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

Rippe, R.C.A.

Genomic waves: where they come from, and how to eliminate them

5

Genomic waves are an undesirable distortion in copy number variation. It is generally assumed that they have real physical existence. We show that this is not true. Fluorescence signal on SNP arrays have a systematic bias, that varies strongly from SNP to SNP, giving the appearance of noise. Smoothing removes high frequency variations and so gives the impression of waves. The bias, and hence the waves, can be estimated and removed by a procedure called SCALA.

5.1 Introduction

Although SNP arrays were originally developed for genotyping of (normal) DNA, they are also a popular tool for studying copy number variations (CNV) and allelic imbalancein tumor samples. When studying CNV a persistent nuisance is the occurrence of “genomic waves", which compromise estimation accuracy due to unclear segment breakpoints. They become visible when the raw signal, the sum of the fluorescence intensities of the two alleles is smoothed sufficiently, as shown in Figure 5.2 for four different arrays.

The existence of waves has been reported frequently in both SNP arrays and aCGH profiles. Several remedies have been proposed. The wave phenomenon was first reported in aCGH profiles by Cardoso et al. (2004) and subsequently by Nannya et al., (2005) and Marioni et al. (2007). Cardoso et

This chapter is submitted as the article:
Rippe, R.C.A. and Eilers, P.H.C. (2012). Genomic Waves: where they come from, and how to eliminate them, *submitted for publication*.

al. referred to waves as a spatial bias, which they thought was due to non-constant specificity in the DNA amplification process. However, this idea was countered when the same pattern was seen in HapMap data. Nannya et al. introduced an algorithm that accounts for GC content (the percentage of nitrogenous bases that are either guanine or cytosine), which was extended in Lepretre et al. (2010). They proposed WACA (waves aCGH correction algorithm) that uses both GC content and size of the DNA fragments to correct for wave bias. However, Marioni et al. concluded after thorough evaluation that fitting a lowess curve through the profile was an improvement over GC correction. Also recently a procedure called NoWaves was proposed (Van de Wiel et al., 2010) to correct for wave bias in tumor profiles without using GC content, using ridge regression on (smoothed) normal profiles.

Genomic waves are also found in CNV profiles from SNP arrays, which are fundamentally different from aCGH profiles, because SNP arrays provide information on the (genotypes of the) two individual alleles. Komura et al. (2006) described genomic waves for this type of array and proposed the Genomic Imbalance Map algorithm that reduces signal noise by accounting for sequence characteristics of both probes and targets. The aCGH model from Nannya et al. proved effective for SNP arrays, too. Diskin et al. (2008) describe an algorithm that first quantifies the genomic waves in terms of GC content and uses this quantification as a predictor in a regression model. They also noted that, although commonly observed, genomic waves are not well understood. Marioni et al. thought it should be seen as spatial autocorrelation.

In reality autocorrelation does not exist, but is created by smoothing. In this paper we show that the cause of these waves is the existence of a systematic bias, characteristic for each allele of each SNP. Without smoothing it appears as noise but in fact it is reproducible, see Figure 5.1, which shows highly similar noise in four different arrays. The bias can be estimated as parameters in a linear model called SCALA (Rippe, Meulman & Eilers, 2012a). The model parameters can be estimated using an initial set of (high quality) arrays and the corresponding genotypes. Once the parameters have been estimated they can be used to correct these arrays and any new array that will become available.

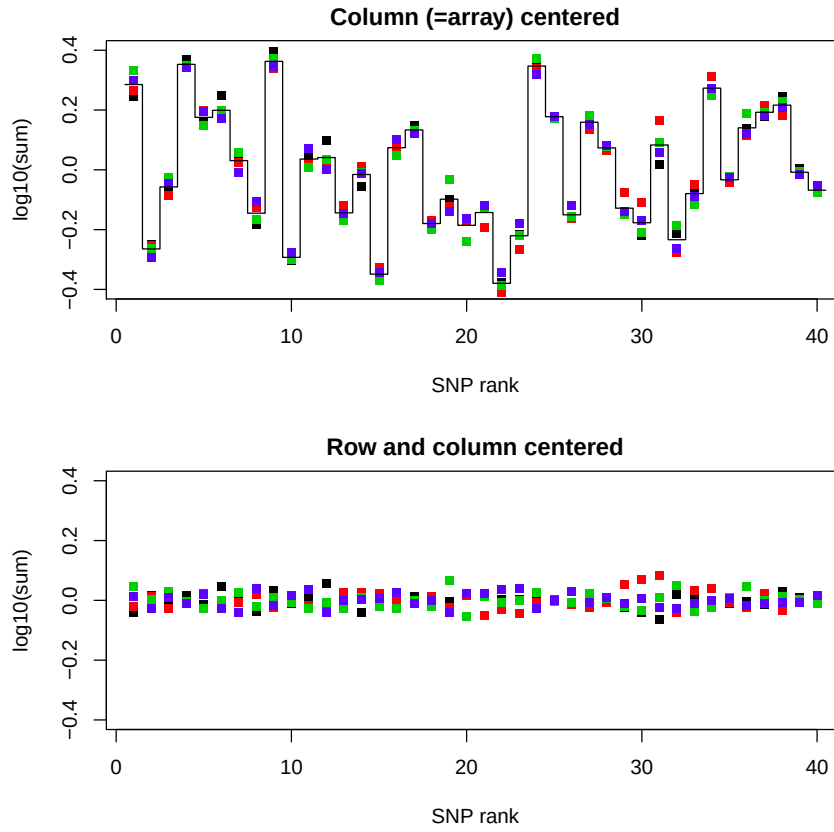

Figure 5.1: The source of the "waves" is systematic bias in the fluorescence signals. Shown are four different arrays, with highly similar noise. Subtracting the mean for each SNP (over arrays) for each SNP essentially eliminates the variation. This only works for normal DNA.

This procedure is easily applicable and therefore we feel it can and should always be applied.

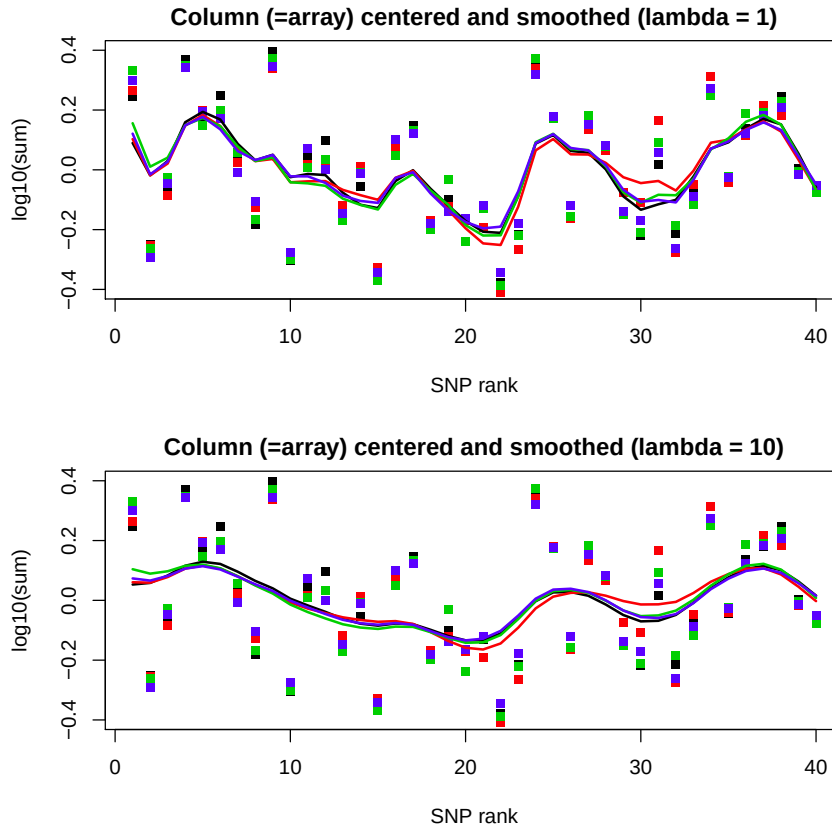

Figure 5.2: An illustration of how smoothing produces "waves", although the raw signals are unstructured. Shown are the first 40 SNPs on chromosome 9. For a clearer display, the positions of the SNPS in the graphs are their ranks, not their physical positions. The Whittaker smoother is used, with two values of the parameter λ.

5.2 Methods

Data and preprocessing

We use Affymetrix 250k NSP tumor profiles from the Erasmus Medical Center (Bralten et al., 2010) and high quality reference profiles from the same

Affymetrix platform. Chromosomes 1 to 22 are analyzed. The non-autosomal chromosomes X and Y are neglected due to the fact that the calibration approach in SCALA requires signals for two alleles, which is impossible to obtain in the Y chromosome. We start from averages of fluorescence intensity over probe sets; information on the individual probes is not used. We transform the signals for the two alleles, a and b, to a single profile $s = log_2(a + b)$.

The origin of the waves

A simple illustration of our claim that each SNP shows a reproducible bias is presented in Figure 5.1. It shows s for the first 40 SNPs (as determined by their position) of chromosome 9. Four high-quality arrays, to which normal DNA was hybridized, were used. Each array was centered by subtracting the mean of s (over the 40 SNPs). To make it easier to see the data for the individual SNPs, their ranks are used for the horizontal coordinate, and not their physical position on the chromosome. From the top panel it is clear that the levels vary strongly from SNP to SNP, but that they are similar within each individual SNP. If we subtract the means per SNP the lower panel is obtained which show much smaller variation and almost no systematic patterns.

This would be a good method to correct data from normal DNA, but copy number variations are not much studied for normal DNA. However, one can compute means per SNP for a set of “normal" arrays and use these values for correcting any other array. In what follows we will present a more advanced allele-specific correction method.

Smoothing

We use the Whittaker smoother (Whittaker, 1923; Eilers, 2003), assuming equally spaced pseudo-positions. This is a simplification, but as we only use the smoothing for illustration, it can do little harm. Results are shown in Figure 5.2, for two values of the smoothing parameter λ. The Whittaker

smoother minimizes the penalized sum of squares

$$Q = \sum_i (s_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2,$$

where z represents the smooth series and Δ is the operator that forms second order differences: $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2})$.

As Figure 5.2 shows, smoothing leads to “waves", even though the unsmoothed data make large jumps from SNP to SNP. Because the “waves" are very similar for the four arrays, it is easy to mistake them for a real spatial pattern, but actually they are an artifact.

The SCALA model

Let Y be a matrix with logarithms of fluorescence intensities for one allele. The rows, indexed by i, represent the SNPs and the columns, indexed by j, the arrays. The SCALA model is defined for any allele signal y_{ij} as

$$y_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{3} \gamma_k h_{ijk} + e_{ij} \tag{5.1}$$

where μ is the grand mean, α_i describes the overall level of SNP i, β_j describes the overall intensity level of array j, k is the genotype code with $1 = AA$, $2 = AB$, $3 = BB$ (we work with normal DNA) and γ_k is a parameter for genotype k. The genotypes are coded in $H = \{h_{ijk}\}$. H is a 3-dimensional indicator array; for each combination of i and j we have a 1 in layer that is indicated by the genotype, and 0 in the other layers. To make the model identifiable we introduce the constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. Details on the estimation procedure are described in Rippe et al. (2012a).

In **correction by SCALA**, the model is fitted for each of the two alleles separately. After fitting, we obtain the parameter vectors $\boldsymbol{\alpha} = [\alpha_i]$. These obtain corrected signals by

$$Y_j^c = Y_j / 10^{\alpha}. \tag{5.2}$$

Correction by NoWaves , which finds regression coefficients for each SNP i by

$$Y = \sum_{i=1}^{n} \beta_i Z_{ij} + \epsilon_j \tag{5.3}$$

with Z_j the smoothed (normal) reference profile.

The coefficients β are estimated using ridge shrinkage on the reference profile parameters, through

$$\beta^* = \text{argmin} \left(\sum_{j=1}^{s} \left(Y_j - \sum_{i=1}^{n} \beta_i Z_{ij} \right)^2 + \delta \sum_{i=1}^{n} \beta_i^2 \right) \tag{5.4}$$

with δ the coefficient shrinkage parameter, which is determined through leave-one-out crossvalidation and hence is sample-dependent. Signal correction is then ensured by:

$$Y_j^c = Y_j - \sum_{i=1}^{n} \beta_i^* Z_{ij}. \tag{5.5}$$

Correction performance

To find a smooth estimate for the CNV profile we use the L_2 norm smoother, as proposed by Whittaker (1923) which minimizes

$$L_2 = \sum_{i=1}^{m} (s_i - z_i)^2 + \lambda \sum_{i=2}^{m} (z_i - z_{i-1})^2, \tag{5.6}$$

where the original signal s is of length m and z is the approximate smooth series of s. The smoothness is determined by λ. Larger λ provides a smoother series z, but has a worse fit to the data y. It is common practice to find the optimal amount of smoothing, but here we do not aim to find an optimal value for λ. We use the P-spline implementation by Eilers & Marx (1996).

To quantify the effect of wave removal we compute the normalized difference $d = \sum_i (s_i - z_i)$ between the raw signal s and the smooth profile z (for each SNP i) on a given chromosome. Formally, we write

$$d = \frac{\sum |s_i - z_i|}{n}. \tag{5.7}$$

For the smooth series z we fix $\lambda = 100$. An increase of d indicates more scatter in the SNP signals, re-lative to the smooth estimate. For detection of constant segments between sharp breakpoints de-dicated (and better) algorithms are available, but here we aim for the removal of waves with gradient properties.

5.3 Empirical results

First we visually illustrate the origin of waves and then numerically compare the models discussed above.

Wave origins

In the left panels in Figure 5.3 the uncorrected signals are shown. Each panel contains two parts: on the left of the dashed line a healthy chromosome 1 is shown, while to the right of the dashed line a tumor chromosome 9 is shown. We display only a small selection of observation from one profile because different tumor patterns in different arrays would clutter the image. The top and middle row show the profiles for allele a and b separately, the bottom row shows the actual copy number signal $s = \log(a + b)$. The right column illustrates SCALA correction by α_i. It can be seen from Figure 5.3 that SCALA calibration with just the SNP parameter α_i is not effective for signals from a single allele a or b, but it is for the (logarithm of the) sum. Also note that all corrections do not remove copy number segments (as seen in the right part of each panel).

Numerical evaluation

We first visually inspect the results for SCALA and NoWaves. The top panels in Figure 5.4 show waves in an Affymetrix 250k tumor sample for two selected chromosomes (1 and 9). It is clear that the wave patterns occur on both healthy (1) and tumor (9) tissue. All panels in Figure 5.4 have the same scales on both the x and y axes. First, Figure 5.4 shows that after SCALA calibration (bottom row), the profiles hardly show any waves. The results for

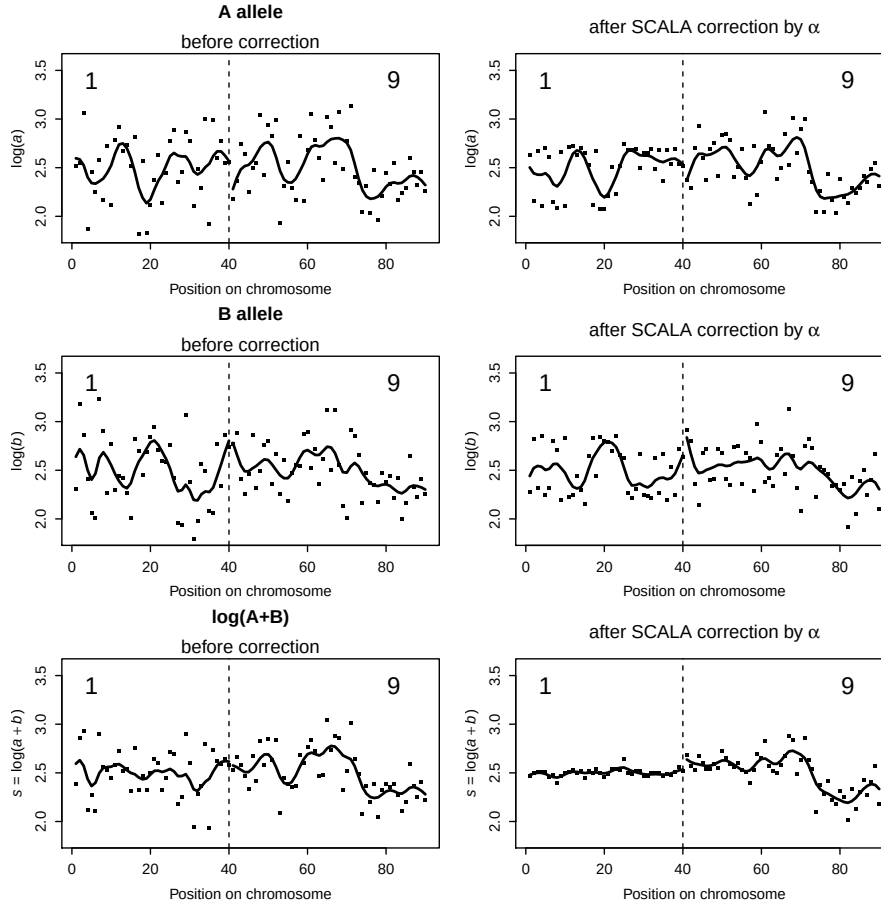

Figure 5.3: Wave patterns in real data. The horizontal axis shows the position of each observation in the sequence. The vertical axis shows either $\log(y)$ with y the allele signal for a or b, or $\log(a+b)$. Left column: uncalibrated signals. Right column: signals after SCALA calibration with α. Top panels: A allele, middle panels: B allele, bottom panels: CNV signal. Left parts of each panel show a healthy chromosome 1; right parts show an unhealthy chromosome 9. Smooth profiles are obtained with the Whittaker smoother ($\lambda = 2000$).

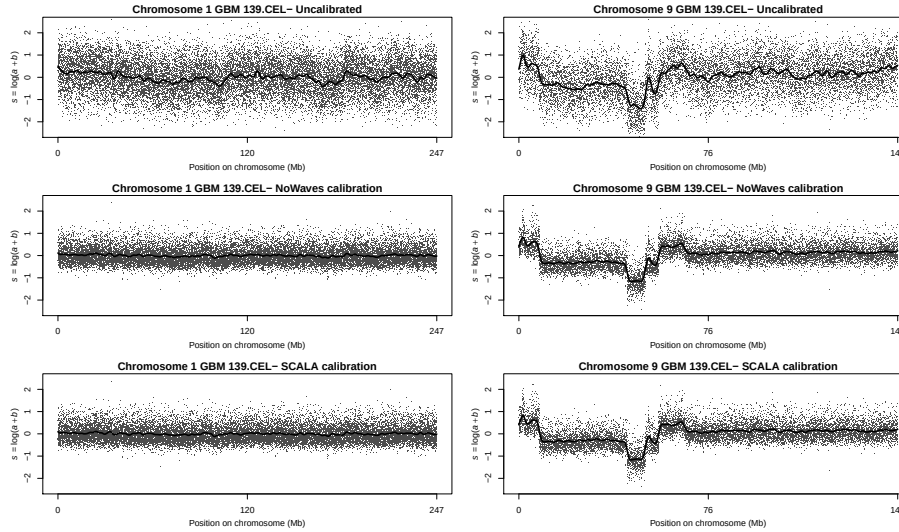

Figure 5.4: Profiles before (top) and after NoWaves (middle) and SCALA (bottom) calibration.

NoWaves (middle row) are similar to equivalent. Removing the waves from the signals clearly keeps CNV segments intact and quantifiable. In fact, the aberrations are the *only* deviations from the reference level $2n$ (2 alleles) that are still visible/detectable.

For the crude data, we find a benchmark d value of 0.60 (0.000-2.662) for chromosome 1 and 0.59 (0.000 -2.922) for chromosome 9. However, applying signal calibration based on SCALA, we find d values for both SCALA and NoWaves of 0.293 with the first ranging (0.000-2.297) and the second (0.000-2.328) for chromosome 1 and 0.295 (0.000-1.786) against 0.297 (0.000-1.745) for chromosome 9. Note that any differences between the latter methods are in the order of 10^{-3}.

Detailed results for chromosome 1 to 22 in several samples, for four different levels of smoothing ($\lambda \in 1, 10, 100, 1000$) are provided in Appendix A. Here too, differences between calibration methods are very small, but large compared to uncalibrated signals.

5.4 Discussion

We have illustrated that the cause of waves in CNV profiles based on SNP fluorescence signals is not spatial autocorrelation. Visual and numerical comparisons between two signal calibration methods, NoWaves and SCALA were made. The first method was developed specifically for single aCGH signals, whereas the second method was developed for two allele channels. The results for the two methods show almost equal improvements. The fact that model-based calibration is effective can be explained by the fact that SNP variation is larger than genotype variation, given that the calibration parameters were computed with only 8 profiles. Therefore, the maximum amount of genotype variation is low by definition.

It can be argued that after transformation to $s = \log_2(a + b)$, the NoWaves correction is already effective, so there is no need for a SCALA correction. However, NoWaves aims solely at wave removal for single channel profiles, while SCALA aims for allele-level correction, which is impossible for NoWaves. Another major advantage of SCALA over NoWaves is that the first calibrates signals with a set of parameters that is calculated only once and can be re-used in later instances, while the latter method needs to recompute the projection for every analysis. The smoothed references profiles can of course be re-used here, too. The SCALA calibration has a very simple nature, subtracting a vector of parameters. Therefore, we argue that it should always be applied, because it require hardly any time, removes waves and leaves segmentation intact.

One of the differences between SCALA and other methods is that for better correction, instead of GC content it exploits genotypes of the reference samples from which the calibration parameters are obtained. This introduces an extra step and thus an extra level of error-proneness. However, since calibration parameters are estimated using high quality reference samples and these data the genotype calls can be made very accurately, this does not pose a threat to the procedure.

It might also be argued that calibration is not necessary when a large amount of smoothing is applied on the uncalibrated data, since this already

removes most of the waves. However, in the right panels (Figure 5.4) we also see that within remaining CNV segments waves still distort the patterns. This problem is absent in the calibrated signals. Furthermore, applying too much smoothing on the raw data will in the end smooth out CNV segments.

In the current work we used a smoother based on the L_2 norm, but in Eilers & DeMenezes (2005) the L_1 norm is illustrated to be more effective in CNV detection. A further refinement to the L_0 norm was proposed by Rippe et al (2012b). The latter norms are much more suitable to detect aberrated regions, since it does not tend to round segment breakpoints (and the L_2 does, true to its quadratic nature). However, both the L_1 and L_0 norm do not respect the wave curvature and hence are not effective in the specific application described here.

Acknowledgements

We acknowledge Mark van de Wiel (VUMC, Amsterdam, The Netherlands) for providing assistance with the NoWaves software.