# Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping

Rippe, R.C.A.

**Author**: Rippe, Ralph Christian Alexander
**Title**: Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping
**Issue Date**: 2012-11-13

# 2 Correction of Fluorescence Bias on Affymetrix Genotyping Microarrays

Fluorescence signals obtained from microarrays for SNP genotyping show systematic strong variations in the levels for SNPs and arrays as well as genotypes. Linear models that take all three effects into account fit very well. Once the model parameters have been estimated for a set of reference arrays, they can be used to calibrate new arrays in a simple way, thereby improving genotyping and analysis of copy number variations and allelic imbalance.

## 2.1 Introduction

Probably the largest scale application of fluorescence these days is the use of microarrays for gene expression, or for genotyping of single nucleotide polymorphisms (SNPs, pronounced as "snip"). A modern SNP microarray contains millions of spots or small beads, called probes, that are covered with small strings of the four nucleotides A, C, G and T that are the building blocks of DNA. Each string is constructed to be the complement (A to T, C to G, and vice versa) of the specific sections of the (human) genome on which SNPs occur.

In a preliminary step, DNA is fragmented by a specific enzyme. The fragments selectively bind (hybridize) to the complementary probes. The amount that hybridizes is, within certain limits, proportional to the concentration of the DNA segments. By preparation with biotin before hybridization, and by

*This chapter was published as the article:*
Rippe, R.C.A., Eilers, P.H.C. and Meulman, J.J. (2012). Correction of Fluorescence Bias on Affymetrix Genotyping Microarrays, *Journal of Chemometrics*, **26**: 191–196. doi: 10.1002/cem.2436.

attaching a fluorophore after, it becomes possible to quantify concentrations by measuring fluorescence intensities. A high-resolution image is formed by scanning the surface of the array with a laser, and the intensities at the probe spots are quantified.

SNPs generally have two variants, called alleles, and the probes are selective to each of the alleles. If we indicate alleles of one SNP by A and B, there will also be two fluorescence signals for each SNP, which we indicate by $a$ and $b$. The DNA of humans (which we consider here), but also that of many other organisms, is contained in two chromosomes. There are three possible combinations of alleles, AA, AB and BB. It is not possible to discern BA from AB, so there is no fourth combination. These combinations are called genotypes.

SNP arrays have two main applications: 1) genotyping of normal DNA, and 2) detection of aberrations, so-called copy number variations (CNV), in tumor DNA. In the first case the result is either AA, AB or BB. Copy number variations allow, in principle, a combination of any number (from zero to five or more) of As and Bs. Usually, if these aberrations occur, they occur in many adjacent positions on the chromosome: whole regions show aberrations in copy number. Franke et al. (2008) describe CNV and its origin in more detail.

We expect the signal $a$ to be proportional to the concentration of the A allele, so only two levels, say $a = a'$ (genotype AB) and $a = 2a'$ (genotype AA) should occur (and a very small background signal in case of genotype BB). For the $b$ signal we similarly expect $b = b'$ (genotype AB) and $b = 2b'$ (genotype BB). Under ideal circumstances, $a'$ and $b'$ should be the same for all SNPs. While working with the fluorescence signals of several types of SNP arrays, we discovered that this is not the case; a strong SNP-dependent bias exists. However, the size of this bias, which is characteristic for each SNP, can be estimated reliably, with a linear regression model, applied to a training set of microarrays. Once the parameters of the model have been estimated, they can be used to correct the bias in new arrays, a procedure we call calibration.

We present two models, one which leads to parameters that can be used to calibrate a new array without knowledge of the SNP genotypes of a new biological sample. An extended model uses this information and allows for somewhat better calibration. However, it can only be used when genotypes are available, which limits its usefulness to special situations, for instance as a building block that iteratively combines calibration with genotype estimation.

The main purpose of this paper is to introduce the model, to show the effect of the calibration procedure, and to illustrate its potential for more precise copy number estimation. Because the regression model is huge (a million parameters or more, derived from approximately 50 million data points), we pay extra attention to efficient calculation.

The organization of the paper is as follows. In the next section we introduce our models, estimation algorithms, model fit and the resulting calibration. We close the paper with a Discussion.

## 2.2 Methods

### Data

We use Affymetrix microarrays. The source of our data is the HapMap (www.hapmap.org) archive (The International HapMap Consortium, 2003, 2007). It provides three types of Affymetrix data files, which are mainly distinguished by the number of SNPs they measure. The oldest platform is the 100k chip, which measures 50.000 SNPs in two different sub-chips: one using the Hind enzyme to cut the DNA into fragments, the other using the Xba enzyme. A newer generation is the 500k chip, which measures 2 x 250.000 SNPs using the NSP or STY enzyme respectively. The final and most recent chip is called SNP6.0, which measures 1.000.000 SNPs in one sample.

We only describe results for the SNP6.0 array, because this is the most recent one. We also analyzed other types of arrays (100k Hind and Xba, 500k NSP and STY) and the results were essentially the same.

### Procedures

The image that is obtained by laser scanning is summarized by averaging the fluorescence intensity over all pixels that belong to one probe. The numbers are collected in a so-called CEL file. Although, to simplify the presentation, we described the technology as if there is one probe per SNP allele, in reality there are four on SNP 6.0 arrays. We simply average the intensities of the four probes to get one number for each of the two alleles of each SNP. This gives us two vectors, each of length $p$.

## 2.3 Models and estimation

In this section we describe the data in more detail. We first explain prior transformation of the raw data. Then we develop two models. One we call “global" because it summarizes the effect of the genotype by just three parameters (per allele) for all SNPs. The second model is called “local" because it has three parameters (per allele) for each SNP. We have to fit the models to very large data matrices (a million SNPs and 90 arrays). We present an efficient semi-symbolic algorithm.

### Two linear models

We denote the number of SNPs by $p$ and the number of arrays (each based on one biological sample) by $n$. The raw fluorescence measurements are contained in two $p \times n$ matrices $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$, one for each allele. A careful study of images of these matrices shows three things:

- Some rows are systematically brighter than others, so each SNP appears to have its own level of brightness.
- Some columns are brighter than others; this is related to the quality of the DNA and its handling in the laboratory. Thus, each array has its own level of brightness.
- Brightness is modulated by the number of alleles (0, 1 or 2).

As a first approximation, it is reasonable to assume multiplicative effects of SNP level, array level and number of alleles. Hence a linear model for the logarithms of the fluorescence intensity is expected to work well.

Let $t_{ij} = \log(a_{ij})$, where the logarithms are to base 10. Let the genotypes be coded in the 3-way indicator matrix $\boldsymbol{H} = [h_{ijk}]$, where $k \in \{1, 2, 3\}$ codes for the genotype; $h_{ijk} = 1$ if SNP $i$ on array $j$ has genotype $k$, otherwise $h_{ijk} = 0$. The model is written as

$$t_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{3} \gamma_k h_{ijk} + e_{ij}, \tag{2.1}$$

where $\mu$ is the grand mean, $\alpha_i$ the effect of SNP $i$, $\beta_j$ the effect of array $j$, and $\gamma_k$ the effect of genotype $k$. For identifiability, we introduce the constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. The error $e = [e_{ij}]$ is assumed to have constant variance. This is a simplifying assumption, but it cannot do much harm. We are only interested in point estimates of the model parameters, not in their standard errors.

We call the model in (2.1) the global model, since it has one set of genotype parameters ($\gamma$) for all SNPs. A refinement is to have separate genotype parameters for each SNP: $\boldsymbol{\Gamma} = [\gamma_{ik}]$. We call this the local model, which is specified as

$$t_{ij} = \mu + \beta_j + \sum_{k=1}^{3} \gamma_{ik} h_{ijk} + e_{ij}, \tag{2.2}$$

where we again require that $\sum_j \beta_j = 0$.

Identical models are used for the B allele, with $t_{ij} = \log(b_{ij})$. As said, we are interested in standard errors per se. Nevertheless, it is good to have a rough estimate. If the estimated $\alpha$ is unreliable, using it for bias correction might introduce additional variance, with detrimental effects. Let us assume that we use 90 arrays, obtained from the HapMap site (www.hapmap.org) to calibrate the model. With many thousands of SNPs, the degrees of freedom consumed by estimating $\beta$ and $\gamma$ are negligible, so $\hat{\alpha}$ for an individual SNP is roughly determined by averaging over 90 arrays. Its variance will thus be approximately 1/90th of the variance of the noise on an individual array. Hence we conclude that we do not have to worry about introducing extra variance.

### Parameter estimation

The arrays we are analyzing here cover up to a million SNPs each. To get parameter estimates, we apply the model to an available set of 90 arrays. Hence we have millions of data points and a huge number of parameters: approximately one million for the global model and triple that number for the local model. Our models can be written as regression models, but explicit construction of the design matrix and invoking a regression procedure is not a good idea: the design matrix would have many billions of elements. However, it is very sparse, so a better solution would be to use sparse matrix software. We have not tried this approach, so we cannot report on its effectiveness. Instead, we have explored block relaxation and symbolic solutions of the regression equations.

In both models (2.1) and (2.2) it is easy to compute one set of parameters if the rest is available. One simply has to average residuals, over SNPs, arrays or genotypes, dependent on the type of parameters. Departing from reasonable starting values (averages over SNPs for $\boldsymbol{\alpha} = [\alpha_i]_{i=1}^{p}$, averages over arrays for $\boldsymbol{\beta} = [\beta_j]_{j=1}^{n}$), one iteratively updates each set of parameters. In the numerical analysis literature this is known as block relaxation.

Alternatively one can build and solve the normal equations symbolically. We illustrate this for the local model (2.2). With appropriate $\boldsymbol{C}$ and $\boldsymbol{D}$, we can write

$$\boldsymbol{t} = \boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\gamma} + e \tag{2.3}$$

where $\boldsymbol{\beta}$ contains the $n$ $\beta_j$ parameters in (2.2) and $\boldsymbol{\gamma}$ = vec($\boldsymbol{\Gamma}$), i.e. the columns of $\boldsymbol{\Gamma} = [\gamma_{ik}]$ stacked below each other, and $\boldsymbol{t} = \text{vec}(\boldsymbol{T})$. The structure of $\boldsymbol{C}$ is simple, it can be written as $\boldsymbol{C} = \boldsymbol{I}_n \otimes \mathbf{1}_p$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\mathbf{1}_p$ is a vector of ones, of length $p$. The structure of $\boldsymbol{D}$ is more complex; it consists of $n$ blocks of diagonal matrices. Each block has three diagonal matrices $\boldsymbol{D}_{jk}$, one for each layer of $\boldsymbol{H}$, and each matrix $\boldsymbol{D}_{jk}$ contains the elements of the $j$th vector in the $k$th layer of the 3-way matrix $\boldsymbol{H}$ on its diagonal. Thus, $\boldsymbol{D}$ has dimensions $(n \times p) \times 3p$.

We don't form $\boldsymbol{C}$ and $\boldsymbol{D}$ explicitly. Instead we study the normal equations

$$\begin{bmatrix} \boldsymbol{C}'\boldsymbol{C} & \boldsymbol{C}'\boldsymbol{D} \\ \boldsymbol{D}'\boldsymbol{C} & \boldsymbol{D}'\boldsymbol{D} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}'\boldsymbol{t} \\ \boldsymbol{D}'\boldsymbol{t} \end{bmatrix}, \tag{2.4}$$

or

$$\begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \end{bmatrix}, \tag{2.5}$$

where $\boldsymbol{V}_{11} = \boldsymbol{C}'\boldsymbol{C}$, $\boldsymbol{V}_{12} = \boldsymbol{C}'\boldsymbol{D}$, $\boldsymbol{V}_{21} = \boldsymbol{D}'\boldsymbol{C}$, $\boldsymbol{V}_{22} = \boldsymbol{D}'\boldsymbol{D}$, $\boldsymbol{f}_1 = \boldsymbol{C}'\boldsymbol{t}$ and $\boldsymbol{f}_2 = \boldsymbol{C}'\boldsymbol{t}$. One can prove that $\boldsymbol{C}'\boldsymbol{C} = p\boldsymbol{I}_n$, $\boldsymbol{D}' = \tilde{\boldsymbol{H}}$ and $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{F}$, where $\tilde{\boldsymbol{H}}$ is a matrix formed by placing the three layers of $\boldsymbol{H}$ below each other. $\boldsymbol{F}$ is a $3p$ by $3p$ diagonal matrix; its first (second, third) $p$ diagonal elements gives, for each SNP, the number of times genotype 1 (2, 3) occurs. Furthermore, $\boldsymbol{C}'\boldsymbol{t}$ contains the sums of the columns of $\boldsymbol{T}$, while $\boldsymbol{D}'\boldsymbol{t}$ is a stack of three vectors; the first (second, third) vectors contain the sum, per SNP of the elements of $\boldsymbol{t}$ corresponding to genotype 1 (2, 3).

From (2.5) follows:

$$\hat{\gamma} = \boldsymbol{V}_{22}^{-1}(\boldsymbol{d}_2 - \boldsymbol{V}_{21}\hat{\boldsymbol{\beta}}) \tag{2.6}$$

and hence

$$(\boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21})\hat{\boldsymbol{\beta}} = \boldsymbol{d}_1 - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{d}_2. \tag{2.7}$$

Because $\boldsymbol{V}_{22}$ is a diagonal matrix, multiplication by $\boldsymbol{V}_{22}^{-1}$ boils down to dividing the elements of a vector or the rows of a matrix by the corresponding diagonal elements of $\boldsymbol{V}_{22}$. Hence, it is not hard to compute $\boldsymbol{V}_{11} - \boldsymbol{V}_{11}\boldsymbol{V}_{12}^{-1}\boldsymbol{V}_{21}$ and to solve for $\hat{\boldsymbol{\beta}}$, a vector of moderate length. Additional efficiency can be realized by exploiting the way $\boldsymbol{V}_{21}$ is formed. Details on the latter suggestion are considered outside the scope of the current paper.

In this analysis we have ignored the fact that the system in (2.5) is singular, because the condition $\sum_j \beta_j = 0$ is not applied. One way to handle this restriction is to introduce a Lagrange multiplier, $\lambda$ and extend the objective function of the model (the sum of squares of differences between observed and fitted values) with $\lambda \sum_j \beta_j$. The system of equations (2.8) becomes

$$\begin{bmatrix} \boldsymbol{C}'\boldsymbol{C} & \boldsymbol{C}'\boldsymbol{D} & \boldsymbol{e} \\ \boldsymbol{D}'\boldsymbol{C} & \boldsymbol{D}'\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{e}' & \boldsymbol{0} & 0 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}'\boldsymbol{t} \\ \boldsymbol{D}'\boldsymbol{t} \\ 0 \end{bmatrix}, \tag{2.8}$$

where $\boldsymbol{e}$ is a vector of ones and $\mathbf{0}$ a matrix of zeros.

An somewhat easier solution is to demand the minimum-norm solution for $\boldsymbol{\beta}$, by replacing $\boldsymbol{C}'\boldsymbol{C}$ in (2.5) by $\boldsymbol{C}'\boldsymbol{C} + \kappa \boldsymbol{I}$ with $\kappa$ a small number. This is the approach we have chosen, using $\kappa = 10^{-6}$.

## 2.4 Results

In this section we describe model fit and effect of possible model-based calibration using parameters from the two models described above. All computations were done with R scripts (R Development Core Team, 2012).

### Model fit

In our experience the speed of convergence is quite good: from 10 to 30 iterations generally suffice to find changes in the updates in the order of $10^{-6}$ (relative size). The constraints on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are applied in each iteration.

Running the implementation of the aforementioned symbolic model on an Intel Core2 Duo 1.4 GHz processor takes about 50 seconds for 90 Affymetrix 100k Hind arrays ($10^5$ SNPs). A larger dataset (63 arrays with $10^6$ SNPs from SNP6) takes 220 seconds.

A typical model fit is shown in Figure 2.1. The standard deviations of the residuals are approximately 0.063 for the global model and 0.051 for the local model. These results are for a set of Affymetrix SNP6.0 arrays. Similar results were obtained for the Affymetrix 100k and 500k arrays. To reduce visual clutter by too many data points in the scatterplots, results for only one chromosome are shown.

The random variation around the fitted line is larger for the BB genotype than for others. The graphs show the fit to the logarithm of the fluorescence signal for the A allele, which is small if the the genotype is BB, as can be seen from the positions of the centers of the clouds of observations. It is well known that constant addtive noise errors appear as increasing relative to low signal values. From the graphs we can deduct that the assumption of

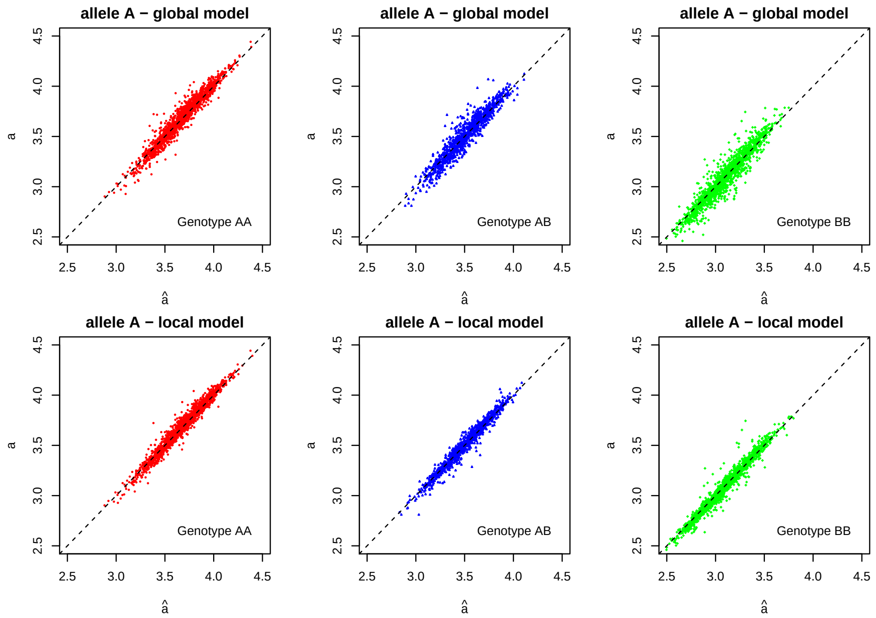

**Figure 2.1:** Results for a selected SNP6.0 array on chromosome 1. Model fit for one allele (A). Top panels show the fitted versus original signals in the global model. Bottom panels show the results for the local model. For allele B, the results are similar.

constant variance of the errors is a simplification, but not an extreme one. We will return to more advanced error models in the Discussion.

## Model-based calibration

From the model we obtain, for each color, either a vector $\hat{\boldsymbol{\alpha}}$ (global model) or a matrix $\hat{\boldsymbol{\Gamma}}$ (local model). We can use them to calibrate the signals of new arrays. Assume that we add one or more columns to our data matrix, representing new SNP arrays, which have not been used for model fitting. Let $l$ indicate one of these columns. For the global model, we compute $t^*_{il} = t_{il} - \hat{\alpha}_i$, to correct for the SNP effects. If we wish to correct for the array effect, we can compute $\hat{\beta}_l$ such that $\sum_i (t^*_{il} - \hat{\beta}_l - \hat{\mu}) = 0$. In our applications we do not need

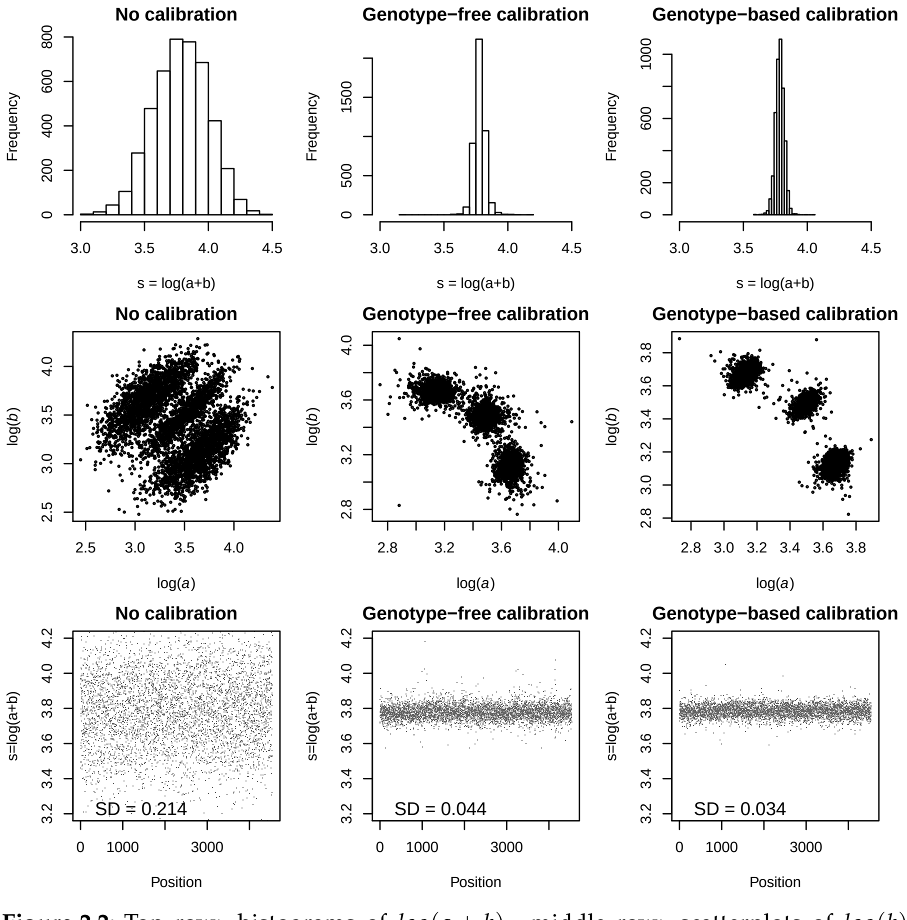

**Figure 2.2:** Top row: histograms of $log(a+b)$. middle row: scatterplots of $log(b)$ against $log(a)$. Bottom row: $log(a+b)$ against chromosomal positions. Standard deviations for raw signal (left), signal after genotype-free calibration (middle) and genotype-based calibration (right). Genotype-free calibration reduces noise considerably, genotype-based calibration provides a small further improvement.

this calibration, because we study single arrays, but this might not be true in other applications.

When we do not use the genotypes of the new array, we call this *genotype-free* calibration. If the genotypes are available we can use the results of the local model, by computing $t^*_{il} = t_{il} - \sum_k h_{ilk}\gamma_{ik}$. We call this *genotype-based* calibration. To calibrate for the array effect, one computes $\hat{\beta}_l$ such that $\sum_i(t^*_{il} - \hat{\beta}_l - \sum_k h_{ilk}\hat{\gamma}_{ik}\hat{\mu}) = 0$.

In Figure 2.2 we show how calibration improves the bandwidth (standard deviation) of the SNP signal. Histograms of $\log(a + b)$ show reduced standard deviations, the middle scatterplots show more condensed (genotype) clusters, and the bottom scatterplots now show $\log(a + b)$ against the position on the chromosome. The improvement from uncalibrated to genotype-free calibrated signal is major, while genotype-based calibration provides a smaller additional improvement.

Genotype-free calibration is less precise, but it can be used for new samples, for which genotypes generally are not available. We propose that the model parameters are estimated once for a set of high-quality DNA samples. The parameters so obtained can be used to calibrate all future arrays.

Genotype-based calibration is less generally useful, but we can envisage a multi-step procedure in the context of genotyping. The first step for any new array is to perform genotype-free calibration. The next step is to determine genotypes using the calibrated signals. Given these genotypes, genotype-based calibration can be performed, followed by a second round of genotyping. Of course, there is the danger of a self-fulfilling prophecy, so only careful testing can show the performance of this recipe. We consider the latter outside the scope of the current paper.

In the next section we show how genotype-free calibration might improve detection of CNV and allelic imbalance along chromosomes.

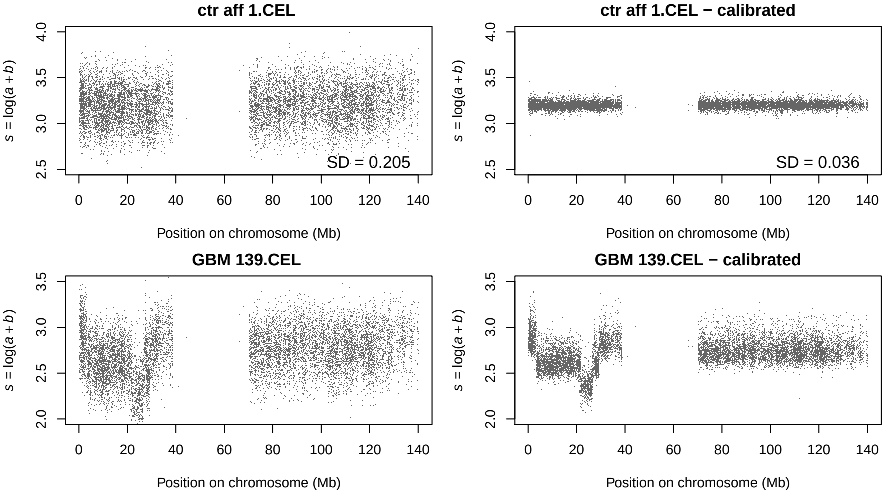

**Figure 2.3:** CNV plot for chromosome 9, based on an Affymetrix 500k NSP array. Top panels: normal tissue; bottom panels: brain tumor. The dots show the (calibrated) signal $s$.

## 2.5 Application: CNV and imbalance maps

An application of SNP signals in tumors is to graph signal levels against their position on the chromosome. Most interesting are copy numbers (the sum of the $a$ and $b$ signals) and allelic imbalance(their ratio). We show possible improvements using genotype-free calibration with our model.

Figure 2.3 shows CNV data for chromosome 9, for normal tissue and for a brain tumor, obtained from the Rotterdam Erasmus Medical Center (Bralten et al., 2010). Figure 2.4 shows allelic imbalance, $\log(b/a)$, before and after calibration. The improvement is evident.

The correction is more effective when the overall signal is strong, because then the systematic SNP effects are strong. In low-quality arrays noise is more dominant, and we cannot correct that with calibration. This can be seen in the figures 2.3 and 2.4. The sample from Affymetrix is of better quality than the tumor sample.

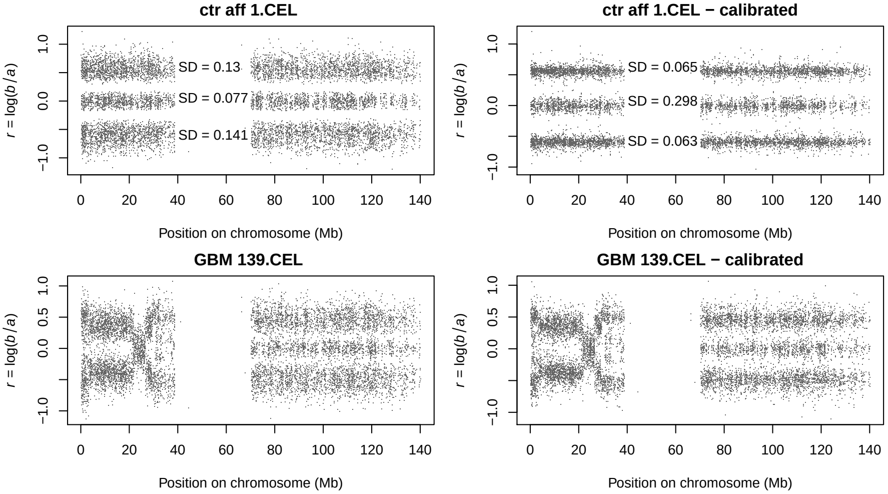

**Figure 2.4:** Data on chromosome 9. Illustration of improvement in signals for allelic imbalance after calibration (right panels). In the top panels, we see three bands (one for each genotype) in a control sample, whereas in the bottom panels we see just two bands in the left side signal (P-arm). Deviations to the three-band signal indicate problematic DNA.

## 2.6 Conclusions and Discussion

We have described two models for systematic effects of SNPs, arrays and genotypes in fluorescence signals on microarrays. The first model contains overall genotype effects and the second contains SNP-specific genotype effects. The parameter estimates following from these models were used to calibrate the raw fluorescence signals. Calibration removes apparent noise in signal maps.

The calibration we propose is simple, fast, and effective. Once parameters for the global model have been estimated, based on a set of high-quality reference arrays, calibration entails only the correction of probe summaries by a single number, per allele of each SNP. This has to be done only once, whether one in interested in genotypes, copy number variations, or both. Each array can be calibrated in isolation, in less than a second.

The idea to model systematic effects and using these to calibrate signals, for improved downstream processing, is not new, see RLMM (Rabbee and Speed, 2006), BRLMM (Affymetrix, 2006) and CRLMM (Carvalho et al., 2007, 2010). These procedures were developed for genotyping and they have in common that they demand a relatively large set of arrays to work reliably and only correct signals for that set. A similar procedure has been developed for CNV (Bengtsson et al., 2008).

We developed our calibration model for, and applied it to, Affymetrix microarrays. It will be interesting to see how it will perform for Illumina arrays. This will be not too hard, because only summary data (per allele, per SNP) are needed.

One of the assumptions of the model is a constant error variance. As Figure 2.1 shows, this is only approximately true: the variance appears to increase for weaker fluorescence signals. This quite common when taking logarithms. A more advanced approach would be to model the relationship between expected value of the model fit and the variance and the error, like in the error model of Rocke and Durbin (2003). Although this would improve the model, we expect little change in the estimated model parameters if estimation is based on a relatively large number of arrays (like the 90 we use here).

Our approach is completely pragmatic: we postulate a model, estimate its parameters, observe a good fit and use the results for calibration. But there should be a fundamental explanation of the very stable systematic patterns we observe. Further research is needed for better understanding.

A software package, named SCALA, is presently available from the corresponding author. It was written for the R system (R Development Core Team, 2012), and has been used for all the analyses mentioned in this paper. We plan to turn it into a Bioconductor package. The software contains a module to convert Affymetrix CEL file data to the aggregated signals that were used in this paper, and a module to estimate the calibration parameters. A module that creates the copy number and allelic imbalance maps including a signal smoother is available. There also is a module for genotyping of single arrays.