BIOINFORMATICS

Vol. 00 no. 00 200x
Pages 1–4

SNP calibration on Illumina BeadArrays

Ralph C.A. Rippe [1], Paul H.C. Eilers [1,3]*and Jacqueline J. Meulman [2]

[1]Department of Educational Sciences, Datatheory Group, Leiden University.
[2]Department of Mathematics, Leiden University.
[3]Department of Methodology and Statistics, Utrecht University.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Fluorescence signals from Illumina BeadArray SNP arrays show strong and persistent systematic patterns, that can be modeled accurately by relatively simple linear statistical models.

Results: Parameters from the model can be used to correct the signals and sharpen genotype clusters.

Availability: R Software is available: SCALA.

Contact: peilers@fsw.leidenuniv.nl

1 INTRODUCTION

The use of SNP arrays has been exploding in recent years. Some highly visible projects are HapMap (Altshuler *et al*., 2005; HapMap Consortium, 2003, 2007) and the CRI breast cancer study study with 5000 cases and 5000 controls (Chin *et al*., 2007; Easton *et al*., 2007; Laframboise*et al*., 2007). In this paper we study Illumina Golden Gate bead arrays (Fan *et al*., 2006; Shen *et al*., 2005). Illumina developed two different formats, both classified as 'Array of Arrays'. The first format is called the 'Sentrix BeadChip'. This type uses a silicon substrate that allows 1 to 16 samples at the same time. The second format, the 'Sentrix Array Matrix' uses an 8x12 array grid that is compatible with standard microtiter plates. We use the latter format, which will be explained in more detail later. The strength of fluorescence signals varies strongly and systematically between SNPs. These stable patterns can be used for signal correction; we call this SNP calibraton. We present two statistical models for observed intensities and given genotypes. We estimate for each SNP a characteristic parameter for the intensity of the fluorescence signal. One model allows calibration before the genotypes have been estimated, using parameters that have been derived from a set of reference arrays. The second model uses initial genotyping to re-estimate improved parameter estimates for new sets of arrays. Calibration is of great value, because it sharpens allele-specific clusters, delivering opportunities for improved genotyping.

As a by-product of our analyses we find a color-dependent anomaly in fluorescence ratios between heterozygote and homozygote genotypes.

*correspondence to peilers@fsw.leidenuniv.nl

2 METHODS

We analyzed data from Illumina Golden Gate BeadArrays grouping 96 arrays in a metal frame. Each array carries 1624 unique bead types, in approximately 30-fold redundancy. Each bead type is covered with SNP-specific dye-labeled oligonucleotides. Therefore, in each array we have a bundle of approximately 50.000 beads. There are four groups of 24 arrays representing four linkage panels containing 1624 SNPs. Each panel covers a number of chromosomes. The data were provided by the Pathology Department of the Leiden University Medical Centre (Lips *et al*., 2005).

Assuming that SNP i has a specific intensity level a_i, and that array j has a normalization factor b_j, a reasonable model for the intensity of the (red or green) fluorescence signal is given by

$$x_{ij} = a_i b_j u_{ij} + e_{ij},$$

where u_{ij} represents the number of alleles of a specific color and e_{ij} represents the error (which might have a complex distribution). This multiplicative model becomes linear if we use log-transformed data. Suppose $X_c = \{x_{ijc}\}$ represents fluorescence for color c and $Y_c = \log(X_c) = \{y_{ijc}\}$ with $i = 1, ..., m$ and $j = 1, ..., n$. We use logarithms to base 10.

In Rippe *et al.* (submitted) the following model is proposed:

$$y_{ijc} = \mu_c + \alpha_{ic} + \beta_j + \sum_{k=1}^{3} \gamma_{kc} h_{ijk} + e_{ijc}, \quad (1)$$

where μ_c is the grand mean, α_{ic} describes the overall level of SNP i, β_j describes the overall level of array j, k is the genotype code with $1 = RR$,$2 = RG$,$3 = GG$ and γ_{kc} is a parameter for genotype k. The genotypes are coded in $H = \{h_{ijk}\}$. H is an indicator array; for each combination of i and j we have a 1 in layer k, and 0 in the other layers. To make the model identifiable we introduce the constraints $\sum_i \alpha_i = 0 \forall c$ and $\sum_j \beta_j = 0$. We call this the global γ model, because all SNPs share the same genotype parameters.

We propose a new model that finds 3 genotype parameters for each individual SNP i. This model equals

$$y_{ijc} = \mu_c + \beta_j + \sum_{k=1}^{3} \gamma_{ikc} h_{ijk} + e_{ijc}. \quad (2)$$

Here γ_{ikc} is a parameter for genotype k of SNP i. To make the model identifiable we introduce the constraint $\sum_j \beta_j = 0$. We call this the local γ model.

For μ_c we take the overall mean (over all SNPs and arrays). We estimate the other parameters using least squares, implemented as block relaxation. This is a natural choice, because it is easy to compute one set of parameters (α, β or γ) by averaging if the others are known. We cyclically update each set in turn. Convergence is quick: less than 10 iterations give six-figure precision.

All computations are performed by a special purpose R program (R Team, 2008), called SCALA (for SNP CALibration Algorithm). A graphical

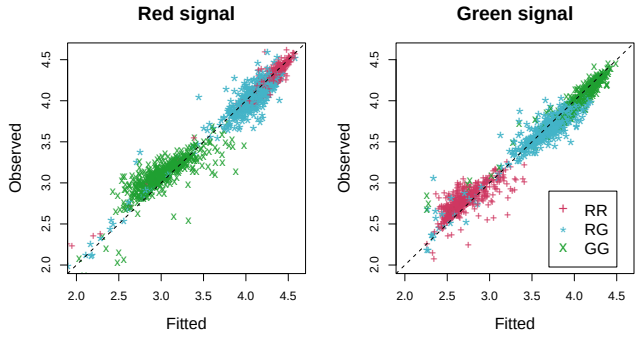

Fig. 1. Data versus fitted values from the global model (1) for a typical array (6).

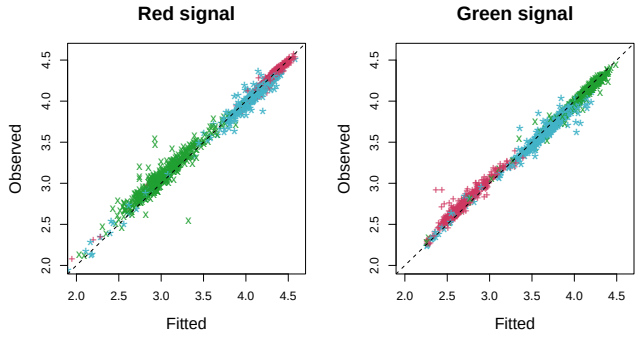

Fig. 2. Data versus fitted values from the local model (2) for the same typical array (6).

user interface allows easy selection of arrays, computation of models, and specification of numerical and graphical output. The program is available on request, from the first author.

3 RESULTS

3.1 Model fit

Figure 1 shows the fit of the global model in (1) for a typical array from linkage panel 1. Figure 2 shows that the local model in (2) gives a much better fit. The local model fits better in all four linkage panels.

3.2 An anomaly

We expect fluorescence intensities to show ratios 0,1 and 2 (2,1 and 0) for GG, RG, RR in the red (green) signal. We never really get 0, due to noise and background signals. But we expect the red RR (green GG) signal to be twice as strong as that for RG. We refer to the ratio RR/RG and GG/RG as the Double-Single Ratio (DSR). The observed values are presented in Table 1. It shows that the γs do not exactly follow the $\log(0, 1, 2)$ pattern. Figure 3 visualizes this effect for Linkage panel 1. Let γ^o be the 'homozygotic' γ (RR for the red signal and GG for green). In this figure we plotted $\gamma - \gamma^o$ against γ^o. Hence, we get differences of all genotype quantifications per SNP wrt the baseline γ^o. The horizontal lines indicate the average difference with γ^o for the heterozygotic genotype. The red

Table 1. The global (top) and averaged local (bottom) Double-Single Ratio (DSR) on linear scale for the red and green signal.

	Panel 1	Panel 2	Panel 3	Panel 4
Global DSR Red	1.517	1.513	1.522	1.515
Global DSR Green	2.247	2.238	2.194	2.101
Local DSR Red	1.547	1.586	1.581	1.582
Local DSR Green	2.289	2.280	2.193	2.132

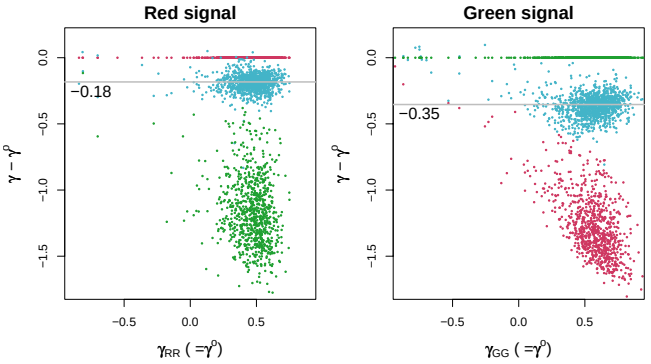

Fig. 3. Linkage Panel 1: Genotype patterns γ_{ikc} versus the homozygotic SNP level (RR for the red signal (left) and GG for the green signal (right)).

difference is smaller than 0.3 (= $\log 2$), whereas the green difference slightly exceeds this value. The anomaly occurs in results from both the global model and the local model. The green signal behaves as expected, but the red signal strongly deviates consistently. A similar effect is also reported by Staaf *et al*. (2008). We can not provide an explanation at this point.

3.3 Cluster sharpening

To calibrate with the global model, we compute $y^*_{ij} = y_{ij} - \alpha_i$, where the index for color has been dropped for clarity. Note that the genotypes are not involved. We could optionally calibrate by computing $y^*_{ij} = y_{ij} - \alpha_i - \beta_j$, to make arrays more comparable, but because we do genotyping on whole arrays, but we do not need this. Results are shown in the second row of panels in Figure 4. The clusters are less diffuse and the stray dots in the SW corner have been moved into the clusters.

To calibrate with the local model, we compute $y^*_{ij} = y_{ij} - \sum_k h_{ijk}\gamma_{ik}$, where again the index for color has been dropped. Note that the genotypes are now involved: $\sum h_{ijk}\gamma_{ik}$ effectively selects the right γ parameter for a genotype. Results of this correction are shown in the third row of panels in Figure 4. The clusters have become even more concentrated.

3.4 Improved genotyping

Genotyping algorithms come in two flavors: SNP-based and array-based. In the former, clusters of arrays are computed for each

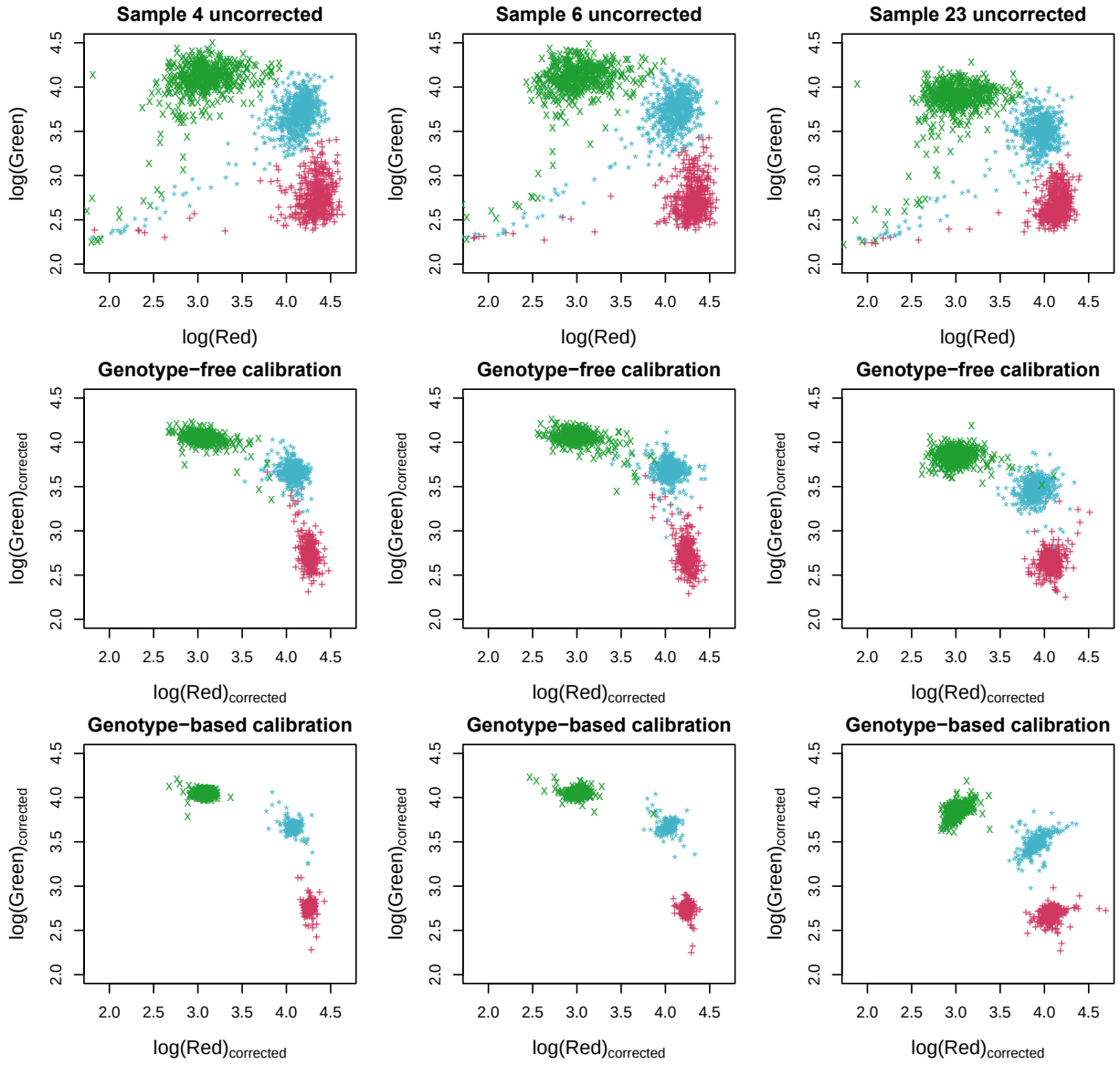

Fig. 4. The top panels show the $\log_{red}$ and $\log_{green}$ signal for three selected arrays from Linkage Panel 1, before calibration. The middle panels show the effect of calibration with α_{ic} from the global model. The bottom panels show the arrays after calibration with the selected γ_{ik}, from the local model proposed in this paper. The bottom panels illustrate the effect of calibration: the genotype clouds show a much better separation.

SNP. This appears to be the standard in commercial software. Alternatively, clusters of SNPs can be computed per array (Teo *et al.*, 2007; Xiao *et al.*, 2007). We prefer the latter approach, because it is faster, allows much better quality control, and even individual arrays can be genotyped. We developed our own algorithm, which we summarize below.

The left panel in Figure 5 shows the log of the ratio of green and red fluorescence against the log of their sum. Three elongated clusters are visible, representing the GG, GR and RR genotypes. As a model for each cluster we assume noisy (normally distributed) observations around a linear regression line. Each cluster has its own slope, offset and noise variance. This mixture of regression lines can be estimated conveniently with the R package FlexMix (Leisch, 2004). In the left panel of Figure 5 the estimated regression lines are shown. From the parameters of the mixture components follow (as standard output of FlexMix), for each SNP, the probabilities of it being a member of each of the clusters. The highest of the three probabilities indicates the most likely cluster and hence the genotype. The colors and the markers of the observation in the graph have been chosen to show to which cluster they have been assigned.

4 DISCUSSION

We have proposed two linear models for (logarithms of) fluorescence intensities. A good to excellent fit to the data was

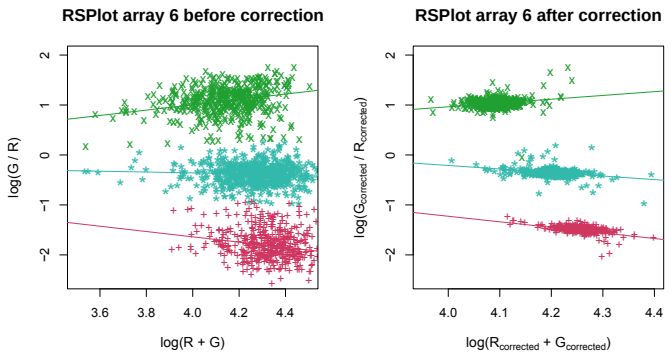

Fig. 5. Left: Ratio-Sum input for the genotype estimation algorithm (FlexMix) before correction. Right: Ratio-Sum input after calibration.

obtained. The simple global model provides (for red and green fluorescence separately) a set of SNP-specific parameters that can be used for calibration before genotyping. The extended local model makes use of estimated genotypes to further improve calibration. Calibration results in sharper clusters in the Red-Green plane. We expect that genotyping algorithms can be improved by calibration. One of our future goals is to investigate this thoroughly.

We observed a persistent anomaly in the ratio of intensities from homozygous to heterozygous genotypes. In Green it is close to the expected value of 2, but for Red it is around 1.5. Staaf *et al*. (2008) report a similar phenomenon. We cannot offer an explanation at this point.

Bengtsson *et al*. (2008) use probe-level data from Affymetrix arrays for SNP calibration. Their method is quite involved and does not apply to Illumina arrays. We plan to compare their method to our straightforward approach, applied to Affymetrix arrays.

ACKNOWLEDGEMENT

This work was supported by the Pathology Department of the Leiden University Medical Centre (LUMC).

REFERENCES

Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., Donnelly, P., et al. (2005) A haplotype map of the human genome, *Nature*, **437**, 1299-1320.

Bengtsson, H.,Irizarry, R., Carvalho, B. and Speed, T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level, *Bioinformatics*, **24(6)**, 759-767.

Chin, S. F., Wang, Y., Thorne, N. P., Teschendorff, A. E., Pinder, S. E., Vias, M., et al. (2007) Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers, *Oncogene*, **26**, 1959-1970.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci, *Nature*, **447**, 1087-1093.

Fan, J. B., Gunderson, K., Bibikova, M., Yeakly, J.M., Chen, J., Wickam-Garcia, E., et al. (2006) Illumina universal bead arrays, *Methods in Enzymology*, **410**, 57-73.

Giannoulatou, E., Yau, C., Colella, S., Ragoussis, J., & Holmes, C. (2008) Genosnp: a variational bayes within-sample snp genotyping algorithm that does not require a reference population, *Bioinformatics*, **24(19)**, 2209-2214.

The International HapMap Consortium (2003) The International HapMap project, *Nature*, **426**, 789-796.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs, *Nature*, **449**, 851-861.

Laframboise, Thomas, Harrington, David, Weir, & Barbara, A. (2007) Plasq: a generalized linear model-based procedure to determine allelic dosage in cancer cells from snp array data, *Biostatistics*, **8**, 323-336.

Leisch, F. (2004) FlexMix: A general framework for finite mixture models and latent class regression in R, *Journal of Statistical Software*, **11(8)**, 1-18.

Lips, E., Dierssen, J., Eijk, R. van, Oosting J., Eilers, P.H.C., Tollenaar, R., et al. (2005) Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays, *Cancer Research*, **65**, 10188-10191.

R Development Core Team (2008) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, **ISBN 3-900051-07-0**, URL http://www.R-project.org.

Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., et al. (2005) High-throughput snp genotyping on universal bead arrays, *Mutat Res*, **573**, 70-82.

Staaf, J., Vallon-Christersson, J., Lindgren, D., Juliusson, G., Rosenquist, R., Höglund, M., Borg, A., & Ringner, M. (2008) Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios, *BMC Bioinformatics*, **9:409**.

Teo, Y.Y., Inoure, M., Small, K.S., Gwilliam, R., Dekoulas, P., Kwiatkowski, D.P., et al. (2007) A genotype calling algorithm for the Illumina beadarray platform, *Bioinformatics*, **23-20**, 2741-2746.

Xiao, Y., Segal, M.R., Yang, Y.H., & Yeh, R.F. (2007) A multi-array multi-snp genotyping algorithm for affymetrix snp microarrays, *Bioinformatics*, **23-12**, 1459-1467.