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

Rippe, R.C.A.

Optimal use of low quality SNP samples by the single array approach

4

SNP samples from any platform can be of low quality due to many reasons. We propose to select the highest quality part of the sample using a probability-based signal threshold. Genotype call rates, call quality as well as visualization and detection of CNV and allele imbalance can be strongly improved using this method. For example, call rates can be increased from around 60% to around 90% or higher. Reprocessing these low quality arrays may therefore be unnecessary, hence improving research efficiency and reducing sampling costs. Since the method approaches single arrays, it is also possible to threshold (parts of) individual chromosomes.

4.1 Introduction

In cancer research, one of two primary goals is to determine the according genotype from obtained Single Nucleotide Polymorphism (SNP) signals (Mao et al., 2007). To quantify SNPs, the strength of fluorescence in two channels (one for each chromosome in a pair) is measured. The ratio of the signals for the two alleles determines the genotypes. The sum of these signals can be used for CNV estimations.

SNPs are measured through different platforms and with different methodologies. Major companies like Affymetrix provide platforms as well as their own software. The major difference between their technologies is found in the way the fluorescence signals for each allele are produced. Platforms of

This chapter is an adapted version of the article:
Rippe, R.C.A., Eilers, P.H.C. and Meulman, J.J. (2012). Optimal use of low quality SNP arrays by the single array approach, *manuscript*.

varying SNP density are available, ranging from 50.000 (Nicolae et al., 2006) to 250.000 (Dunbar et al., 2008) to 1.000.000 SNPs (Nishida et al., 2008). Of course, other platforms exist (e.g. Illumina), but this paper focuses on the Affymetrix platforms. Throughout the current work we use signals a and b that are averaged over all probes for the A and B allele.

Genotyping

The most procedure to genotype SNPs is to do clustering per SNP, using multiple arrays. From a practical point of view this is very unattractive, because it is almost impossible to monitor the influence of individual arrays on the results. Rippe et al. (2010) developed a single array genotyping algorithm with excellent performance on normal tissue arrays (e.g. as used in epidemiology) and as provided by HapMap. Single arrays are easy to judge in terms of quality, and along the same lines, are easier to have the low quality part of the signal removed.

Commercial software generally calls genotypes for one SNP, using information from multiple arrays. However, taking an alternative perspective by calling genotypes for all SNPs in an individual array might be more fruitful than single SNP - multi-array calling (Tikhomirov, Konkashbaev & Nicolae, 2008; Rippe et al., 2010). One advantage is that there are enough data points available to perform very stable model fitting to obtain the genotype calls (e.g. 4600 SNPs on chromosome 1 on a 100k Hind array). Problems with low minor allele frequencies do not occur. Another major advantage is that besides genotyping, this provides an indication of array quality. For the left sample in Figure 4.1, genotyping is a clear matter, but for the right one it isn't. The SNPs are mostly cluttered in the low signal area, restricting any genotype separation.

In the described single array approach, a SNP genotype generally follows from the ratio of two signals ($r = \log_{10}(b/a)$) with a the signal for allele A and b for allele B. Since the ratio r of the two fluorescence intensities (one for each allele) determines the genotype, signal noise can deteriorate the quality of the genotype calls. It is based on a two-dimensional mixture of log-concave densities (along the lines of Eilers & Borgdorff, 2007), fitted on smoothed 2-

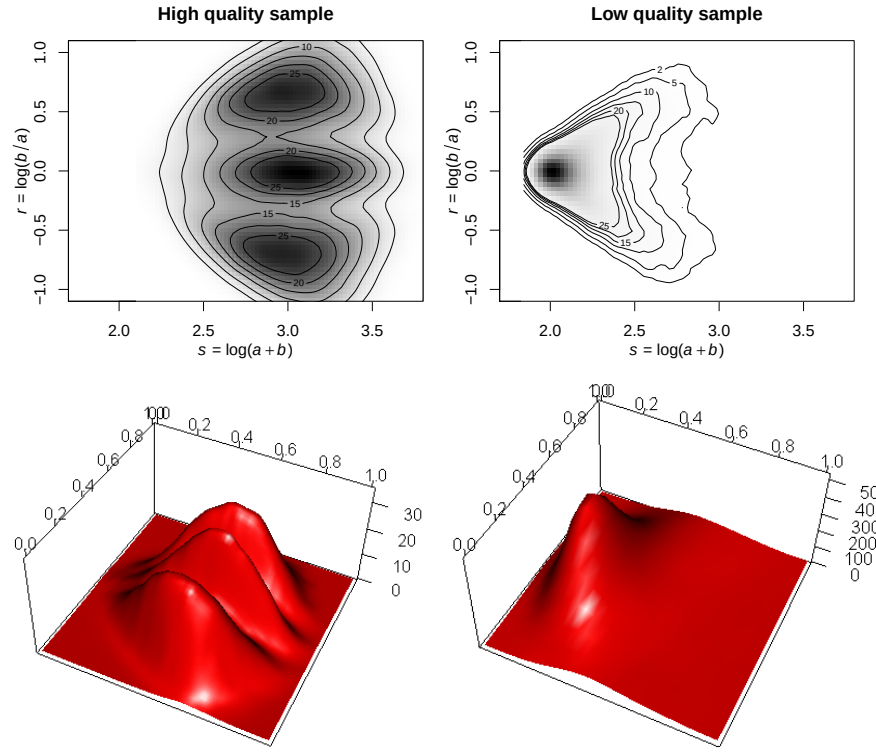

Figure 4.1: Top row: left panel shows a high quality Affymetrix SNP6.0 control sample. Right panel shows a low-quality control sample. On the horizontal axis the sum $s = \log_{10}(a + b)$ is shown; higher values (total signal) indicate higher quality. On the y-axis, the ratio of the signals for the two alleles $r = \log_{10}(b/a)$ is shown; this indicates the relative concentration of fluorescence signals for the two alleles. Genotype cluster contours are superimposed to illustrate cluster separation; better separation results in higher call rates. Bottom row: 3-dimensional replication of the same data. The left panel shows three clear peaks in an otherwise flat landscape, while the right panel shows one peak in a hilly environment: noise.

dimensional histograms (Eilers & Marx, 2007).

To estimate a mixture with three smooth components in two dimensions, we use the familiar EM (expectation-maximization) algorithm. Two steps are

repeated until convergence: 1) split the counts y into three vectors of pseudo-counts, proportional to the current estimate of the mixture components; 2) apply smoothing to the pseudo-counts. Decent starting estimates for the components are needed.

Let $\boldsymbol{Y} = \{y_{ih}\}$ be an $n_1 \times n_2$ matrix of counts in a two-dimensional $n_1 \times n_2$ histogram. The center of bin (i, h) is given by (u_i, v_h). The expected values are modeled by sums of tensor product B-splines. Two bases are computed, $\boldsymbol{B_1}$, with c_1 columns, based on $\boldsymbol{u}$ and $\boldsymbol{B_2}$, with c_2 columns, based on $\boldsymbol{v}$. The bases are combined with a $c_1 \times c_2$ matrix $\boldsymbol{\Theta}$ of coefficients, and the matrix of expected values is computed as

$$\boldsymbol{M} = \exp(\boldsymbol{B_1}\boldsymbol{\Theta}\boldsymbol{B_2}'). \tag{4.1}$$

A penalized Poisson log-likelihood is then optimized. The penalty is complex, because both rows and columns of $\boldsymbol{\Theta}$ are penalized. If $||\boldsymbol{X}||_F$ indicates the Frobenius norm of the matrix $\boldsymbol{X}$, i.e. the sum of the squares of its elements, the penalty is

$$\text{Pen} = \lambda_1 ||\boldsymbol{D_1}\boldsymbol{\Theta}||_F/2 + \lambda_2 ||\boldsymbol{\Theta}\boldsymbol{D_2}'||_F/2, \tag{4.2}$$

where $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ are matrices of the proper dimensions ($c_1 - 3 \times c_1$ and $c_2 - 3 \times c_2$) that form third differences. We use third order differences because for larger λ, the values in Θ move towards a quadratic trend, since in such a trend third differences no longer exists and hence the penalty becomes zero. Given the right circumstances we end up with a mixture of log-concave unimodal shapes.

The final mixture components give three expected values for bin (i, h) of the histogram: μ_{ih1}, μ_{ih2} and μ_{ih3}. From these numbers follow, after division by their sum, three membership probabilities. The largest of the three, which we indicate by $\hat{p}_{ih}$ points to which cluster all the observations in the bin should be assigned.

CNV and Allelic Imbalance

If we plot SNP signals according to their chromosomal position, it is easy to visualize any aberrations in high-quality samples. The s-signal can illustrate

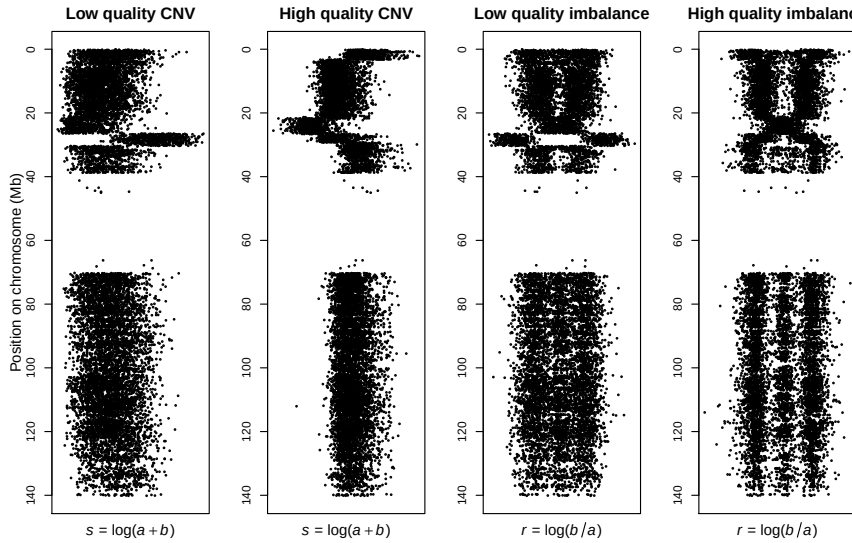

Figure 4.2: High and poor quality examples from Affymetrix 250k NSP tumor arrays. The two left panels show CNV by plotting the *s*-signal from Figure 4.1 against chromosomal position. Right two panels show the *r*-signal against its chromosomal position. Higher quality samples show clearer patterns. Here, higher quality was induced using signal calibration.

copy number variations (CNV) (Sebat et al., 2004; Taylor et al., 2008; McCarrol et al., 2008), while the *r*-signal can indicate allelic imbalance or loss of heterozygosity (LOH) (Beroukhim et al., 2006). Again, having low quality SNP measurements included, besides inducing low call rates (in normal tissue), they can distort any clear pattern in either CNV or LOH. Figure 4.2 illustrates the latter effect. From left to right it shows CNV and allelic imbalancefor a high and low quality array.

Sample quality and signal calibration

With increases in the number of simultaneous SNP probes (currently up to 10^6 SNPs) in recent platforms, the absolute number of errors has also increased. In lower-quality samples this effect is boosted even more. This

problem is illustrated in Figure 4.1 and Figure 4.2, where the left panel shows a high quality Affymetrix SNP6.0 sample compared to a low quality one on the right.

Many researchers that work with SNP samples have come across low quality samples coming in from the lab. Lower sample quality makes it harder to call genotypes and increase the probability (proportion) of wrong calls. In commercial software SNPs below a certain probability threshold would not be called at all (No Call). Individual arrays with too many 'No Calls' may be rejected for further analysis. Unfortunately, processing these samples and their processing still costed money and time.

Rippe et al. (2010, 2012a) described a procedure called SCALA in which single array genotyping and array calibration are implemented. From their linear models follow, for each allele, either a parameter vector $\boldsymbol{\alpha} = [\alpha_i]$ or a three-column array $\boldsymbol{\Gamma} = [\gamma_{ij}]$. See Rippe et al. (2012a) for more information on how the calibration parameters are obtained. We can use them to compute corrected signals and - if wanted - repeat the estimation of the mixture model using these 'new' signals. We call this calibration. In the first case this translates to $x^*_{ij} = x_{ij}/10^{\alpha_i}$. It does not use the genotypes of a new sample, so we call it *genotype-free calibration*. In the second case we have $x^*_{ij} = x_{ij}/10^{\psi}$, where $\boldsymbol{\psi} = \sum_k h_{ijk}\gamma_{ik}$, which we call *genotype-based calibration*. Analogous formulas hold for signal b.

The result of genotype-free calibration for an illustrative array is shown in Figure 4.3. After calibration the cluster are more condensed.

They also discussed a quality control measure similar to the QC criterion used in e.g. Affymetrix Genotyping Console. Our approach is also suitable to save as much information as possible from low quality arrays. In this paper we propose to select only a part of a low quality array and to perform genotyping and calibration methods within the SCALA framework on the remaining data.

Samples can be selected as high or low quality based on their initial call rate. However, since some call rates are either not provided or not obtained from the different data sources, we use the single-array method described above to get an initial call for all samples based on a single method.

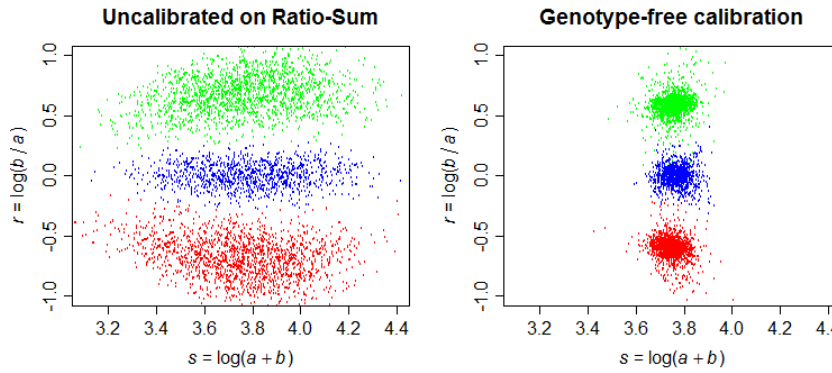

Figure 4.3: Effective genotype-free calibration in a high quality Affymetrix 250k NSP array.

Threshold SNP selection

To tackle the problem of analyzing low overall quality arrays, we propose a signal threshold for non-amplified low quality samples (Mead et al., 2008; Ziegler et al., 2009) to select only those SNPs for which the combination ($s = \log_{10}(a + b)$) of two allelic signals both exceed a certain threshold value. This part of the data shows a much clearer genotype separation. This implies that sample quality is inversely related to the number of SNPs with a low s-value. The suggested method implies that only a part of the probed SNPs are analyzed in depth, but still we obtain more information that nothing at all. Furthermore, in such a sample it is still possible to see larger regions of aberrations on chromosomal regions. Platforms with a higher SNP density benefit more from this approach, since after threshold selection the remaining SNP density is still high enough to perform analyses on.

It is recommended to threshold-select SNPs for one individual array at a time, since observations removed from one array do not match those selected in other array, for a very large part.

We will illustrate our approach using samples obtained from GeoDB, HapMap and the Erasmus Medical Center.

4.2 Methods

Data

In this work we use data from three different platforms. The Affymetrix 100k (Hind and XBA) samples are obtained from GeoDB and include samples of both high and low quality. The Affymetrix 250k NSP (Bralten et al., 2010) and SNP6.0 (Gravendeel et al., 2009) clinical samples of varying quality were provided by the Erasmus Medical Center. Poor quality samples were omitted in these studies, but are included in the current study. The 500k NSP set also contains 8 high quality reference samples. SNP6.0 reference samples were obtained from Affymetrix. 500k STY samples are not evaluated due to the simple fact that no low quality array were avaible.

Genotyping

For the genotyping procedure we chose not to use commercial software, since this software can contain differences in signal processing (Rabbee & Speed, 2006; Xiao et al., 2007) as well as internal SNP and sample comparison. To overcome the problem of platform comparability, we use our own genotyping implementation that is applied on a single array as described in section 4.1.

Call probability as a sample quality measure

Low quality samples can be identified simply by counting the number of SNPs with an s-value below a certain threshold value, even without genotype call rates. However, this requires an objective threshold, which is hard to determine since array base levels may differ for each batch.

Second, the quality of a sample can be measured by the proportion of called genotypes with a large probability. In commercially available software, low-probability calls are rejected and hence result in NC (No Call). We will use an equivalent measure, based on the genotyping results $Q = \max(p_1, p_2, p_3)$ from the SCALA software. This measure is formulated as the proportion of called genotypes with a call probability lower than 0.90 ($Q90$);

the more low-probability calls are obtained, the higher $Q90$ and the lower the sample quality. If a platform collection doesn't contain such samples, 6 samples with the lowest call rate for that platform were selected. High quality arrays were available for all platforms.

Thresholding

To perform threshold signal selection, we use the $\log_{10}$ of the sum of the two allelic signals ($s = log_{10}(a + b)$). This is the x-axis in Figure 4.1. In practice, this means that any SNP with an s-value below a given threshold is completely excluded from further analysis. This can be done for each chromosome separately or for all array observations at once. We choose to perform the computations per chromosome, to have better control.

There are two ways to go from here. First, we can use a straightforward 'hard threshold', which is intuitively very effective. Since the signal quality is specific for a sample, there is no control on the amount of data that is removed: in one sample a threshold value of $s = 3.0$ may exclude 30% of the data, whereas in a very low quality sample the same threshold may remove 100% of the signals.

A different way to apply thresholding is to use the empirical distribution function (edf, which is trivial to determine: for a random sample it is the cumulative distribution function of the values obtained in the sample. It is a staircase function that is equal to 0 for $s < s_1$ and is equal to 1 for $s \geq s_n$ or

$$\hat{F}_n(s) = \frac{\mathrm{z} \leq s}{n} = \frac{1}{n} \sum_{i=1}^{n} I(S_i \leq s) \tag{4.3}$$

with z the number of measured elements in the sample.

Suppose we apply a certain threshold value for the s-signal on a given chromosome. We can restrict the SNP removal to just the one-sided $p = 0.90$ part of the data. If according to the threshold value all SNPs have to be removed, we can force the best $p = 0.10$ below that threshold to always be kept, so that for allele A we have $a^* = a_{s>p}$ where a^* and a are of different lengths (analogous for allele B). This way, there is always some available data

to provide information about genotypes and/or aberrations. We will use the latter approach for further analyses.

Calibration

We will assess the effect of threshold selection on calibration (and downstream call rates).

Furthermore we will look at calibration effects *after* we removed low-signal SNPs and *then* apply precomputed calibration parameters for the remaining SNPs. From these calibrated signals we determine the genotypes. Given these genotypes, we obtain calibration parameters based on the new data. Finally, for illustrational purpose, we apply a genotype-based calibration method to obtain the best possible separation and call rates, using estimated genotypes obtained after genotype-free calibration.

4.3 Results

Call rates for different threshold values

The results for the call rates on uncalibrated chromosome 4 in each of the platforms, for each percentile in $[0, 10, ..., 80, 90]$ are given in Figure 4.4. For high quality SNP6 arrays, it is clear that removal of the lower signals is detrimental to the overall call rate. This pattern holds for all chromosomes (not shown). For the other platforms however, call rates remain more or less constant after removal of the lower signals.

In the left panel (high-quality samples) it can be seen that for all platforms the call rates are high at all threshold levels, but the rates for SNP6 slightly decrease if too much signal is removed. In the right panel (low quality samples) we see lower call rates in all platforms. The SNP6.0 platform seems to benefit the most. This may be due to the fact that in this platform the last 10% still contains 100.000 SNPs. However, both of the Affymetrix 100k (Hind and XBA) call rates do not seem to suffer very much: they are still very high, though lower than for their high-quality counterparts at the same threshold

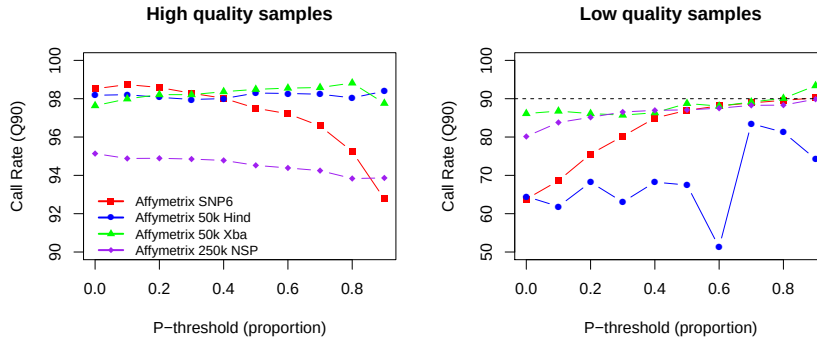

Figure 4.4: Call rates on chromosome 4 using four different platforms. Left: high quality samples. Right: low(er) quality samples. Note that call rates in the left panel ranges from 90 to 100, whereas the right panel ranges from 60 to 100. The dotted line in the left panel indicates the lowerbound of the y-axis in the left panel.

level. Affymetrix 250k NSP and Affymetrix SNP6.0 have very low call rates. All three benefit strongly from threshold selection, which seems to be most effective for the SNP6.0 data. However, the best low-quality SNP6.0 call rate after thresholding (at 90% cut-off) is still below 90%, and this is lower than the call rate for a non-thresholded high-quality sample.

Calibration effects

The effects of thresholding on the same low quality samples as used in section 4.3, but now after genotype-free calibration, are summarized in Figure 4.5. In Table 4.1 these call rates are compared with their uncalibrated counterparts. All platforms show (strong) improvements in call rates using calibration. The Affymetrix improvements are as expected from the calibrated clusters. An example for a 250k NSP sample is shown Figure 4.3.

Furthermore, the 'jump' in call rates at $p = 0.30$ for low quality Affymetrix SNP6 samples indicates after removal of the worst 20% the remaining signal contains much clearer genotype information, which is however still viable for

Table 4.1: Results (median error rates) for low-quality samples after genotype-free calibration. P = (P)roportion of data removed, B = (B)efore calibration, A = (A)fter calibration.

	SNP6		100k Hind		100k Xba		250k NSP	
P	B	A	B	A	B	A	B	A
.0	36.2	19.7	35.5	25.0	13.9	20.7	19.9	5.9
.1	31.4	18.3	38.1	31.3	13.2	17.6	16.2	5.1
.2	24.4	17.0	31.8	28.1	13.9	19.0	14.9	4.7
.3	19.8	4.3	36.9	26.5	14.3	18.1	13.5	4.4
.4	15.1	3.9	31.8	25.4	13.7	15.8	13.0	4.2
.5	13.0	3.7	32.5	24.5	11.2	16.5	12.9	3.9
.6	11.9	3.6	48.8	22.9	11.9	13.9	12.5	3.5
.7	11.1	3.6	16.5	22.9	10.9	12.2	11.7	3.1
.8	10.4	3.6	18.8	17.7	9.9	6.4	11.7	2.8
.9	9.8	3.6	25.8	15.4	6.6	6.1	10.1	1.9

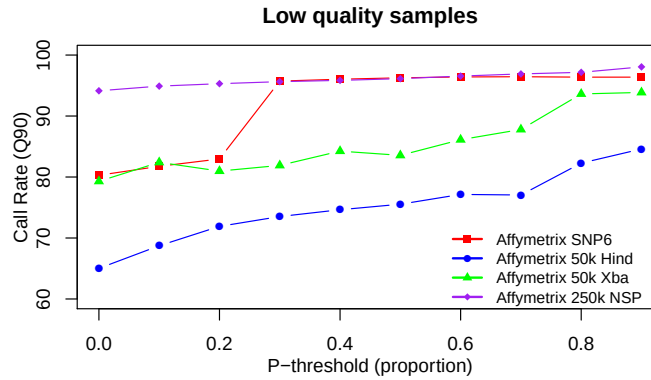

Figure 4.5: Call rates for low quality samples *after* genotype-free calibration. Same samples on chromosome 4.

further improvements. The best call rate is still below the lowest call rate for high quality samples.

Some further investigation shows that the effective combination of calibration and thresholding on the call rates for the 100k Hind, 100k Xba and 250k NSP samples is due to a higher overall quality of the selected low quality samples, compared to the low quality samples from SNP6.0. High quality samples however benefit strongly (not shown), with median call rates consistently approaching 99% for all five platforms.

The effects in Table 4.1 also support the assumption that genotype-based calibration (see section 4.1) is unwise to do: genotype-based calibration requires called genotypes, but these calls can only be made using the original, poor quality signals. This may introduce erroneous calibration due to erroneously called genotypes.

Inspecting aberration profiles

Figure 4.6 shows the effect of threshold selection on signals for profiles of CNV and allelic imbalance. In this particular section we used data on chromosome 9, since we know that aberrations occur there. The left panels show CNV detection, and imbalancepatterns are represented in the right panel. At first sight, in the complete sample the noise in both the left and right panels clearly dominate the data and no patterns can be distinguished at all. However, with an increase in threshold level, we can see deviations patterns (like in Figure 4.2) arise. It may seem like we are introducing bias in the CNV panel. However, matching the remaining signal to the imbalance panel, we can see that we are in fact removing signal indicating "complete loss" or noise. From $p = 0.70$ and above, aberrational patterns on the ratio-signal r are visible: in the right panel in which we look at possible imbalance patterns the left (P) arm clearly lacks a heterozygous genotype and ends with a complete loss of information (looks like a vertical bar). The right arm (Q) has the expected three genotype bands (homozygous on top and bottom, heterozygous in the middle, but it also seems to end with complete loss. These patterns were not (clearly) visible. Since the thresholding is performed on the s-signal, this is just signal that is cut off starting from the bottom and there-

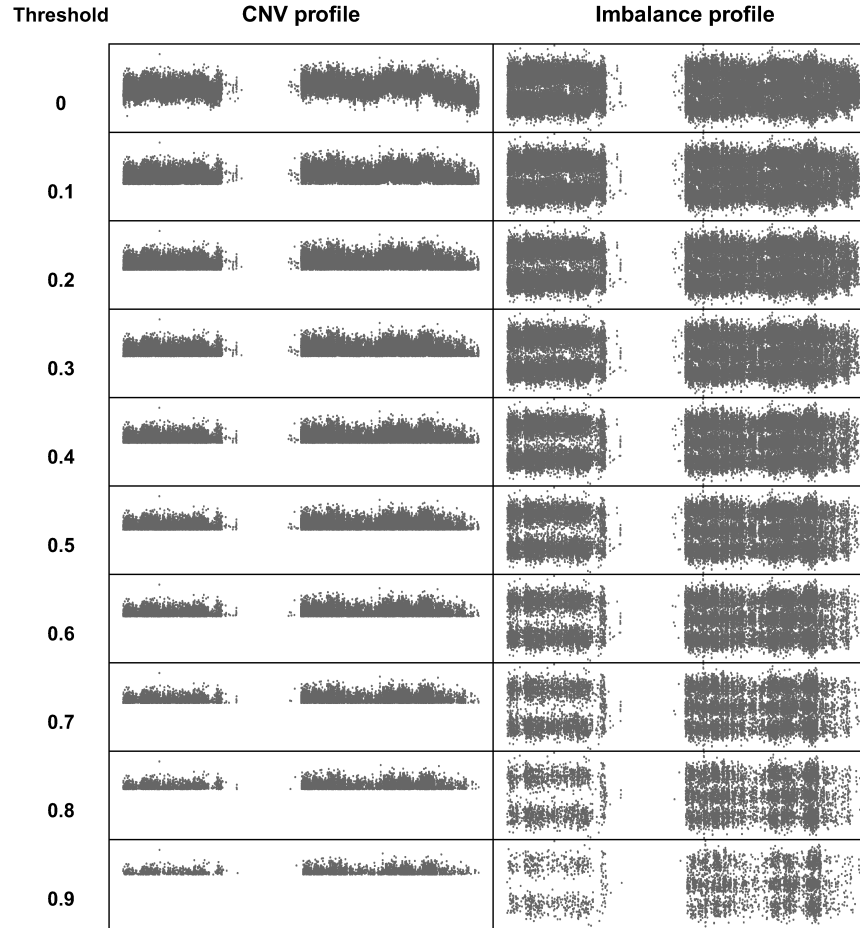

Figure 4.6: Aberration detection. Visual (and statistical) detection in a very low quality sample improves only after removing 80% of the originally low *s*-signal. Left: thresholded CNV signal. Right: thresholded imbalance signal. For CNV profiles, signal selection is not useful, while it is for allelic imbalance.

fore it doesn't benefit from threshold selection. Figure 4.7 shows that when a lower CNV signal (a deletion) is removed, this does not delete information

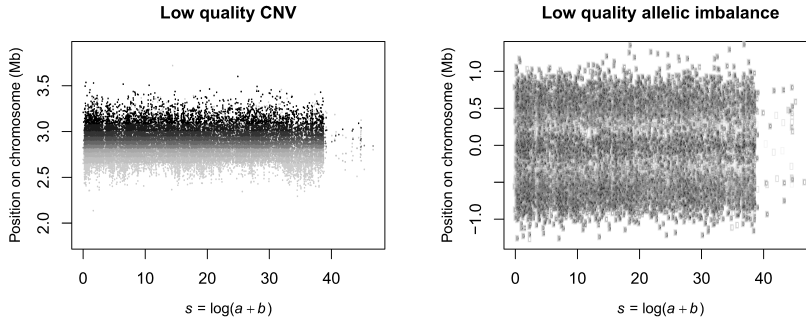

Figure 4.7: Gradient representation of the threshold levels. Lighter dots are removed in low(er) threshold levels. Left: CNV signal threshold. Right: imbalance signal thresholds. This illustrates that when a lower CNV signal (a deletion) is removed, this does not delete information about allele loss: it can still be retrieved from the imbalance panel.

about allele loss: it can still be retrieved from the LOH panel.

4.4 Discussion

We proposed a signal threshold approach to low-quality SNP samples for any given platform. A distribution-based approach provides a wayto ensure that one can never accidentally remove all SNPs; the best $(1 - p)$ % of the SNPs remain available. Applying such a threshold selection on low quality samples (strongly) improves array quality, defined by call rate. Threshold selection on high-quality samples doesn't have the same strong effect, but fortunately these samples do not *need* thresholding. The same effect holds for the effects of both genotype-free and genotype-based calibration; if sequentially applied the sample size is reduced, but the call rate and structural quality improve (strongly). The same holds for the detection of aberrational CNV andLOH patterns in low quality data; with increased threshold, the detection also improves, hopefully improving to a high enough quality level so that patterns are detectable. However, to provide more insight, numerical analyses should

be performed.

With respect to the above results and the fact that low-quality samples still have to be paid for, we suggest to make the best of the situation. The proposed threshold recovery procedure should be applied and evaluated before re-sending samples to the lab again. This approach saves money, time and effort and can be sufficient in many situations.

In this chapter we didn't include Illumina (Infinium) samples, since these arrays have asymmetric homozygous clusters, which is probably due to a dye saturation effect. We are currently looking into this; these results will be discussed elsewhere.