BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 3 2013, pages 299–307
doi:10.1093/bioinformatics/bts683

Genome analysis

Advance Access publication November 29, 2012

PAIR: paired allelic log-intensity-ratio-based normalization method for SNP-CGH arrays

Shengping Yang[1], Stanley Pounds[2], Kun Zhang[3,*] and Zhide Fang[1,*]

[1]Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, LA, [2]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN and [3]Department of Computer Science, Xavier University of Louisiana, New Orleans, LA, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Normalization is critical in DNA copy number analysis. We propose a new method to correctly identify two-copy probes from the genome to obtain representative references for normalization in single nucleotide polymorphism arrays. The method is based on a two-state Hidden Markov Model. Unlike most currently available methods in the literature, the proposed method does not need to assume that the percentage of two-copy state probes is dominant in the genome, as long as there do exist two-copy probes.

Results: The real data analysis and simulation study show that the proposed algorithm is successful in that (i) it performs as well as the current methods (e.g. CGHnormaliter and popLowess) for samples with dominant two-copy states and outperforms these methods for samples with less dominant two-copy states; (ii) it can identify the copy-neutral loss of heterozygosity; and (iii) it is efficient in terms of the computational time used.

Availability: R scripts are available at http://publichealth.lsuhsc.edu/PAIR.html.

Contact: zfang@lsuhsc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 20, 2012; revised on November 13, 2012; accepted on November 22, 2012

1 INTRODUCTION

Copy number alterations (CNAs) have been associated with many genomic disorders (Fanciulli *et al*., 2007; Yang *et al*., 2007). Insertion or deletion of DNA sequences can directly alter the gene expression levels, and thus potentially cause genetic diseases (McCarroll *et al*., 2007). For example, in a St. Jude study of >200 B-progenitor and T-lineage acute lymphoblastic leukemia (ALL) patients, >50 recurring regions of CNAs were identified. In addition, many of these recurring CNAs were specific to different subtypes of ALL, associated with prognosis, indicating that resistance to therapy in ALL might be caused by specific genetic changes (Mullighan, 2009). In another St. Jude study, deletion, amplification and other types of CNAs were found to be in 40% of B-progenitor ALL cases (Mullighan *et al*., 2007).

Facilitated by the Human Genome Project, the development in comparative genomic hybridization (CGH) array technology has made it indispensable for CNA study. Among many types of CGH arrays, the single nucleotide polymorphism-CGH (SNP-CGH) array is widely used because of its high resolution and its ability to provide genotype estimates (Curtis *et al*., 2009; Scharpf *et al*., 2008; Ylstra *et al*., 2006). Although advances in next-generation sequencing are rapidly changing the landscape of genetics, the SNP-CGH array offers a cost-effective way in CNA studies, complementary to next-generation sequencing (Przybytkowski *et al*., 2011). In addition, the SNP-CGH array has advantages in voluminous, publicly available data and a wealth of data analysis tools. Integrative analysis of genomic data generated from different platforms requires SNP-CGH data of high quality.

The raw signals from the SNP-CGH array are inherently noisy because of sampling and/or technical variation. To reduce such noise, the logarithms of signal intensity ratios between tumor and normal samples from the same individual are usually used for CNAs detection (Mullighan *et al*., 2007). A large portion of the sampling variation may be removed this way because of strong similarities between paired tumor and normal samples. In addition, tumor/normal pairs are expected to be run in the same laboratory and in the same batch to remove experimental variations as much as possible.

Loss of heterozygosity (LOH) describes the loss of normal function of one allele in a gene, and LOH detection plays an important role in discovering DNA segments that harbor tumor suppression genes (Staaf *et al*., 2008). Copy-neutral LOH (cnLOH) refers to duplication of one of the two alleles and concurrently loss of the other allele, so that there is no change in a DNA copy number. As a result, it is difficult to detect cnLOH by conventional copy number measurements. On the other hand, the methods for accurate cnLOH detection often require both tumor and control samples from the same patient.

Many statistical models and algorithms have been proposed for CGH array-based CNA analysis (Fridlyand *et al*., 2004; Hupé *et al*., 2004; Marioni *et al*., 2006; Olshen *et al*., 2004). Normalization is critical for obtaining biologically meaningful results. An appropriate normalization method should be able to substantially remove systematic variations, while signal changes because of biological alterations are preserved. Some of the early CGH data normalization methods directly adopt those developed for microarray gene expression, such as Quantile normalization, Rank-invariant set normalization, Lowess normalization and so forth (Chen *et al*., 2008; Pounds *et al*., 2009). A fundamental assumption in these algorithms is

*To whom correspondence should be addressed.

that the distributions of the log-signal intensities from all samples are the same. However, this assumption may not hold in a CNA study. For example, nearly all solid tumor cancers exhibit whole-chromosome gains/losses (De Vita *et al*., 2008), and the proportion of gain/loss probes in the tumor genome can vary substantially from sample to sample. These translate into substantial differences in distributions of log signal intensities of different tumor samples. Therefore, directly applying normalization algorithms developed for gene expression data might result in 'over-normalization' of segments of CNAs (van Houte *et al*., 2009). To address this problem, several methods have been proposed specifically for CGH array normalization. For example, based on genotype calls and log-ratio of signal intensities, a reference alignment method identifies chromosomes in two-copy state, which are then used as internal references in normalization/alignment (Pounds *et al*., 2009). One problem in this method is that the number of reference chromosomes identified is usually small, and thus they might not represent the two-copy probes well. Another method, popLowess, identifies probes in two-copy state using k-means clustering on the segmented mean values, and subsequently normalizes the data by Lowess regression using probes identified as references (Staaf *et al*., 2007). The weakness of this approach lies in that only coarse separation can be achieved because of probe pre-selection, and that k-means clustering requires a pre-determined number of clusters. The ridge-tracing normalization method normalizes CGH data by applying ridgeline regression on the highest ridgeline of a 2D log-intensity distribution and then using the expectation maximization algorithm to centralize the copy number ratio. This method assumes that the most frequently occurring copy number probes correspond to the two-copy states (Chen *et al*., 2008). More recently, iterative normalization strategies, such as ITALICS (for one-channel array) and CGHnormaliter (for two-channel array), have been proposed (Rigaill *et al*., 2008; van Houte *et al*., 2009). In both cases, segmentation is the key step, and the subsequent normalization depends on the results of segmentation. TumorBoost (Bengtsson *et al*., 2010) was developed to normalize the allele B fraction in tumor, whereas total copy number signals are not normalized.

In this article, we propose a new normalization procedure for SNP array data. This is motivated by the following reasons. (i) Although applying paired samples in experimental design could substantially reduce noise in data, estimating the true copy number is often complicated by other experimental and biological factors, such as imperfect dissection of the primary tumors, the existence of the multiple clones in the tumor cell population, ploidy of a sample, the spatial signal variation and so forth. These factors can reduce the expected signal–noise ratio (SNR) and make it difficult to accurately estimate the true underlying copy number (Fridlyand *et al*., 2004). As a result, the conventional segmentation methods such as circular binary segmentation (*CBS*) (Olshen *et al*., 2004), which works well in partitioning a SNP sequence into segments of the same underlying copy number, do not provide estimation of the underlying copy number for each segment. (ii) Almost all current normalization methods do not take into consideration the existence of cnLOH. Since the mean value of probes in a cnLOH segment may be slightly different (but not significantly) from the mean value of probes in segments with true normal two-copy state, it is difficult to identify and subsequently remove them from the reference probe set. Therefore, it is necessary to develop new strategies that have strong statistical power in distinguishing probes in two-copy state from those having CNAs and simultaneously take into account the existence of cnLOH.

We will briefly present Affymetrix SNP arrays in the next section, and then describe the proposed algorithm in two parts. The first part is to partition probes into the segments of *UMS* (uni-modal state, defined in Section 2) or *BMS* (bi-modal state) using a Hidden Markov Model (*HMM*). The second part is to apply the *CBS* segmentation method on the piecewise average log-ratio (*PAlog-Ratio*, defined in Section 2.3) to separate probes in two-copy state and other even-number copy number states. The application of our method on a publicly available dataset, followed by a comparison of our method with two currently available methods, is presented in Section 3. At the end of the article, we discuss the advantages and limitations of our method and possible improvements.

2 METHODS

2.1 Affymetrix GeneChip Mapping 500 K array

There are several widely used SNP-CGH array platforms, including Affymetrix genome-wide human SNP Array 5.0, GeneChip® Mapping 500 K Array and Illumina Infinium whole-genome SNP array. In this article, we will focus on the Affymetrix GeneChip® Mapping 500 K array. But the method we are proposing can be applied to data obtained from other platforms as well.

The GeneChip® Human Mapping 500 K Array Set consists of Nsp and Sty arrays. These two arrays are digested with NspI restriction and StyI restriction enzymes separately, each having the capacity to interrogate ~250 000 SNPs. At the SNP level, each SNP is interrogated by 6–10 probe quartets, each of which is composed of a 25 base pair (bp) perfect match (*PM*) oligonucleotide probe and a mismatch (*MM*) probe for alleles A and B (Rigaill *et al*., 2008). There are 24–40 different 25 bp oligonucleotide probes per SNP.

We first define the log ratios of the *PM* probes A and B for SNP i of the tumor and normal samples as,

$$T_i = \log_2 \frac{\sum_{j=1}^{n_i} T_{(A)ij}}{n_i} - \log_2 \frac{\sum_{j=1}^{n_i} T_{(B)ij}}{n_i} \text{and}$$
$$C_i = \log_2 \frac{\sum_{j=1}^{n_i} C_{(A)ij}}{n_i} - \log_2 \frac{\sum_{j=1}^{n_i} C_{(B)ij}}{n_i}$$

where $T_{(A)ij}$ and $T_{(B)ij}$ ($C_{(A)ij}$ and $C_{(B)ij}$) are the raw intensities of the *PM* probes A and B in the *jth* probe quartet for SNP i of the tumor (paired normal) sample, respectively. As explained in Irizarry *et al*. (2003), we will only use the *PM* probe data in our method. Further, using the *PM* probes only will allow our method to be readily used on newer generations of SNP arrays, where only *PM* probes are included (Carvalho *et al*., 2007).

Genotyping calls for normal samples usually can achieve 99% accuracy (Scharpf *et al*., 2008), thus, it is feasible to accurately identify heterozygous probes in normal samples. Hence, we can define the allelic log ratios (*alog-Ratio*) of the *PM* probes A and B for heterozygous SNP i of the paired tumor and normal samples as:

$$T_i^{(hz)} = \log_2 \frac{\sum_{j=1}^{n_i} T_{(A)ij}^{(hz)}}{n_i} - \log_2 \frac{\sum_{j=1}^{n_i} T_{(B)ij}^{(hz)}}{n_i} \text{and}$$
$$C_i^{(hz)} = \log_2 \frac{\sum_{j=1}^{n_i} C_{(A)ij}^{(hz)}}{n_i} - \log_2 \frac{\sum_{j=1}^{n_i} C_{(B)ij}^{(hz)}}{n_i}$$

respectively. The difference of *alog-ratios* (*dalog-Ratio*) between the paired tumor and normal samples is defined as, $O_i^{(hz)} = T_i^{(hz)} - C_i^{(hz)}$.

Additionally, we define the log-intensities of the paired tumor and normal samples for SNP i as:

$$T_i' = \log_2 \frac{\sum_{j=1}^{n_i}\left(T_{(A)ij} + T_{(B)ij}\right)}{n_i} \text{and} \quad C_i' = \log_2 \frac{\sum_{j=1}^{n_i}\left(C_{(A)ij} + C_{(B)ij}\right)}{n_i}$$

and the *log-Ratio* as: $Y_i = T_i' - C_i'$. The corresponding log-ratio for heterozygous SNP i is denoted as $Y_i^{(hz)}$. Note that, in calculating *alog-Ratio* of the *PM* probes, we first take the average of the raw signal intensities of all quartets of a probe, then take the logarithm, instead of averaging after taking the logarithm of the raw intensities. We explained in the Supplementary Data that the normality assumptions on *dalog-Ratio* and *log-Ratio* hold in both cases (see S1 in Supplementary Data for details). In practice, by averaging the raw signal intensities first, we observed less pronounced saturation effects (data not shown).

In addition, under an ideal situation (Gardina *et al.*, 2008), the relationship between *dalog-Ratio* and *log-Ratio* for heterozygous probes with one-copy gain/loss can be expressed as (see S2 in Supplementary Data for details):

$$E\left(O_{\pm A}^{(hz)}\right) \approx \log_2\left(2^{E\left(Y_{\pm A}^{(hz)}\right)+1} - 1\right),$$

where $E\left(O_{\pm A}^{(hz)}\right)\left(\text{or } E\left(Y_{\pm A}^{(hz)}\right)\right)$ is the expected change(s) in *dalog-Ratio* (or *log-Ratio*) of heterozygous probes with one-copy gain(+)/loss(−) comparing with normal two-copy heterozygous probes. By using *dalog-Ratio* instead of B allele frequency (*BAF*) in our proposed method, we can avoid the unnecessary data manipulation as required for computing *BAF* values; in the meantime, no useful information is lost.

Next, we will present our proposed method, the PAIR algorithm, which consists of two parts. Part I applies a two-state *HMM* model on *dalog-Ratio* values to coarsely identify probes in two-copy state, and then Part II refines the results obtained from Part I by using CBS segmentation.

2.2 The PAIR algorithm—Part I: partition probes into *UMS* and *BMS* using *HMM* model

The *HMM* approach is intrinsically suitable for analyzing data with unobservable (missing) variables—the hidden states. In general, a discrete state *HMM* with continuous output includes the following components (Fridlyand *et al.*, 2004; Rabiner, 1989).

(i) Let K be the number of hidden states (discrete). In this article, we will consider a Markov process with two states ($K=2$), with S_1 being the *UMS*, including all genotypes that have equal number of A and B alleles (i.e. 00, AB, AABB, AAABBB and so forth), and S_2 being the *BMS* including all other genotypes (see S3 in Supplementary Data for details).

(ii) Let π_1 and π_2 be the initial probabilities of the two states in (i), where $\pi_1 + \pi_2 = 1$. It has been shown that for *HMM*, with a continuous distribution output, the initial values for π and the state transition probability matrix T (defined next) can be chosen arbitrarily (Rabiner, 1989). Thus, we can assign, for example, $\pi_1 = 0.8$, and $\pi_2 = 0.2$ as the initial probabilities for the two states, given that the two-copy state in most cases dominates. However, assigning any other moderate values would be acceptable.

(iii) Denote $T_{mn} = P(s_t = S_m | s_{t-1} = S_n), 1 \le m, n \le K$, and the state transition probability matrix $T = \{T_{mn}\}$. We assigned a small value (0.0001) to the probability of changing from *UMS* to *BMS* (that is,$T_{12} = 0.0001$) on the assumption that there are on average a few to around dozens of CNAs in the whole genome of a tumor sample. Consequently, the probability of staying in the *UMS* is 0.9999. Further, to avoid 'over-normalization' of probes with copy number gain/loss, and also to take into account the often large number of reference two-copy state probes used in normalization, we are prone to be conservative in calling a probe in *UMS* versus *BMS*. Thus, we assigned an even smaller value (0.00001) to the probability of changing from *BMS* to *UMS* ($T_{21} = 0.00001$).

Denote $b_1(O)$ and $b_2(O)$ as the emission probabilities for *UMS* and *BMS*, respectively. In our *HMM* model, the observation O is the difference of allelic *log-Ratio* of the heterozygous probes. Further, under the normality assumption of *dalog-Ratio*, $b_1(O)$ is a normal density with $\mu = 0$, and variance σ^2, and $b_2(O)$ is the density of a mixture of two normal distributions (*BMS*) with means μ and $-\mu$, and the same component variance σ^2. Here, we also assume that the two modes of the *BMS* have equal weight because we do not expect an association between allele type and CNAs, and thus allele A is affected as often as allele B (see S3 in Supplementary Data for details). In addition, we assume the variance of $b_1(O)$ and the component variances of $b_2(O)$ are the same for the case of one-copy gain situation for simplification. Note that the component variances of $b_2(O)$ can be larger for other types of CNAs. However, this is not a concern for our model, as in those situations, μ is much larger, and a moderate change in component variance does not affect *UMS* and *BMS* partition (see Supplementary Fig. S1).

Given the *HMM*, we used *Viterbi* algorithm (Rabiner, 1989; Viterbi, 1967) to compute the probability of the most probable state sequence.

The distribution of probes that are in the *BMS* may be complicated. For example, there may be tumors that simultaneously have, say, segments of one-copy gain, one-copy loss, cnLOH and two-copy gain. As a result, a bi-modal distribution would not be sufficient to model it. However, among these gains/losses and cnLOH, one-copy gain is the most difficult one to detect (see S4 in Supplementary Data), except for some minor clone caused by CNAs that might not be detectable at all. Thus benefiting from this inequality, algorithms that detect one-copy gain will automatically detect other types of odd-number gain/loss, cnLOH and certain types of even-number gain/loss.

To detect segments of one-copy gain, we applied a *HMM* with three iterations. In the first iteration, we set the initial values for the variance of $b_1(O)$ and the component variance of $b_2(O)$ in the *HMM* equal to the sample variance (σ_0^2) of the whole sequence of observations (*dalog-Ratio* values). When there is none or only focal gains/losses, this sample variance is supposed to be close to the true variance. Otherwise, when there exist large segments of gain/loss, the value of σ_0^2 will be bigger than the true variance of probes in *UMS*, but smaller than the true mixture distribution variance of probes in *BMS*. As a result, probes with larger variance tend to fit the *BMS* better, and probes with smaller variance tend to fit the *UMS* better. Thus, a meaningful initial partition can be obtained with such choice of the initial values of variances. In the second iteration, we used the estimation of the variance of probes in the *UMS* (detected from iteration 1) as the input in the *HMM* model. And the third iteration is a repeat of the second iteration. Note that, when the *SNR* is high, the two states can be easily separated, and when the *SNR* is low, the power for distinguishing the two states from each other depends more on the means (μ *and* $-\mu$) of the two component distributions. Theoretically, additional iterations may be run until a convergence criterion is met. However, in practice, we observed that two or three iterations are usually sufficient to get a decent normalization result.

We comment that to have large power to identify one-copy gain, it makes sense to set μ as small as possible while controlling false-positive rate. Based on our simulation, by setting μ at ~2 standard deviations of a sequence of random normally distributed observations, we can control the false-positive rate at a 5% level. In practice, we noticed that setting μ close to any value between 1.9 and 2 standard deviations of the whole sequence (for first iteration), or observations in *UMS* (detected by previous iteration), can achieve decent results of separation between *UMS* and *BMS*, even for noisy samples. And minor changes in μ, such as ± 0.05, would not affect the separation result except for some focal gains/losses. This is in concordance with Supplementary Data S7, where we show that

the expected *dalog-Ratio* value of one-copy gain is >1 standard deviation of the two-copy probes if the sample contamination rate is ≤12%. Note that, we made conservative assumptions in Supplementary Data S7. In practice, we were able to obtain good separation of *UMS* and *BMS* for all samples we tested.

Detection of homozygous deletions (HDs) is much easier and can be achieved by applying the same *HMM* model to the *dalog-Ratios* of all probes (both homozygous and heterozygous probes). The variances of HD segments could be as small as 1/5 of the variances of other segments, thus these HDs could be easily identified in one iteration (see Supplementary Fig. S2).

As the final step of Part I, we mapped the probes in *BMS* back to the original whole sequence (including both homozygous and heterozygous probes) and considered the probes within two consecutive *BMS* probes as CNA probes. After excluding these CNAs and HD probes, all remaining probes are called the initial reference set.

2.3 The PAIR algorithm—Part II: detecting probes in two-copy state by applying *CBS* on piecewise average *log-ratio*

Let $\tau_i, i = 1, 2, ..., l,$ be the sequence of points where copy number changes occur. Assume that Y_i is a sequence of independent random variables with normal distribution $Y_i \sim N(\mu_i, \sigma^2)$, where $\tau_i < t \leq \tau_{i+1}, i = 0, 1, 2, \ldots, l,$ and $0 = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = T$. In this article, the variables Y_i represent the logarithm of signal-intensity ratio described in Section 2.1.

Divide the whole sequence of random variables into subsequences of equal size, with m random variables (elements) in each subsequence (the last one might have less than m random variables). We then obtain a new sequence of random variables,

$$M_p = \sum_{j=1+(p-1)m}^{\min(pm,\, T)} Y_j, p = 1, 2, ..., \left\lceil \mathrm{T} \Big/ \mathrm{m} \right\rceil$$

the average of all random variables within each subsequence, and term it the piecewise average log-ratio (*PAlog-Ratio*). Here, M_p was calculated by taking the average of m (= 100) consecutive *log-Ratio* values in the whole sequence, including homozygous ones. Note that the reasons to use $m = 100$ are 2-fold: (i) the computation is much faster and (ii) the bias in segment-mean estimate introduced by using *PAlog-Ratios* instead of raw *log-Ratios* is negligible.

The execution of Part II started with a modality test (Hasselbla, 1966) on *PAlog-Ratios* for all probes in the initial reference set. Based on *BIC* (Schwarz, 1978), if a uni-modal distribution fits well, all probes in the initial reference set would be two-copy probes and will be used as the references in subsequent normalization. Otherwise, if a multiple-modal distribution fits well, the estimated mean *log-Ratio* of the component distribution with the smallest mean would be used as the mean of two-copy probes. Next, after applying *CBS* segmentation to *PAlog-Ratio* values (HD and one-copy loss probes have been excluded by Part I), we removed any segments whose mean *log-Ratio* values are greater than the median of the mean log-ratios of one-copy gain segments (see S5 in the Supplementary Data for the criteria to detect segments of one-copy gain/loss) and used the probes remaining in the initial reference set as the reference set. In cases where only segments of one-copy loss exist, we imputed the one-copy gain value by multiplying the absolute change of one-copy loss by 0.58. If neither a one-copy gain nor a one-copy loss segment was detected in Part I, we simply used a 95% confidence interval about the two-copy mean as the cut-off values.

2.4 Normalization via spline smoothing

After identifying probes in the two-copy state, normalization was carried out using an *M–A* plot, where the log-Ratio (*M*) was plotted against the log mean intensity (*A*). Note that *A* is the average of tumor and the paired normal sample log intensities (Smyth and Speed, 2003). We obtained *M* and *A* values for each probe, and set the median of M values of the probes (all probes for normal sample, and two-copy probes only for tumor sample) as *0*. Then, we used the *M* and *A* values of probes in the two-copy state to create a spline smoothing curve (Chambers *et al.*, 1992). The corrected *M* values for all probes based on the smoothing curve are the normalized *log-Ratio* values, where the median of normalized *M* values of the probes in the two-copy state equals *0*.

2.5 An option for normalizing noisy data

We have assumed that the distributions of the log-intensities of A and B alleles of heterozygous SNPs on the same array are the same. In other words, the log-intensities of probes with genotypes AA, AB and BB are supposed to have the same distribution. However, in practice, they may be somewhat different because of unexpected sources of variation. For example, cnLOH (exclusively AA or BB genotypes) should have the same mean *log-Ratio* as those of normal two-copy segments (including probes with AA, AB and BB genotypes), but minor differences do exist because of the fact that log-intensities of probes with AB genotype can be higher/lower than those of probes with AA or BB genotype. To correct this, especially for arrays with choppy signals, we used spline smoothing to normalize tumor and normal log-intensities separately before normalizing the *log-Ratios*. To normalize log-intensities, we plotted the log intensity of the tumor (or normal) sample versus the log-ratio of A and B alleles of the tumor (or normal) sample.

2.6 PAIR algorithm flow chart

A flow chart outlining our procedure can be seen in Figure 1.

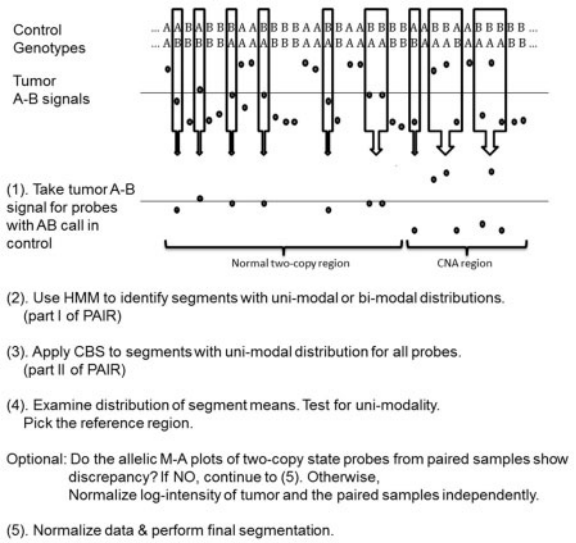

Fig. 1. Flow chart of the PAIR normalization

2.7 Simulation study

To evaluate how the proportion of two-copy probes affects the normalization result, we simulated the SNP array data by re-sampling one-copy gain/loss and two-copy probes from a tumor-control pair, where three normalization methods, CGH-normaliter, PAIR and popLowess, gave similar normalization results. Specifically, we first randomly filled the whole genome with one-copy gain/loss probes, such that all probes in each chromosome (or chromosome arm) have the same copy number—either one-copy gain or one-copy loss. To maintain the correlation

structure, the log-intensities of alleles A and B for the paired tumor and normal samples and the genotype for the normal sample of the same SNP were sampled simultaneously. Then, we randomly replaced some of the one-copy gain/loss segments with the two-copy probes to have a pre-defined proportion of two-copy probes in the genome. That is, we randomly picked a chromosome (or arm) and filled it with two-copy probes and repeated this process until the pre-defined proportion was reached. Finally, we applied all three normalization methods to each simulated data, and calculated the rate of the two-copy probes correctly identified by each method. These rates are presented and compared in Section 3.6.

3 RESULTS

We demonstrated the performance of PAIR algorithm by applying it to a public dataset, where 114 multiple myeloma samples were analyzed by Affymetrix 500K array set (Nsp + Sty). Among them, 80 samples have matched peripheral blood DNA performed on the same array types. These data were generated in Walker *et al.* (2010) and can be downloaded from the public domain GEO with access ID, GSE21349.

3.1 Partitioning heterozygous probes into segments of *UMS* and *BMS*

By Part I of the PAIR algorithm, a two-state *HMM* model was applied to *dalog-Ratio*, and the heterozygous probes were partitioned into segments of *UMS* and *BMS*. The partition result is presented in Figure 2. In Figure 2a, probes in black color are those that follow a uni-modal distribution (the *UMS* segments), and probes in red color are those that follow a bi-modal distribution (the *BMS* segments). It is clearly indicated by the 95% confidence lines (in red) that probes in the *UMS* segments have smaller variance than those in the *BMS* segments do (see S6 in Supplementary Data for details). Note that any two-copy probe, if it exists, has to be within one of *UMS* segments. This is the key to ensure no misclassification of probes in two-copy state or in one-copy gain/loss. On the other hand, a *UMS* segment might include probes with certain types of even-number gain, such as AABB, AAABBB and so forth. This justifies the necessity of Part II of the PAIR algorithm. For example, from Figure 2c—the plot of *log-Ratios* of the whole-genome with *BMS* probes being colored in red, the majority of probes on chromosomes 2–4, 7–12, 14, 16–17 and 19–22 are in *UMS*, but the probes on chromosome 9 are in two-copy gain instead of two-copy state. The explanation for chromosome 9 being partitioned into the *UMS* segment is that the genotypes of heterozygous probes on this chromosome are presumably AABB, instead of AAAB/ABBB.

3.2 Identifying probes in two-copy state from all probes in the initial reference set

By Part II of the PAIR algorithm, we aim to separate probes in two-copy state from the segments of the initial reference set. Based on the fact that the mean *log-Ratio* of probes in a two-copy gain segment is higher than that of probes in a one-copy gain segment, we concluded by Part II that chromosome 9, which was originally partitioned into the initial reference set, was a two-copy gain chromosome. In general, the segments with copy number gain of higher magnitude can be excluded

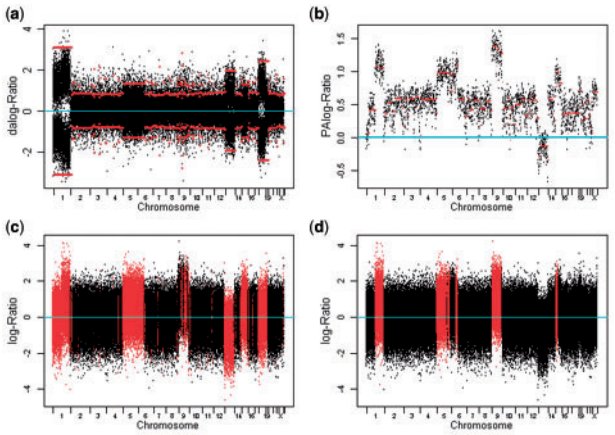

Fig. 2. An illustration of applying PAIR algorithm to one of the paired samples. (**a**) The application of Part I of the PAIR algorithm—segments of *UMS* and *BMS* were outlined by the two standard deviation confidence interval lines of each segment (in red). Probes in segments of *UMS* have the smallest confidence interval (variance), that is, chromosomes 2–4, 7–12, 14, 16, 17 and 19–22. Probes in segments of *BMS* have larger variances of different scales owing to the different nature of gain/loss, that is, probes on chromosome 5 and 15, and part of chromosome 6 has one-copy gain, probes on chromosome 13 have one-copy loss and chromosome 18 and part of chromosome 1 are segments of cnLOH. (**b**) The application of Part II of the PAIR algorithm—CBS segmentation was applied to the *PAlog-Ratio*. (**c**) The reflection of Part I *UMS* and *BMS* partition on log-ratio values. The black and red colors represent segments of *UMS* and *BMS* detected in Part I, respectively. Probes on chromosome 9, which have two-copy gains, were largely in UMS. (**d**) The reflection of applying CBS on log-ratio values. Chromosome 9 was re-identified as a segment of two-copy gain by CBS segmentation. Chromosome 18 and part of chromosome 1, both of which were cnLOH, were not identified as CNAs

from the reference set by Part II because their mean *log-Ratios* are higher than those of one-copy gain segments. After applying Part II, segments remaining in the reference set predominately consist of two-copy probes. By subtracting the median of *log-Ratios* of these two-copy probes from the raw *log-Ratios* of all probes, we obtained the centralized *log-Ratio* values.

We comment that (i) it may be difficult to accurately identify CNAs by conventional segmentation methods alone. For example, chromosome 18 is in cnLOH (Fig. 2a), but it cannot be detected by conventional methods (Fig. 2c) because cnLOH is also two-copy. (ii) Although chromosomes 2–4, 16 and 17 are all two-copy chromosomes, there is a drift in segment mean along chromosomes (Fig. 2b). As a result, the segment means of *PAlog-Ratio* values of probes on chromosomes 16 and 17 are lower than those of probes on chromosomes 2–4 (the reference cyan line represents 0). This may result in assigning incorrect copy numbers to those chromosomes. In contrast, the *dalog-Ratio* plot (Fig. 2a) is symmetric about 0 for all the chromosomes, and the variations of *dalog-Ratio* values of probes on chromosomes 2–4, 16 and 17 are among the smallest, indicating that all these chromosomes are in *UMS*, and potentially in two-copy state.

We applied Part I of the PAIR algorithm to achieve two goals: (i) to identify segments that contain probes in two-copy state; (ii) to detect segments of cnLOH, such as chromosome 18 and

part of chromosome 1 in the aforementioned example. In the extreme cases where there are no two-copy probes in the whole genome, some additional work is needed. A brief discussion can be found in Section 4. Based on the fact that two-copy heterozygous probes follow a uni-modal normal distribution, by partitioning probes into *UMS* and *BMS*, we can confidently exclude probes in *BMS* from the reference set. On the other hand, conventional segmentation methods, such as *CBS*, were neither designed to accurately assign the true underlying copy number to segments that have the same copy number, nor to detect segments of cnLOH. For example, the mean *log-Ratio* values of probes on chromosome 18 and a part of chromosome 1 are close to those of probes in normal two-copy segments, thus it is difficult to distinguish them from the normal two-copy probes. Nevertheless, *CBS* (adopted in Part II) has greater power for detecting higher-magnitude change of mean *log-Ratio* value. In the aforementioned example, *CBS* provided satisfactory result for detecting two-copy gains of chromosome 9.

3.3 Normalization by an M–A plot

The result of normalization (by an M–A plot) is presented in Supplementary Figure S3. As a quicker alternative to Lowess regression, spline smoothing was used in computing the correction curve. The advantage of using spline function over polynomial regression is to avoid distortion, especially at the left and right tails (Chen *et al*., 2008).

3.4 Comparing with CGHnormaliter, popLowess normalization methods and others

The *R*/Bioconductor package 'CGHnormaliter' and a standalone version of *R* code for popLowess were used to normalize the data for the purpose of comparison. Normalized data were then segmented using *CBS* to visualize the results. In many cases, three algorithms, CGHnormaliter, PAIR and popLowess, generated almost the same normalized data. However, when the proportion of two-copy probes is not predominantly high, both CGHnormaliter and popLowess may have difficulties in correctly centralizing the normalized data. Using the sample with ID 303 as an example, our method (Fig. 3b) centralized the data around the mean *log-Ratio* values of chromosomes 2, 3, 5, 7, 9, 15, 19 and 20. CGHnormaliter (Fig. 3c), however, centralized the data around the mean *log-Ratio* values of chromosomes 1, 4, 6, 8, 10, 12–14, 16, 21 and 22. We highlighted those probes that have AB genotype calls in normal, but AA or BB in tumor (see the probes in red color in Figure 3; probes that do not have this change are in black color). As most of the probes that have heterozygosity loss are located in chromosomes 1, 4, 6, 8, 10, 12–14, 16, 21 and 22, we assume that these chromosomes are the one-copy loss chromosomes.

PopLowess (Fig. 3d) was able to centralize the data similarly to our method, the mean log-Ratio values of the reference chromosomes, especially chromosomes 9, 15, 19 and 20, were slightly different from 0. The possible reason is that the k-mean cluster approach used in popLowess does not provide precise estimates of cluster means.

In another sample, CGHnormaliter and PAIR obtained similar normalization results, but popLowess picked different chromosomes as the reference (see Supplementary Fig. S4).

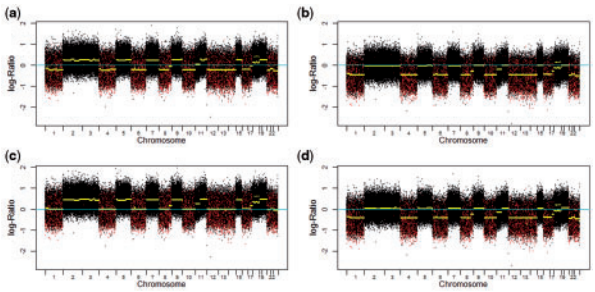

Fig. 3. The comparison of normalization results from CGHnormaliter, PAIR and popLowess algorithms. CBS segmentation profile was superimposed in yellow color. Probes in red color are those with AB genotype in the paired normal, but AA or BB in tumor. (**a**) Applying CBS segmentation on the raw log-ratio. The total number of segments was 70, and the mean log-ratio of neither of the two dominating copy number states was close to 0. (**b**) The PAIR normalized data. The total number of segments was 40, and the mean log-ratio values of the segments in two-copy state, that is, chromosomes 2, 3, 5, 7, 9, 15, 19 and 20, were ~0. (**c**) The CGHnormaliter normalized data. The total number of segments was 65, the mean log-ratio values of chromosomes, that is, 1, 4, 6, 8, 10, 12–14, 16, 21 and 22, were ~0. (**d**) The popLowess normalized data. The total number of segments was 50; however, the horizontal cyan line, which indicates no CNAs, was slightly different from the mean log-ratio values of chromosomes, that is, 9, 11, 15, 19 and 20, indicating there might be some problems with the centralization process

We also compared *PAIR* to other methods, such as *Quantile* (Bolstad *et al*., 2003), *Invariant Set* (Li and Wong, 2001) and *ITALICS* (see Supplementary Figure S5). The figure indicates that *PAIR* is the method that is able to set '0' at approximately the median of a copy number population.

3.5 Optional normalization of tumor and normal log-intensity separately

The allelic M–A plot of probes in two-copy state is theoretically centered about a horizontal line, as the probes with AA, AB and BB genotypes are supposed to have the same log-intensity values. However, the real data often show that the pattern of log-intensity values deviates from a horizontal line, the superimposed spline smoothing white lines in Figure 4a for a tumor sample and c for the paired normal sample. Using paired samples could normally neutralize the effect of such differences if the pattern for the tumor sample is coincident with that of the paired normal sample. Nevertheless, should the patterns for the paired samples be more divergent, optionally normalizing the log-intensities of tumor (based on the PAIR identified two-copy probes) and the paired sample separately before normalizing *log-Ratios* might be necessary to improve the quality of normalization.

The effects of the optional normalizations are presented in Figure 5, where we plotted the PAIR normalized *log-Ratio* with or without the optional individual sample normalization in a or b. Figure 5c and d are the normalization results after applying CGHnormaliter and popLowess. Comparing a and b, we can see that applying the optional normalization resulted in a large reduction in the total number of segments detected by *CBS*. The reduction was also substantial if comparing with CGHnormaliter and popLowess normalization results.

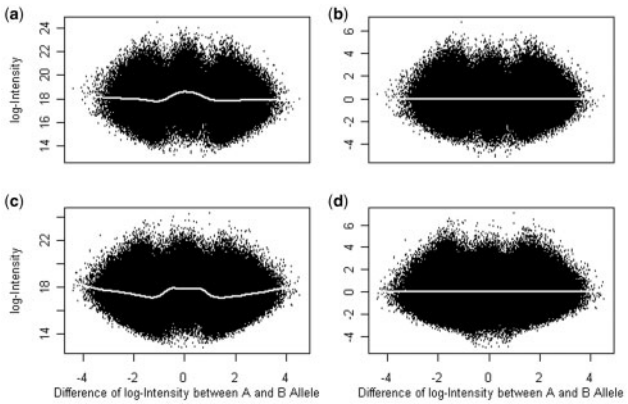

Fig. 4. The M–A plots for the allelic log-intensities of the paired samples—M is the log sum intensity of the two alleles, and A is the difference of the log-intensities of the two alleles. (**a**) Tumor sample (probes in two-copy state only) before log-intensity normalization. (**b**) Tumor sample, after log-intensity normalization based on PAIR identified two-copy probes. (**c**) The paired normal sample, before log-intensity normalization. (**d**) The paired normal sample, after log-intensity normalization

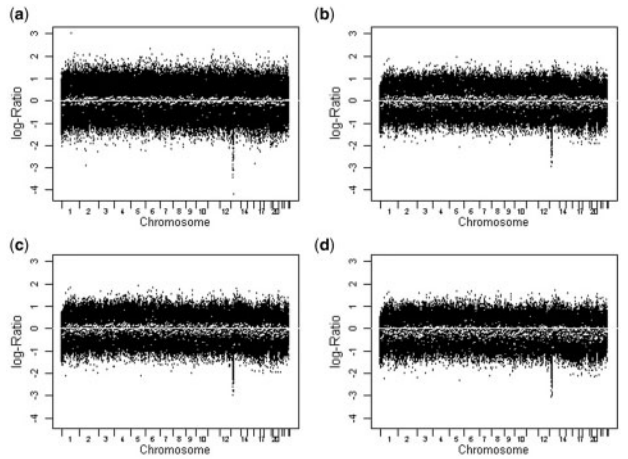

Fig. 5. The comparison of normalizations with/without the optional individual sample log-intensity normalization, as well as CGHnormaliter and popLowess normalization. CBS segmentation profile was superimposed in white color. (**a**) The PAIR normalized log-Ratio with the optional individual sample normalization. (**b**) The PAIR normalized log-ratio without the optional individual sample normalization. (**c**) The CGHnormaliter normalized log-ratio. (**d**) The popLowess normalized log-ratio

Furthermore, the white band (of mean *log-Ratios*) in Figure 5a is narrower than those in b, c and d, indicating that optional normalization may also be able to reduce the signal noise.

Furthermore, the effect of the optional individual sample log-intensity normalization could also be seen from the change in the mean log-Ratio of segments of cnLOH. We compared calculating results by different methods (see S1 in Supplementary Data) in Figure 6. For Figure 6a and b, the intensities of A/B alleles were calculated by taking the averages of the raw A/B allele intensities of all quartets for a probe, and

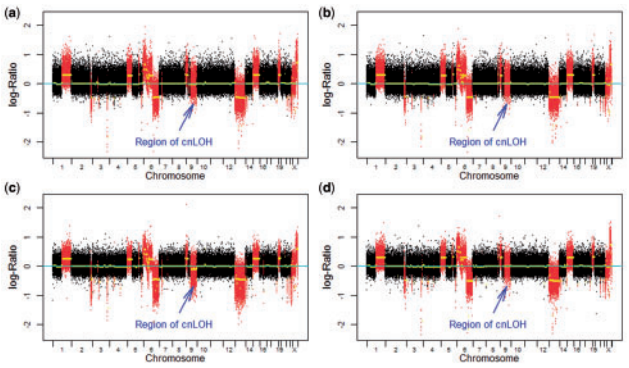

Fig. 6. The effect of the optional individual sample log-intensity normalization on the segment of cnLOH. The CBS segmentation profile was superimposed in yellow. (**a**) The segmentation result of PAIR normalized data. (**b**) The segmentation result of PAIR normalized data after applying the optional individual sample log-intensity normalization. (**c**) The segmentation result of PAIR normalized data. (**d**) The segmentation result of PAIR normalized data after applying the optional individual sample log-intensity normalization

Table 1. The numbers of two-copy probes identified by three algorithms

2n probe (%)	PAIR			CGHnormaliter			popLowess		
	F	T	pct.	F	T	pct.	F	T	pct.
5	30	162	84	192	0	0	192	0	0
10	0	225	100	221	4	2	225	0	0
20	0	205	100	199	6	3	203	2	1
30	0	140	100	97	43	31	75	65	46
40	0	200	100	14	186	93	0	200	100
50	0	219	100	0	219	100	0	219	100
60	0	102	100	0	102	100	0	102	100
70	0	75	100	0	75	100	0	75	100
80	0	50	100	0	30	100	0	30	100

Note: Simulations which caused program crash were not included.
F, false; T, true; pct, percentage of correctly identified probes.

then the total log-intensity was the logarithm of the sum of A and B allele intensities. For c and d, the log-intensities of A/B alleles were calculated by taking the average of logarithms of the raw A/B allele intensities of all quartets for a probe, and the total log-intensity was the sum of the A and B allele log-intensities. Figure 6a showed that the mean log-ratio of the cnLOH segment was slightly different from those of the normal two-copy segments before individual normalization. The optional individual normalization mitigated this difference (Fig. 6b). This effect was much striking when comparing d and c. Note that for samples that performed the optional normalization, segments of cnLOH could be included in the reference set.

3.6 Simulation result

In Table 1, we present the simulation results of the numbers of two-copy probes correctly and incorrectly identified by the

methods. All three methods perform equally well when the proportion of two-copy probes is >50%. However, when this proportion drops, neither CGHnormaliter nor popLowess provides satisfactory normalization. The main reason is that the assumption of these two methods is violated. In contrast, PAIR can provide satisfactory normalization even when the proportion of two-copy probes is as low as 10%.

4 DISCUSSION

Correctly identifying reference probes in two-copy state from the SNP array is not trivial. We propose an algorithm, PAIR, that can accurately distinguish two-copy state probes from those with CNAs.

The proposed PAIR algorithm first partitions the whole genome into segments of *UMS* or *BMS*. The normal two-copy probes, if they exist, can then be narrowed down to the segments of *UMS*. This algorithm does not assume that two-copy probes dominate the whole genome. This is an advantage compared with the methods that require this assumption.

In contrast to the conventional methods, PAIR has the advantage of being able to identify segments of cnLOH. This is preferable, as the mean log-ratio values of probes in some cnLOH segments can be different from those in normal two-copy segments, and this could cause a bias in the step of centralization should the cnLOH segment be large. Furthermore, with the optional allelic normalization, we are able to equalize the log-intensities of A and B alleles, and thus remove one source of false discovery.

We did not integrate *log-Ratio* and *dalog-R* analyses into the *HMM* model, but instead, we adopted a two-step process. In Part II, we applied *CBS* segmentation to *PAlog-Ratios* instead of the raw *log-Ratios*. By doing so, we could substantially reduce the computational time. Meanwhile, comparing with the methods using *log-Ratio*, the risk of false detection can be reduced if signal drift and uncorrected genomic wave exist (Diskin *et al*., 2008).

We obtained the monotone relationship between *log-Ratio* and *dalog-Ratio*, and proposed to use *dalog-Ratio* instead of *BAF* in cnLOH detection and underlying copy number estimation to avoid unnecessary data manipulation.

The proposed method reduced the computation time when comparing with CGHnormaliter. For the samples with a large number of segments, it takes only 3–5 min to complete the normalization process using our method, but usually 20–30 min using CGHnormaliter, when the algorithms were run on Lenovo T400 laptop.

CNAs are different from sample to sample, and sometimes can be complicated. It may be difficult to fully specify all CNA forms by a *HMM* model. Bearing this in mind, the proposed two-state model, although it might miss some focal CNAs, can greatly reduce the risk of misclassification of large segments of CNAs. This, in our opinion, is sufficient for a normalization method to be acceptable.

One of the assumptions of the PAIR algorithm is that there are at least some normal two-copy heterozygous probes in the genome that need not to be dominated. The simulation result shows that the PAIR algorithm has almost 100% power for correctly detecting the normal two-copy states when the proportion of two-copy state probes are as low as 10%. In the extreme cases where virtually no two-copy probes exist, that is, the whole genome is in the same state, we could still determine whether the whole genome is in one-copy gain/loss from two-copy. It will be more challenging to accurately centralize the probes. However, this is not the focus of this article.

ACKNOWLEDGEMENTS

The authors are thankful to Ms Alice LeBlanc for her help in editing the article, which greatly improves the presentation of the article, and to Dr Cheng Cheng for valuable comments and discussions. We also thank the Associate Editor and anonymous reviewers for their insightful comments.

Funding: NIH (NIMHD-RCMI 8G12MD007595-04); US Department of Army (W911NF-12-1-0066); NSF (EPS-1006891); Louisiana Cancer Research Consortium (to K.Z.). National Institute of General Medical Sciences of the National Institutes of Health (1 U54 GM104940 to S.Y. and Z. F.) which funds the Louisiana Clinical and Translational Science Center of Pennington Biomedical Research Center.

Conflict of Interest: none declared.

REFERENCES

Bengtsson,H. *et al.* (2010) TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, **11**, 245.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Carvalho,B. *et al.* (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.

Chambers,JM. *et al.* (1992) *Statistical Models in S*. Wadsworth & Brooks/Cole, CA.

Chen,H.I. *et al.* (2008) A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics*, **24**, 1749–1756.

Curtis,C. *et al.* (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*, **10**, 588.

De Vita,V.T. *et al.* (2008) A history of cancer chemotherapy. *Cancer Res.*, **68**, 8643–8653.

Diskin,S.J. *et al.* (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, e126.

Fanciulli,M. *et al.* (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.

Fridlyand,J. *et al.* (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

Gardina,P.J. *et al.* (2008) Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. *BMC Genomics*, **9**, 489–505.

Hasselbla,V. (1966) Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.

Hupé,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA segments. *Bioinformatics*, **20**, 3413–3422.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Li,C. and Wong,H.W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, **2,** RESEARCH0032.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

McCarroll,S.A. *et al.* (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.

Mullighan,C.G. (2009) Genomic analysis of acute leukemia. *Int. J. Lab. Hematol.*, **31**, 384–397.

Mullighan,C.G. *et al.* (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukemia. *Nature*, **446**, 758–764.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pounds,S. *et al.* (2009) Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics*, **25**, 315–321.

Przybytkowski,E. *et al.* (2011) The use of ultra-dense array CGH analysis for the discovery of micro-copy number alterations and gene fusions in the cancer genome. *BMC Med. Genomics*, **4**, 16.

Rabiner,L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–287.

Rigaill,G. *et al.* (2008) ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. *Bioinformatics*, **24**, 768–774.

Scharpf,R.B. *et al.* (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Stat.*, **2**, 687–713.

Staaf,J. *et al.* (2007) Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics*, **8**, 382.

Staaf,J. *et al.* (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.*, **9**, R136.

Smyth,G.K. and Speed,T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

van Houte,B.P.P. *et al.* (2009) CGHnormaliter: an iterative strategy to enhance normalization of array CGH data with imbalanced aberrations. *BMC Genomics*, **10**, 401.

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.

Walker,B.A. *et al.* (2010) A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, **116**, e56–e65.

Yang,Y. *et al.* (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.*, **80**, 1037–1054.

Ylstra,B. *et al.* (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res.*, **34**, 445–450.