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

Rippe, R.C.A.

Discussion

8

This chapter provides a short review of the preceding chapters and restates why single array genotyping should be applied more widely. Furthermore, it addresses some open problems, illustrates new ideas in extension of the current chapters and points out benefits of the proposed methods.

8.1 Advantages of single array analysis

One of the main themes in this thesis was the propagation to switch to single array genotyping, as opposed to single SNP, multi-array genotyping, which is the current common practice. The approach has a number of advantages, which are summarized below.

Single array genotyping is fast and flexible, due to its semi-parametric approach. It is insensitive to differences in sample size, and depends only on user-chosen dimensions of the underlying histogram. The process is very easy to monitor, since it requires tracking only one sample at a time.

Along the same lines, it also allows for better quality control, because the overall level of the signals is an indication of data quality. Because quality control is easy, the procedure is also highly suitable for use in development of small series of chips, for example when devising new layout to research "new" organisms. Furthermore, the procedure is readily available in open source software, in the SCALA software suite.

8.2 A short review

The theme of this thesis can be summarized in a few words: "better data analysis for SNP arrays". The five main chapters present efficient and effective solutions to many problems that are encountered in practice. They are reviewed concisely.

SNP platforms provide a variety of opportunities as well as challenges. Fluorescence signals from these platforms have structural properties: overall fluorescence levels differ not just between arrays, but also between SNPs within one array. The SCALA model, discussed in Chapter 2, contains parameters for estimating the systematic effects of SNPs, arrays and genotypes. This large regression model is applied to both alleles separately, and delivers a million parameters or more. However, due to its extremely sparse structure, a specialized semi-symbolic algorithm allows exact estimation in a very short time. Model fit is highly adequate in terms of (standard deviation) of residuals. Once the parameters of the model have been estimated, they are used to eliminate the systematic effects, thereby greatly enhancing the quality of the fluorescence signals. We call this calibration and apply it in a later chapter.

In Chapter 5 it is shown that the signal calibration is also useful for correction of genomic waves, visible as a systematic pattern when plotting fluorescence signals along chromosomes and smoothed. Calibration removes these waves. Because the model used to obtain calibration parameters does not model spatial autocorrelation, the results of calibration imply that wave patterns in reality are not caused by not spatial autocorrelation. Furthermore, noise in the signals is reduced. When compared to a dedicated wave correction model, NoWaves, performance is equal, but the proposed calibration is more efficient. NoWaves requires reference samples for each array subject to correction, while SCALA applies calibration parameters that were estimated at some prior point in time.

One application of SNP fluorescence signals is to determine SNP genotypes. In Chapter 3 we break with common practice and perform genotyping for all SNPs on individual arrays. A semi-parametric mixture model is

estimated, with three component densities, one for each of the AA, AB and BB genotypes. Comparison to results of SNP by SNP algorithms (CRLMM) as well as a de-facto standard, as found on the HapMap archives, show equal or better performance. Furthermore, where traditional methods do not provide reliable estimates for all scenarios, i.e. low probabilities, for low Minor Allele Frequencies (MAF) due to small or missing components, the estimates from the single array model have higher probabilities and additionally provide genotypes for SNPs that were not called by HapMap. The current model is suitable for different platforms as well as chips with different densities.

Throughout the chapters, genotyping is based on a display of the ratio of the A and B signals versus their sum (on logarithmic scales). Low signals on the sum scale, as well as unclear separation between the three genotype groups on the ratio scale indicate low(er) chip quality. Applying calibration before single array genotyping, as described in Chapter 4, allows us to exploit this knowledge to select only the SNP observations of the highest quality, by a user-defined threshold. This results in higher genotyping probabilities for the selected high-quality observations on low(er) quality arrays.

Another application of the fluorescence signals is the estimation of profiles of copy number changes. These changes generally occur in a segment-wise manner along chromosomes. There is a large literature on smoothing and segmentation of CNV signals, all with the goal to obtain the boundaries of the segments and their levels. A new smoothing algorithm was presented in Chapter 6. The model uses a so-called L_0 penalty on jumps between smoothed values and is therefore referred to as the Zero Exponent Norm, ZEN. The result is an extremely sharp segmentation. A similar segmentation also holds for allelic imbalance signals. However, it is not possible to apply the same smoother to allelic imbalance signals, because several parallel data bands occur. Therefore, we modified an existing scatterplot smoother to use the L_0 penalty in one direction and the L_2 norm in the other, in order to get sharp segmentation here too.

All models and algorithm are written in R, and are combined in a software suite, The SCALA suite (Chapter 7). It provides both command-line functions (for estimation and calibration, as well as genotpying) and a graphical user interface for interactive (simultaneous) smoothing and plotting of CNV and

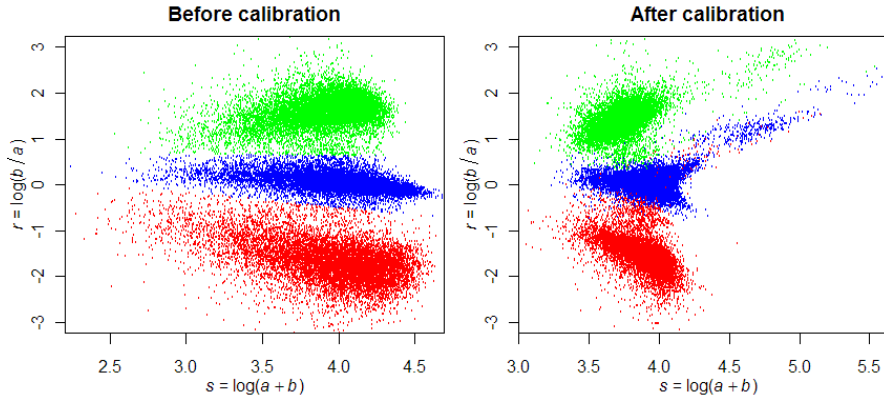

Figure 8.1: An Illumina array with asymmetric signals before and after calibration. Clusters are not condensed, but tails appear.

allelic imbalance.

8.3 Ideas for future research

Although the previous chapters have addressed specific questions and problems, there are still open questions and new directions to be explored. Below, a few are discussed.

Calibration of asymmetric fluorescences signals

In Chapter 3, a method was developed for single array signal calibration. This method was tested extensively for Affymetrix arrays, which have strong symmetric properties when looking at a single array. However, it was also mentioned that e.g. the asymmetric signals from the Illumina *Infinium* platform was not evaluated due to problems with signal calibration. Unfortunately, no explanation was provided as for why the calibration doesn't work, except for the fact that Illumina (but others too) uses two-color fluorescence, where Affymetrix uses just one. It seems that the resulting wavelength differences are at the heart of the asymmetry. In the near future we aim to provide more

insight into the positive and negative aspects of asymmetric signals and propose a solution for the less desirable ones. After calibration, the clusters are not condensed like for Affymetrix, but seem to obtain a swallow-like shape (Figure 8.1). The tails appear after calibration and compromise quality of the genotype calls.

Staaf et al. (2008) used quantile normalization using reference arrays to overcome the asymmetry. However, in practice their approach is not effective in a single ratio-sum transformation since they use a set of arrays to find a symmetric transformation within the given set (Bolstad, Irizarry & Speed, 2003). Still, a part of the solution for asymmetry in fluorescence signals before calibration may be found here.

Extended models

A possible model extension is to perform simultaneously modeling of genotypes, copy number profiles and calibration parameters. An example in which independent estimations for genotypes and CNV have been combined in a single representation is shown for chromosome 9 in Figure 8.2. We refer to the model as the Michelin model, because this representation of the data has similarities to the profile on a (car) tire. However, calling all genotypes at once for such a sample will induce errors. For better clarification, the complete chromosome is split into the tumorous P-arm and healthy Q-arm in Figure 8.3. The top panel shows the healthy tissue with constant CNV and full allelic balance, and has clear genotypes. The three separate views are shown in the left panels in Figure 8.4. The bottom panel however shows the tumor tissue, showing CNV and allelic imbalance. These are shown in the right panels in Figure 8.4. Genotyping this arm at once will be largely incorrect, because one number of genotype clusters is estimated, while a different number of clusters for each segment in this arm would be more appropriate.

The core principle behind this idea is that it is possible to have a different genotype component mixture for each copy number or allelic imbalance segment. Mixtures of 1, 2 or 3 components can occur in different segments. It then is possible to fit a log-concave component mixture per segment, as a fundamental approach to “interactions” between CNV and genotypes. An

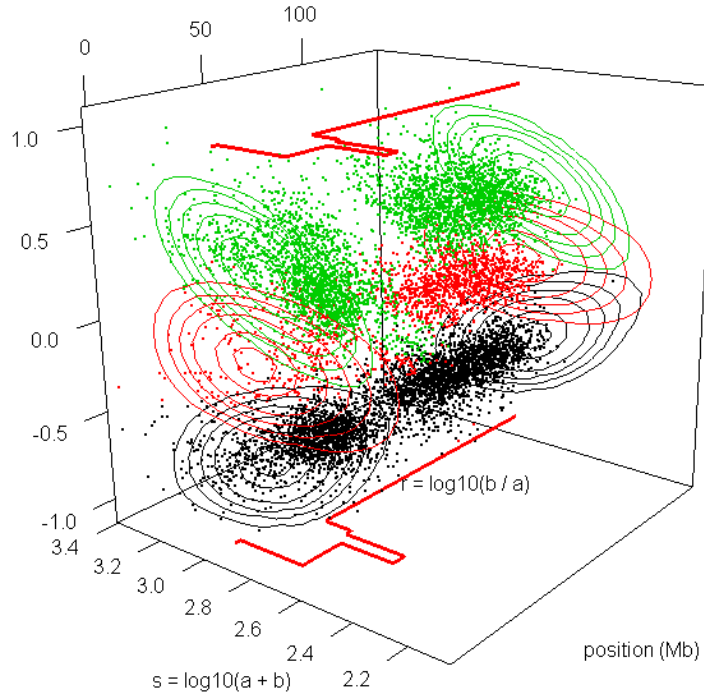

Figure 8.2: Three models combined in one SNP signal representation: 1) CNV profiles (top view), 2) Allelic imbalance (right side view) and 3) Genotyping (front side view).

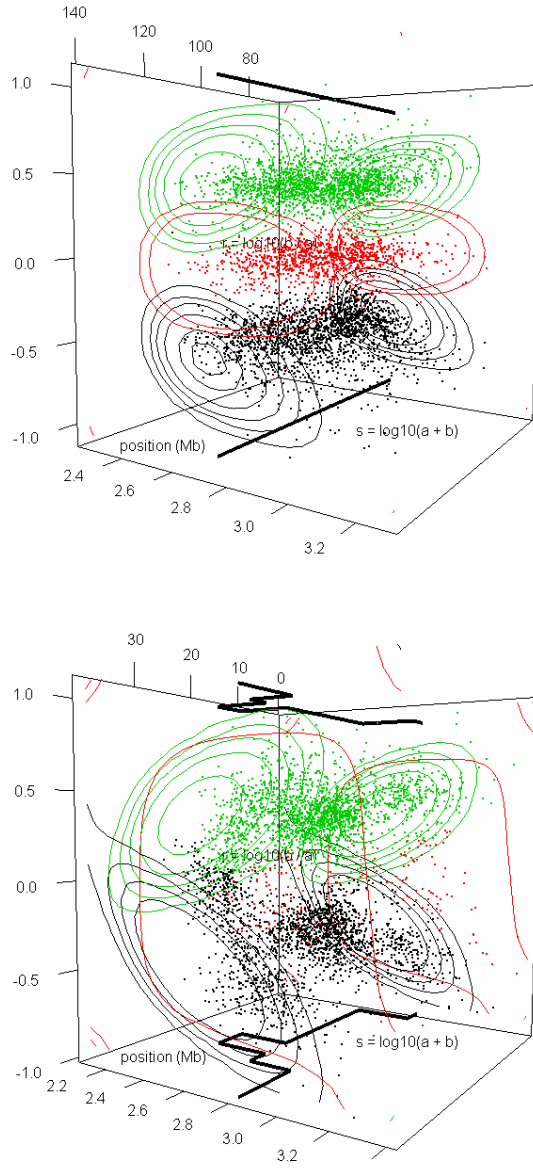

Figure 8.3: Combined SNP model for normal (top) and diseased (bottom) tissue.

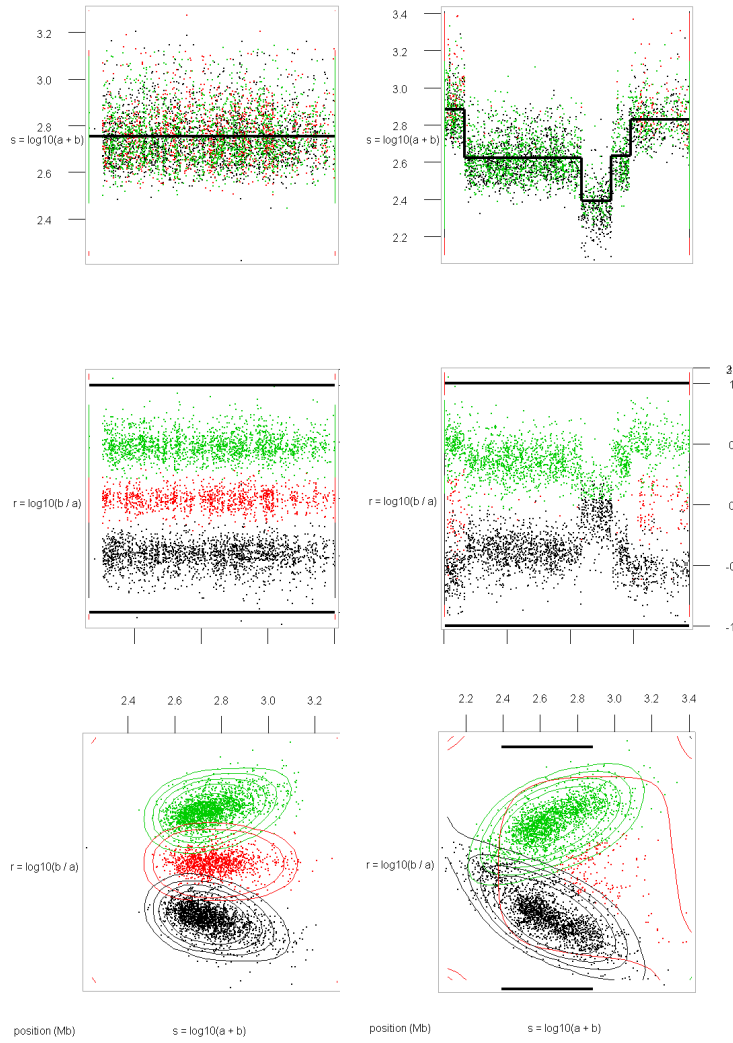

Figure 8.4: Combined SNP model in single view orientations. Left column shows healthy tissue; right column shows tumor tissue. The top panel shows a CNV profile, the middel panel shows allelic imbalance, and the bottom panel shows genotypes.

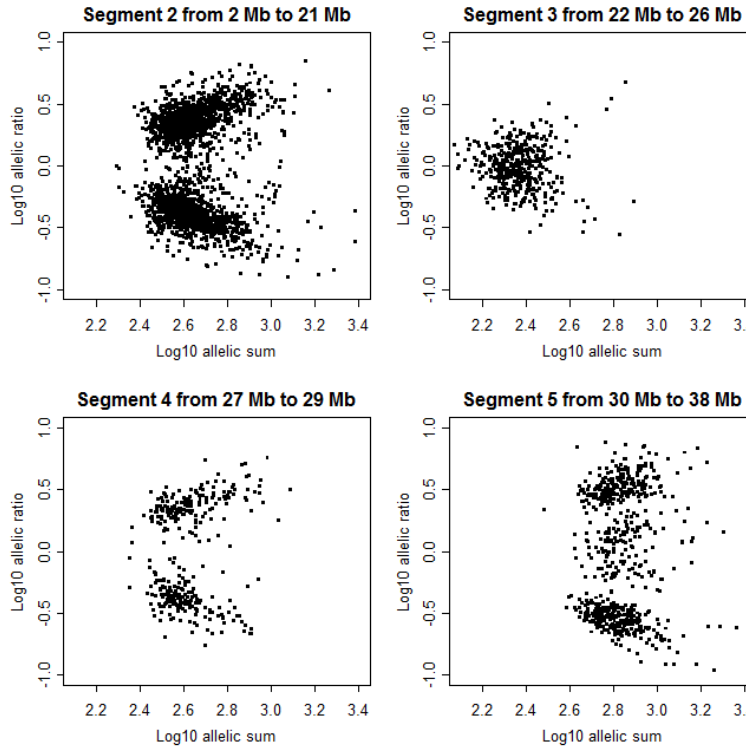

Figure 8.5: Genotype representation of the data within segments. Mixtures of one (top right), two (top left and bottom left) and three (bottom right) components can be distinguished.

illustration is given in Figure 8.5. This approach will provide some mathematical challenges in terms of overfitting or overparametrization.

Figure 8.5 also indicates why it is better to use both the ratio and the sum dimension for genotyping, instead of just the ratio, because the latter would provide genotype densities that are too wide. The bottom right panel provides a clear demonstration. Using the sum dimension in addition allows for more accurate estimations.