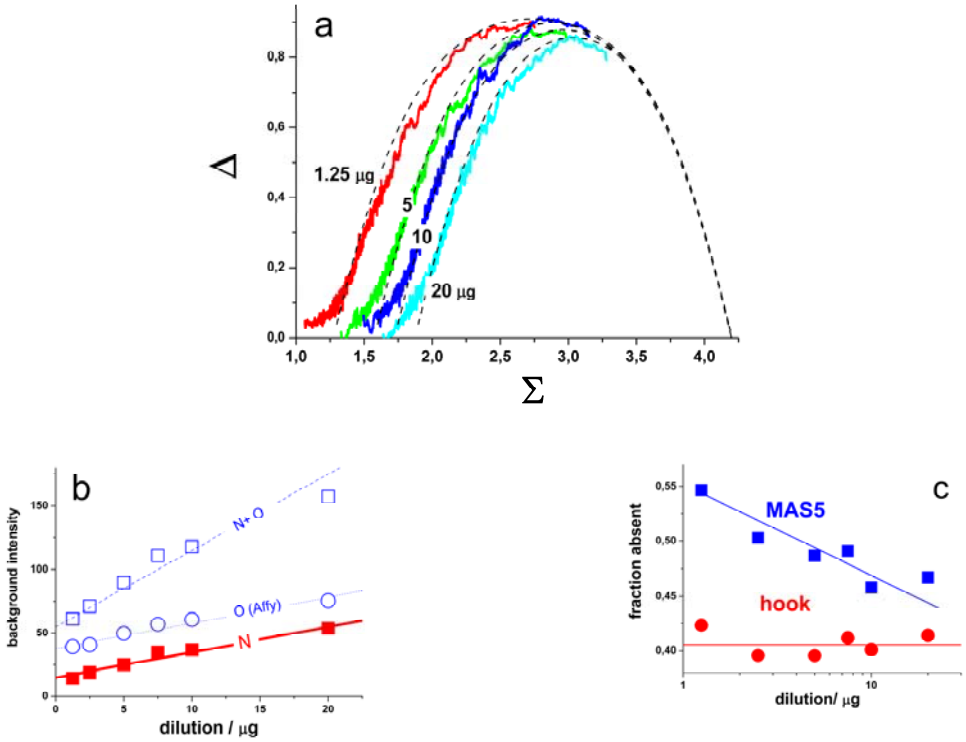



**Figure 4.6:** Part a: Hook curves of the dilution experiments for different amounts of RNA (see figure). The dashed curves are calculated using the two-species Langmuir isotherm assuming a common asymptotic maximum intensity value. Upon dilution, the position of the left branch of the hook shifts to smaller abscissa values indicating the decrease of non-specific hybridization. Part b of the figure shows the background level upon dilution: The total background (N+O) decomposes into contributions due to the optical effects (O) and non-specific hybridization (N). Part c shows that the hook method provides a virtually constant fraction of absent probes upon dilution whereas MAS5 probably progressively overestimates absent calls.

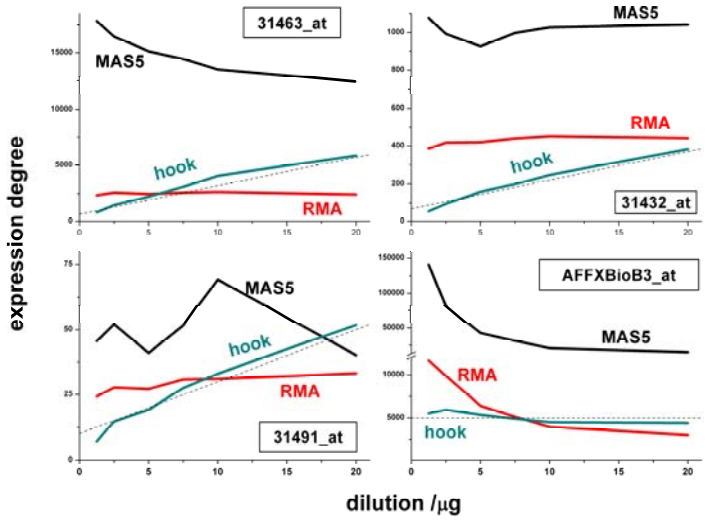



**Figure 4.7:** Expression values of selected probes and methods upon dilution: The concentration of the specific transcripts linearly decreases as reflected by the hook estimate. The other methods provide different, mostly constant expression estimates owing to normalization. Note that AFFXBioB3 is a hybridization control which is spiked-into the hybridization solution with constant concentration. Again the hook-method well reproduces this behaviour.

Part b of Fig. 4.6 shows that the background intensity indeed changes almost linearly with dilution. The mean non-specific background (N) is the log-intensity-average over the N-range of the respective hooks. The optical background (O) referring to 2% of the darkest probes was obtained in step 1 of the algorithm. The total background (N+O) was independently obtained by omitting this optical background correction in the hook-algorithm. The relation between the background levels indicates that the optical contribution gradually decreases with increasing transcript concentrations. Moreover, the residual slope of the O-data shows that the "optical" background correction probably comprises also small contributions from non-specific hybridization.
Simple dilution doesn't change the component-composition of the hybridization solution. Consequently the amount of absent probe-sets is expected to remain invariant in the different dilution steps. The respective fractions of absent probes obtained from the hook curves confirm this expectation (part c of Fig. 4.6). Contrarily, MAS5 provides an increasing amount of absent probes at smaller transcript concentrations, probably because the underlying algorithm converts probes with smaller intensities progressively into absent ones. The hook-method uses the N-region as classificatory criterion for absent probes. Obviously it is more robust against dilution effects than the probe-intensity criterion used by MAS5 (32).
Fig. 4.7 illustrates the effect of dilution on the expression levels of selected probe-sets. The expression data obtained from the hook-algorithm correctly reflect the linear decrease of transcript concentration upon dilution in contrast to the MAS5- and RMA-expression levels, which remain virtually constant. The latter effect is the result of the used normalization algorithms which for MAS5 (global mean normalization) and RMA (quantile) balance the probe-level data relatively to a mean characteristics over all dilution steps. This relative scale remains virtually invariant in this type of experiment. In contrast, the hook-method uses an absolute scale which sensitively responds to dilution effects. A set of special probes, the so-called hybridization controls, are spiked into the hybridizations with equal concentrations. The global normalizations pretend variable expression degrees for these probes over the dilution series (e.g. AFFXBioB3_at, see Fig. 4.7) whereas the hook-expression values remain virtually constant as expected.
Note that also another effect is revealed in the expression data shown in Fig. 4.7: The mean expression levels of the selected transcripts differ by more than three orders of magnitude. These absolute changes are accompanied by distinct variations between the expression levels provided by the different methods. For example, one gets RMA>hook at intermediate expressions (31432_at in Fig. 4.7) but partly hook>RMA at high (31463_at) and low (31491_at) levels. These trends can be attributed to the better linearity of the hook-method over the whole concentration range which reduces systematic biases due to background and saturation effects compared, e.g., with RMA (see also Fig. 4.5).

### 4.5. Download

The beta-version of the hook-program can be downloaded from www.izbi.de. The stand-alone JAVA program processes single-chips and chip-series in a batch-mode according to the scheme given in Fig. 4.1. Chip and probe-set related characteristics such as expression degrees, hook-curves and sensitivity profiles are exported in tabular form and jpg-graphics.

## 5. Conclusions

The improvement of microarray calibration methods is an essential prerequisite for obtaining absolute expression estimates which in turn are required for the quantitative analysis of transcriptional regulation. Benchmark studies indicate that the correction for non-specific background intensity contributions is the crucial preprocessing step. Here mismatched MM probes provide essential information not available from PMonly approaches. Among established linear calibration approaches gcRMA emerges as the method which makes the best compromise between accuracy and precision across the whole intensity range. The Langmuir-hybridization model provides a physically adequate and computationally feasible approach for microarray intensity calibration with the potency to improve existing linear methods. Our hook-calibration method uses this model together with the positional-dependent nearest-neighbour affinity correction. Although related to single-chip analysis, hook performs roughly as well as the multi-chip method gcRMA method in estimating expression values. The hook method in addition provides a set of chip summary characteristics which evaluate the

performance of a given hybridization in terms simple parameters such as the mean non-specific background intensity, its saturation value, the mean PM/MM-sensitivity gain and the fraction of absent probes.

***Acknowledgements:*** We thank Anke Wendschlag for performing some of the data calculations. The work was supported by the Deutsche Forschungsgemeinschaft under grant no. BIZ 6/4. H. Berger was supported by the Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe (grant no. 70-3173-Tr3) to which we are grateful for using the MMML-gene expression data.

## 6. References

1. Affymetrix spiked-in data set: http://www.affymetrix.com/support/technical/sample_data/datasets.affx.
2. GeneLogic dilution data: http://www.GeneLogic.dilution.com/.
3. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. (2005), Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms, *Nucl. Acids Res.* **33,** 5914-23.
4. Eklund, A. C., Turner, L. R., Chen, P., Jensen, R. V., deFeo, G., Kopf-Sill, A. R., and Szallasi, Z. (2006), Replacing cRNA targets with cDNA reduces microarray cross-hybridization, *Nature Biotechnology* **24,** 1071-73.
5. Deng, V., et al. (2007), FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice, *Hum. Mol. Genet.* **16,** 640-50.
6. Binder, H. (2006), Thermodynamics of competitive surface adsorption on DNA microarrays - theoretical aspects, *Journal of Physics Condensed Matter* **18,** S491-S523.
7. Hummel, M., et al. (2006), A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling, *N Engl J Med* **354,** 2419-30.
8. Juhasz, A., Markel, S., Gaur, S., Wu, X., and Doroshow, J. (2007), Inhibition of NOX1 Gene Expression with shRNA in Human Colon Cancer, *Gene Expression Omnibus* **GSE4561**.
9. Malek, S. N., and Ouilette, P. N. (2007), Chronic lymphocytic leukemia (CLL) gene expression comparison, *Gene Expression Omnibus* **GSE 9250**.
10. Furge, K. A., Chen, J., Koeman, J., Swiatek, P., Dykema, K., Lucin, K., Kahnoski, R., Yang, X. J., and Teh, B. T. (2007), Detection of DNA Copy Number Changes and Oncogenic Signaling Abnormalities from Gene Expression Data Reveals MYC Activation in High-Grade Papillary Renal Cell Carcinoma, *Cancer Res* **67,** 3171-76.
11. zur Nieden, N. I., Price, F. D., Davis, L. A., Everitt, R. E., and Rancourt, D. E. (2007), Gene Profiling on Mixed Embryonic Stem Cell Populations Reveals a Biphasic Role for {beta}-Catenin in Osteogenic Differentiation, *Mol Endocrinol* **21,** 674-85.
12. Stepanova, A. N., Yun, J., Likhacheva, A. V., and Alonso, J. M. (2007), Multilevel Interactions between Ethylene and Auxin in Arabidopsis Roots, *Plant Cell* **19,** 2169-85.
13. Li, C. M., and Klevecz, R. R. (2006), From the Cover: A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change, *Proceedings of the National Academy of Sciences* **103,** 16254-59.
14. Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A. K., and Khurana, J. P. (2007), F-Box Proteins in Rice. Genome-Wide Analysis, Classification, Temporal and Spatial Gene Expression during Panicle and Seed Development, and Regulation by Light and Abiotic Stress, *Plant Physiol.* **143,** 1467-83.
15. Hekstra, D., Taussig, A. R., Magnasco, M., and Naef, F. (2003), Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays, *Nucl. Acids. Res.* **31,** 1962-68.
16. Burden, C. J., Pittelkow, Y. E., and Wilson, S. R. (2004), Statistical Analysis of Adsorption Models for Oligonucleotide Microarrays, *Statistical Applications in Genetics and Molecular Biology* **3,** 35.

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

18. Binder, H., and Preibisch, S. (2006), GeneChip microarrays - signal intensities, RNA concentrations and probe sequences, *Journal of Physics Condensed Matter* **18,** S537-S66.

19. Burden, C. J., Pittelkow, Y. E., and Wilson, S. R. (2006), Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays, *Journal of Physics Condensed Matter* **18,** 5545-65.

20. Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A., and Vingron, M. (2002), Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics* **1,** 1-9.

21. Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002), A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics* **18,** 105-10.

22. Wu, Z., and Irizarry, R. A. (2005), A Statistical Framework for the Analysis of Microarray Probe-Level Data, *John Hopkins University, Dept. of Biostatistics Working Paper* **73,** 1-31.

23. Binder, H., and Preibisch, S. (2005), Specific and non-specific hybridization of oligonucleotide probes on microarrays, *Biophysical Journal* **89,** 337-52.

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

25. Affymetrix (2001), Affymetrix Microarray Suite 5.0, *in* "User Guide", Affymetrix, Inc., Santa Clara, CA.

26. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003), Summaries of Affymetrix GeneChip probe level data, *Nucl. Acids. Res.* **31,** e15-.

27. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics* **4,** 249-64.

28. Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2003), A Model Based Background Adjustment for Oligonucleotide Expression Arrays, *John Hopkins University, Dept. of Biostatistics Working Paper* **1**.

29. Li, C., and Wong, W. H. (2001), Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proc Natl Acad Sci U S A* **98,** 31-36.

30. Affymetrix (2005), Guide to probe logarithmic intensity error (PLIER) estimation.

31. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias, *Bioinformatics* **19,** (9).

32. Affymetrix (2002), Statistical Algorithms Description Document, Santa Clara.

33. Zhang, L., Miles, M. F., and Aldape, K. D. (2003), A model of molecular interactions on short oligonucleotide microarrays, *Nature Biotechnology* **21,** 818-28.

34. Shedden, K., et al. (2005), Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data, *BMC Bioinformatics* **6,** 26.

35. Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006), A New Summarization Method for Affymetrix Probe Level Data, *Bioinformatics* **22,** 943-49.

36. Havilio, M. (2005), Signal Deconvolution Based Expression-Detection and Background Adjustment for Microarray Data, *Journal of Computational Biology* **13,** 63-80.

37. Choe, S., Boutros, M., Michelson, A., Church, G., and Halfon, M. (2005), Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, *Genome Biology* **6,** R16.

38. Qin, L.-X., Beyer, R., Hudson, F., Linford, N., Morris, D., and Kerr, K. (2006), Evaluation of methods for oligonucleotide array data via quantitative real-time PCR, *BMC Bioinformatics* **7,** 23.

39. Ploner, A., Miller, L., Hall, P., Bergh, J., and Pawitan, Y. (2005), Correlation test to assess low-level processing of high-density oligonucleotide microarray data, *BMC Bioinformatics* **6,** 80.

40. Verhaak, R., Staal, F., Valk, P., Lowenberg, B., Reinders, M., and de Ridder, D. (2006), The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies, *BMC Bioinformatics* **7,** 105.

41. Zakharkin, S., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K., Parrish, R., Allison, D., and Page, G. (2005), Sources of variation in Affymetrix microarray experiments, *BMC Bioinformatics* **6,** 214.

42. Freudenberg, J., Boriss, H., and Hasenclever, D. (2004), Comparison of Preprocessing Procedures for Oligo-nucleotide Microarrays by Parametric Bootstrap Simulation of Spiked-in Experiments, *Methods in Inf. Med.* **5,** 434-38.

43. Irizarry, R. A., Wu, Z., and Jaffee, H. A. (2006), Comparison of Affymetrix GeneChip expression measures, *Bioinformatics* **22,** 789-94.

44. Affymetrix (2001), Array Design for the GeneChip Human Genome U133 Set.

45. Affymetrix (2003), GeneChip Human Genome U133 Arrays.

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