BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 15 2010, pages 1827–1833
doi:10.1093/bioinformatics/btq300

Genome analysis

Advance Access publication June 6, 2010

ACNE: a summarization method to estimate allele-specific copy numbers for Affymetrix SNP arrays

Maria Ortiz-Estevez[1], Henrik Bengtsson[2] and Angel Rubio[1,*]

[1]Group of Bioinformatics, CEIT and TECNUN, University of Navarra, San Sebastian, Spain and [2]Department of Statistics, University of California, Berkeley, CA, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Current algorithms for estimating DNA copy numbers (CNs) borrow concepts from gene expression analysis methods. However, single nucleotide polymorphism (SNP) arrays have special characteristics that, if taken into account, can improve the overall performance. For example, cross hybridization between alleles occurs in SNP probe pairs. In addition, most of the current CN methods are focused on total CNs, while it has been shown that allele-specific CNs are of paramount importance for some studies. Therefore, we have developed a summarization method that estimates high-quality allele-specific CNs.

Results: The proposed method estimates the allele-specific DNA CNs for all Affymetrix SNP arrays dealing directly with the cross hybridization between probes within SNP probesets. This algorithm outperforms (or at least it performs as well as) other state-of-the-art algorithms for computing DNA CNs. It better discerns an aberration from a normal state and it also gives more precise allele-specific CNs.

Availability: The method is available in the open-source R package ACNE, which also includes an add on to the aroma.affymetrix framework (http://www.aroma-project.org/).

Contact: arubio@ceit.es

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

Received on February 25, 2010; revised on May 13, 2010; accepted on June 3, 2010

1 INTRODUCTION

Genomic aberrations are involved in the pathogenesis of different diseases, especially cancer (Pinkel *et al.*, 1998; Pollack *et al.*, 1999). These genomic abnormalities may contribute to a cell becoming cancerous, especially if tumor suppressor genes or oncogenes are affected.

DNA copy number aberrations (CNAs) are genomic aberrations that involve amplifications or deletions of a part of the genome (a whole chromosome, an arm or a segment). The DNA copy numbers (CNs) in a CNA may be larger (amplifications) or smaller (deletions and homozygous deletions) than the normal state (CN equals to two). Loss of heterozygosity (LOH) is another genomic aberration that occurs in a segment of the genome that lacks one of the two parental chromosomes. For each SNP in an LOH zone, one of the alleles has CN equals to zero. CN for a zone with LOH may be one (loss of one of the parental chromosomes), two (copy neutral LOH) or even larger than two. In the latter two cases, the additional copies of the region with LOH have been regenerated by the cell using the present copy.

The single nucleotide polymorphism (SNP) arrays can be used to study the CNAs and also the LOH cases. Although the initial application of these arrays was genotyping, they can also be utilized for estimating the number of copies at the locus where the interrogated SNP is located. There are also genomic aberrations that do not affect total CNs or genotypes, e.g. translocations and reversions. Such aberrations cannot be identified using SNP arrays.

Several companies have developed this type of arrays but we will focus in the Affymetrix (Affymetrix Inc., 2009) platform. Affymetrix has a family of SNP arrays starting from the 10K arrays, which interrogates 10 000 SNPs. Afterward they developed the 100K and 500K chip sets, both using two different restriction enzymes. The more recent genome wide human SNP array 5.0 (GWS5) and genome wide human SNP array 6.0 (GWS6) arrays also use two enzymes, but they differ from their predecessors in various characteristics. The GWS arrays have non-polymorphic probes in addition to SNP probes, which are used for studying the DNA CN variations (CNVs) and cover parts of the genome where there are no SNPs. The GWS5 has the same number of SNPs as the 500K as well as 420 000 non-polymorphic probes, while GWS6 has around 900 000 SNPs and a bit less than 950 000 non-polymorphic probes. The probes in each chip are grouped in probesets corresponding to single SNPs. Probes complementary to both alleles (A and B) in the same SNP are known as probe pairs. There are probes in the array that are complementary to both strands of the SNP. The length of each probe is 25 nt. The probesets from the 10–500K arrays are formed by perfect match (PM) and mismatch probes (MMs), where the MMs differ from the PMs in the central nucleotide. To give space for more PMs, the GWS arrays do not have MMs.

Several processing methods have been proposed to obtain the DNA CN information from SNP arrays. The first algorithms used were directly inherited from gene expression analysis, as dChip (Li and Hung Wong, 2001) and RMA (Irizarry *et al.*, 2003). There also exist more recent ones that are specific to the CN data, e.g. CNAG (Nannya *et al.*, 2005), PLASQ (LaFramboise *et al.*, 2007), Italics (Rigaill *et al.*, 2008), CRMA (Bengtsson *et al.*, 2008b), CN5 (Affymetrix Inc., 2008) and CRMA v2 (Bengtsson *et al.*, 2009).

The whole pipeline to treat SNP arrays consist of several low-level methods, namely background removal, normalization (to equalize the signal levels for several arrays), summarization (to extract a signal proportional to the CN of each of the alleles), genomic post-processing (to deal with signal bias related to size and other

*To whom correspondence should be addressed.

properties of the sequences where the probes are located) and segmentation (to identify contiguous regions of the genome with the same aberration). Here, we focus on the probe-summarization step of a low-level analysis of SNP arrays and present allele-specific CN estimation (ACNE). ACNE is a multi-sample summarization method based on non-negative matrix factorization (NMF), which directly deals with the cross hybridization problem, giving as result the allele-specific CNs (ASCNs). ACNE uses the data given by CRMA v2 after background and offset removal and can be used in substitution of the probe-summarization models already implemented in CRMA v2.

Using SNP arrays it is possible to get ASCNs, i.e. the number of copies of each allele for any SNP. ASCN is, on its own, biologically relevant for some diseases (Duffy *et al.*, 2008). ASCN is also important to identify regions that present LOH. LOH has shown to be important since it helps to identify aberrant regions of the genome that inactivates tumor suppressors. As will be shown, ASCNs also help the interpretation of CN plots when there is some contamination of normal tissue, i.e. tumor is not 100% pure.

Some of the aforementioned methods provide ASCNs (a brief description of the summarization methods from dChip and CRMA v2 is included in the Supplementary Material). However, results are not as precise as total CNs. The main reason for this poorer performance is crosstalk between the signals for each allele. In turn, this occurs because of cross hybridization. Although cross hybridization (signal added to one probe due to the hybridization with a sequence of the genome different from its target) occurs on all chip types regardless of where the underlying target comes from, it is different depending on the type of array. In SNP arrays, there exist cross hybridization between alleles, since the probes within a probe pair are almost identical (they differ only in the nucleotide that corresponds to the SNP). Special constraints between the two allele signals can be used to calculate and control for this particular crosstalk. Similar constraints are not possible with, for example, expression arrays. CRMA v2 performs a global crosstalk calibration. Here, we propose to correct for cross hybridization also on a per SNP basis. This approach, as shown in Section 3, provides more reliable estimates of total and allele-specific CNs.

2 METHODS

As already stated, the MMs are not included in the latest Affymetrix SNP arrays and they will not be taken into account here. Therefore, we only refer to the PMs when dealing with probesets.

2.1 Data

The data here used (CEL files) come from the International HapMap Project and two different publicly available datasets. The first dataset is formed by 31 samples from the international HapMap Project (Altshuler *et al.*, 2005; The International HapMap Consortium, 2003) hybridized on Mapping250K_Nsp (listed in the Supplementary Material).

The second dataset is formed by a subgroup of 24 tumor samples of a breast cancer study (Haverty *et al.*, 2008) (the data is deposited in NCBI-GEO under accession numbers GSM182833-44, GSM182848-59 and GSE7545). These arrays were hybridized on Mapping250K_Nsp.

Finally, a third dataset is from a study on prostate cancer and it has 74 tumor samples hybridized on GWS6 (Liu *et al.*, 2009) (GSM374529-44, GSM374552-609 and GSE14996).

The initial low-level steps (background and offset removal) are performed using the methods from CRMA v2 (using default parameters). They consist of a calibration for offset and global crosstalk between alleles, and a normalization for probe-sequence effects.

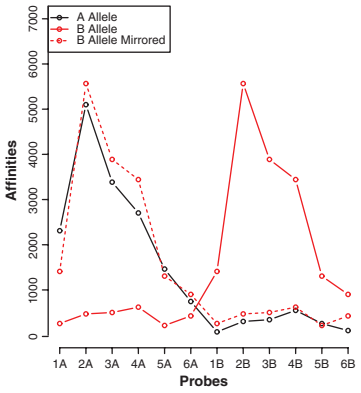

Fig. 1. The affinities of the 12 probes of the SNP_A-1807167 of the breast cancer study. To obtain this affinity matrix all the samples from the dataset are used. The probes have been sorted so that the first half targets A allele (they are PM_A) and the other half targets B allele (PM_B). Moreover, they are also sorted so that the first PM_A (probe 1A) belongs to the same probe pair as the first PM_B (probe 1B). It can be seen that the PM_A probes have higher affinity with A allele, although they also give signal with B allele due to cross hybridization. The target-specific affinities (affinities of A probes for A allele and B probes for B allele) are quite similar to each other as shown by the dotted line. This line represents the affinity of B allele with the position of the probes from the same probe pair mirrored. This is a general trend but there are also SNPs that show probes whose affinities do not behave similarly. Data is for the Affymetrix Mapping250K_Nsp platform.

2.2 Summarization model

Once the data have been preprocessed using CRMA v2, some of the original data y_{ki} (probe intensity of probe k in sample i) can be negative because of offset and cross allelic calibration. In these cases, these data are truncated to a small positive number. Using these data, ACNE summarizes the probes providing two values for each SNP, the estimated CNs for each allele. For this, we propose a linear model in which observed probe intensities are modeled as linear combinations of the true allele-specific CNs. Considering a given SNP, we propose that the observed probe intensity y_{ki} for probe $k=1,\ldots,K$ and sample $i=1,\ldots,I$ of that SNP can be modeled as

$$y_{ki} = \phi_{kA} C_{Ai} + \phi_{kB} C_{Bi} + \varepsilon_{ki}, \tag{1}$$

where (C_{Ai}, C_{Bi}) are the true allele-specific CNs and ε_{ki} a probe-specific random error. The (ϕ_{kA}, ϕ_{kB}) are allele- and probe-specific affinities. For example, for a probe that is complementary to the A allele, we expect its affinity ϕ_{kA} to be larger than ϕ_{kB} and vice versa for a probe complementary to B allele. Ideally, ϕ_{kB} would be zero, but it is not because the alleles cross hybridize to each other. The model in Equation (1) resembles Li-Wong's multiplicative model (Li and Hung Wong, 2001) available in dChip and elsewhere. However, in ACNE we consider two affinities taking into account the cross hybridization between alleles.

As stated before, the probes of a SNP's probe pair are nearly identical differing only in the nucleotide at the SNP locus. Therefore, if probes p and r correspond to the same probe pair, then ϕ_{pA} will be close to ϕ_{rB}. Figure 1 illustrates how probes that belong to the same probe pair have similar affinities.

The matrix representation of Equation (1) is

$$\mathbf{Y} = \begin{bmatrix} \phi_{1A} & \phi_{1B} \\ \phi_{2A} & \phi_{2B} \\ \cdots & \cdots \\ \phi_{KA} & \phi_{KB} \end{bmatrix} \begin{bmatrix} C_{A1} & C_{A2} & \cdots & C_{AI} \\ C_{B1} & C_{B2} & \cdots & C_{BI} \end{bmatrix} + \mathbf{E}, \tag{2}$$

where $\mathbf{E}$ is the error matrix, and $\mathbf{E}$ and $\mathbf{Y}$ matrices are of dimension $K \times I$. Equation (2) suggests an approximated factorization of the intensity matrix ($\mathbf{Y}$), for which both factors are unknown. The entries of both factors must be non-negative, because affinities and CNs are non-negative entities. This characteristic lets Equation (2) to be estimated using NMF (Lee and Seung, 1999) techniques, where the internal factorization dimension is 2.

More precisely, NMF is a group of algorithms where a matrix $\mathbf{Y}$ is factorized into two matrices, $\mathbf{\Phi}$ and $\mathbf{C}$, as

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{C} + \mathbf{E}. \tag{3}$$

NMF enforces the constraint that all the entries of the factors $\mathbf{\Phi}$ and $\mathbf{C}$ must be non-negative. In our case, $\mathbf{\Phi}$ is the *allele-specific affinity matrix* with dimensions $K \times 2$ and $\mathbf{C}$ is the *ASCN* matrix with dimensions $2 \times I$.

2.2.1 Initialization of $\mathbf{\Phi}$ *and* $\mathbf{C}$ NMF factorization can be stated as the following optimization problem

$$\begin{array}{l} \min_{\mathbf{\Phi},\mathbf{C}} ||\mathbf{Y} - \mathbf{\Phi}\mathbf{C}|| \\ \text{subject to} \\ \phi_{ku} \geq 0 \\ c_{ui} \geq 0 \end{array}, \tag{4}$$

where $||\mathbf{Y} - \mathbf{\Phi}\mathbf{C}||$ is a norm of the factorization error, e.g. the Froebenius norm with $u = A, B$ corresponding to the two alleles. The Froebenius norm has been chosen in detriment of the Kullback–Leibler (KL) divergence because, in this particular case, the algorithm to minimize the the KL divergence takes a longer time to obtain the same results. Different authors (Lee and Seung, 1999; Zdunek and Cichocki, 2008) suggest iterative algorithms to solve this problem. Since it is a non-convex optimization problem, a careful selection of the initial estimate is needed in order to avoid local minima.

Initial ASCNs. Our algorithm starts with an initialization of the $\mathbf{C}$ matrix based on a naive estimation of the genotypes. The CN matrix is calculated using the intensity matrix as argument and assuming that $C_{Ai} + C_{Bi} = 2$, i.e. assuming that total CN is 2. Then, for each SNP, if the intensities of the A allele probes are higher than 2 times the intensities of the ones of B allele for a majority of probe pairs, then the SNP will be assigned $C_{Ai} = 2$ and $C_{Bi} = 0$ (genotype AA) and vice versa for genotype BB. Otherwise, the SNP is assigned $C_{Ai} = C_{Bi} = 1$ (genotype AB). The parameter of *2 times the signal of the other allele* was chosen based on empirical results using available HapMap genotype data. This setting provides the smallest percentage of genotype error calls when comparing the predicted genotypes with validated HapMap genotypes.

Initialization of probe affinities. After the initialization of $\mathbf{C}$, we derive a robust initialization for $\mathbf{\Phi}$. The procedure is as follows: choosing two different samples that have different allele-specific CNs, we solve the system of equations for $\mathbf{\Phi}$. The same process is repeated a number of times using different samples. At the end, the median of the computed $\mathbf{\Phi}$s is calculated. We have found that using 50 such random pairs, is a good trade off between computing time and accuracy, while providing a good initial estimate of $\mathbf{\Phi}$. Using the median helps to withstand the presence of outliers in the data (probes that give a wrong signal) and samples that do not have neutral CN as initially assumed.

Dealing with exceptions to total CNs equal to 2. In the above initialization step, the total CNs are assumed to be equal to 2. Using a similar procedure to the initialization of probe affinities and changing the role of the different samples with different probes, we have refined the initialization of the CN matrix.

The results of these robust initializations are $\mathbf{\Phi}^{(0)}$ and $\mathbf{C}^{(0)}$. This initialization step is stochastic so it could give slightly different results for every run. However, in the implementation, the random seed is fixed so that the results are numerically deterministic. Supplementary Figure S3 shows that the accuracy of the solutions does not change if the seed is changed. Standard initialization procedures using uniform positive random numbers do not work well in this particular application. Supplementary Figure S4 shows the quality of the initial values of affinities and CNs obtained in this section.

2.2.2 Pruning of outliers Sometimes outliers appear in the intensity matrix, which NMF algorithms typically are sensitive to. To identify and control for outliers, an error matrix is calculated as the difference between the initial intensity matrix and an estimated initial intensity matrix equal to $\hat{\mathbf{Y}}^{(0)} = \mathbf{\Phi}^{(0)}\mathbf{C}^{(0)}$. Afterward, the SD across samples for each probe is estimated robustly using the median absolute deviation estimator. Any probe intensity error that is 10 times (roughly six SDs for a normal distribution) larger than the across-sample SD is considered an outlier. Such outliers are assigned the corresponding value of $\hat{\mathbf{Y}}^{(0)}$.

2.2.3 NMF optimization The NMF optimization consists of a number of iterations where the CNs and the affinity matrices are recalculated. There exists a number of algorithms for estimating $\mathbf{\Phi}$ and $\mathbf{C}$ (Lee and Seung, 1999; Zdunek and Cichocki, 2008). We have found that the 'projected least squares' method (Zdunek and Cichocki, 2008) works well in this particular case. For each iteration $n = 1, \ldots, N$ of this algorithm, $\mathbf{C}^{(n)}$ is assumed to be an accurate estimate of the true CN matrix. An estimate of $\mathbf{\Phi}$ is then

$$\mathbf{\Phi}^{(n+1)} = \max(\mathbf{Y}\mathbf{C}^{(n)^+}, 0), \tag{5}$$

where $\mathbf{C}^{(n)^+}$ represents the pseudoinverse of $\mathbf{C}^{(n)}$. Next, the same assumption is applied to the $\mathbf{\Phi}^{(n+1)}$ matrix and an updated estimate of $\mathbf{C}$ is

$$\mathbf{C}^{(n+1)} = \max(\mathbf{\Phi}^{(n+1)^+}\mathbf{Y}, 0). \tag{6}$$

In each iteration, $\mathbf{\Phi}^{(n+1)}$ is normalized to have balanced allele-specific affinities, and the sum of the columns of $\mathbf{C}^{(n+1)}$ to be close to 2, as shown in Steps 6 and 8 in the algorithm shown below.

We consider the estimates to have converged when the maximum of the absolute differences between two consecutive matrices, $|\mathbf{C}^{(n+1)} - \mathbf{C}^{(n)}|$, is smaller than ϵ (defaults to 0.01). In rare cases ($\sim$1%), where the algorithm does not converge, we abort the algorithm after $n_{\max}$ (defaults to 10) iterations. The pseudocode of the above NMF optimization is

1. $n \leftarrow -1$
2. **repeat**
3. $\quad n \leftarrow n+1$
4. $\quad \mathbf{\Phi}^{(n+1)} \leftarrow \max(\mathbf{Y}\mathbf{C}^{(n)^+}, 0)$
5. $\quad \mathbf{C}^{(n+1)} \leftarrow \max(\mathbf{\Phi}^{(n+1)^+}\mathbf{Y}, 0)$
6. $\quad$ Normalize $\mathbf{\Phi}^{(n+1)}$ so that the sum of its columns are identical
7. $\quad$ Recalculate $\mathbf{C}^{(n+1)}$
8. $\quad$ Scale $\mathbf{C}^{(n+1)}$ so that the median of the column sums is 2
9. $\quad$ Recalculate $\mathbf{\Phi}^{(n+1)}$
10. **until** $(\max(|\mathbf{C}^{(n+1)} - \mathbf{C}^{(n)}|) < \epsilon$ **or** $n \geq n_{\max})$

The complexity of each of the steps of the algorithm is linear in number of probes (K) and number of samples (I). For actual benchmarking results, see the help pages of the *ACNE* package.

2.2.4 Application to GWS arrays The latest generation of the Affymetrix arrays includes three or four probe pairs for each SNP. In these GWS arrays, the probes are technical replicates, i.e. they have the same sequence. Therefore, their affinities (for each allele) are expected to be identical. A possible model for these arrays is then (assuming three probe pairs per SNP)

$$\mathbf{Y} = \begin{bmatrix} \phi_{1A} & \phi_{1B} \\ \phi_{1A} & \phi_{1B} \\ \phi_{1A} & \phi_{1B} \\ \phi_{2A} & \phi_{2B} \\ \phi_{2A} & \phi_{2B} \\ \phi_{2A} & \phi_{2B} \end{bmatrix} \begin{bmatrix} C_{A1} & \ldots & C_{AI} \\ C_{B1} & \ldots & C_{BI} \end{bmatrix} + \mathbf{E}, \tag{7}$$

which is equivalent to the *cascaded NMF factorization* (Zdunek and Cichocki, 2008)

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \phi_{1A} & \phi_{1B} \\ \phi_{2A} & \phi_{2B} \end{bmatrix} \begin{bmatrix} C_{A1} & \ldots & C_{AI} \\ C_{B1} & \ldots & C_{BI} \end{bmatrix} + \mathbf{E}. \tag{8}$$

There are different adaptations of the NMF algorithms to deal with the additional fixed matrix that appears in Equation (8). One obvious way to perform the optimization is pre-multiplying the probe intensity matrix by the pseudoinverse of this first matrix. This method gives the intuitive solution of substituting the values of each probe by their corresponding means of the replicates, which can be robustified using a median estimator, cf. the CRMA v2 method.

Even in the case of technical probe replicates, it can be argued that the affinity of the probes is affected by the surrounding probes (Langdon *et al.*, 2009). Therefore, each affinity can be assumed to be independent and computed using the algorithm. In addition, the knowledge of having perfect replicates can be used to validate the result of the optimization. For these reasons, we chose to treat perfectly replicated probes independently and to process GWS arrays just like the earlier generations of chip types.

Equation (8) closely resembles the model of PLASQ (LaFramboise *et al.*, 2007). Although, PLASQ assumes that the affinity of the probes that belong to certain groups (forward strand, reverse strand, different alleles and different number of mismatches) are identical, which ACNE does not. Moreover, the main difference between ACNE and PLASQ is that, in ACNE the affinities and the CNs are computed for every dataset, whereas in PLASQ control samples are needed in order to compute the affinities and afterward these are used to compute the CNs in other datasets.

2.3 Robust scaling

ACNE provides a scaling step (Step 8 in the pseudocode) in order to obtain total CNs close to two. However, the results can be improved by finishing the algorithm with a more sophisticated method as follows. This step is not included within the main iteration loop in order to diminish the computing time.

The suggested adjustment succeeding the main algorithm fits a scale factor to each of the alleles by assuming that the sum of the total CN is equal to two for most of the samples. Usually, other methods such as CRMA v2 or dChip use the median or a trimmed mean of the total CN of the references (or all the samples) to achieve an analog scaling.

To do the scaling, we use the characteristic that NMF is not unique and that there always exists a family of matrices $\mathbf{T}$ that provides different factorizations

$$\mathbf{Y} \simeq \mathbf{\Phi C} = (\mathbf{\Phi T}^{-1})(\mathbf{TC}) = \tilde{\mathbf{\Phi}}\tilde{\mathbf{C}}, \tag{9}$$

where $\tilde{\mathbf{\Phi}}\tilde{\mathbf{C}}$ is a valid factorization if $\tilde{\phi}_{ku}$, $\tilde{c}_{ui} \geq 0$. If $\mathbf{T}$ is a 2×2 diagonal matrix with positive elements, its inverse is also a diagonal matrix with positive elements and, therefore, $\tilde{\mathbf{\Phi}}$ and $\tilde{\mathbf{C}}$ are positive because they are the sums and products of positive numbers. Matrix $\mathbf{T}$ can be used to scale $\mathbf{C}$ so that most of the samples, or at least a subset of them, have a CN close to two.

We can assume that most of the samples, or at least a subset of them, have neutral CNs. The column sums of the $\tilde{\mathbf{C}}$ matrix should be close to two, that is

$$[1\ 1]\tilde{\mathbf{C}} \simeq [2 \ldots 2]. \tag{10}$$

In turn,

$$[1\ 1]\begin{bmatrix} t_{11} & 0 \\ 0 & t_{22} \end{bmatrix}\begin{bmatrix} c_{11} & \ldots & c_{1I} \\ c_{21} & \ldots & c_{2I} \end{bmatrix} \simeq [2 \ldots 2], \tag{11}$$

or equivalently,

$$[t_{11}\ t_{22}]\mathbf{C} \simeq [2 \ldots 2], \tag{12}$$

which can also be written as

$$\mathbf{C}^T\begin{bmatrix} t_{11} \\ t_{22} \end{bmatrix} \simeq \begin{bmatrix} 2 \\ \ldots \\ 2 \end{bmatrix}. \tag{13}$$

Note that Equation (13) is a linear system of equations that can be solved for t_{11} and t_{22}. The corrected matrix $\tilde{\mathbf{C}}$ that provides ASCN is then

$$\tilde{\mathbf{C}} = \begin{bmatrix} t_{11} & 0 \\ 0 & t_{22} \end{bmatrix}\mathbf{C}. \tag{14}$$

Since there can be samples with loci whose total CNs are not equal to two (because there are CNAs in the region where the SNP is located in some samples), the linear system shown in Equation (13) must be solved robustly to withstand the presence of outliers. The outliers are, in this case, samples whose CNs are different to two. We use iteratively reweighted least squares (IWLSs) to solve the system of equations. If a group of samples in the experiment are known to be normals, they can be used as references and $\mathbf{C}^T$ should include only the rows that correspond to these samples.

This normalization step works well if, for any particular SNP, most of the samples have neutral CN, although all of them may be tumor samples. If an aberration occurs for a large percentage of the samples, the robust normalization method is not able to identify the normal samples. For these cases, additional normal samples must be added to the experiment, or the CN estimates will be biased in the affected region. We have tested in a simulation (data not shown) that the breakdown point is located ∼30%, i.e. at least 70% of the samples must behave normally *at a particular SNP* (not genome wide).

2.4 Downstream steps

Once the raw CN estimation is performed, a number of post-processing methods can be applied. For example, fragment length compensation can be executed in order to correct for different effectiveness of the restriction enzyme depending on the length of the DNA fragment (Nannya *et al.*, 2005). In addition, total CNs may also be corrected for GC-content effects (Nannya *et al.*, 2005). Finally, a segmentation method is utilized to divide the sample into regions with common CNs, e.g. circular binary segmentation (CBS) (Olshen *et al.*, 2004). All these processes can be applied after the summarization step but they are not part of ACNE and are not used in Section 3.

2.5 Implementation

The proposed ACNE method is available in the *ACNE* package implemented in *R* (R Development Core Team, 2010). In addition to providing a low-level estimator, the *ACNE* package also includes an add on to the high-level *aroma.affymetrix* framework (Bengtsson *et al.*, 2008a), which allows ACNE to be applied to very large Affymetrix SNP datasets. More information is available at http://www.aroma-project.org/.

3 RESULTS

To show the accuracy of the results, we have compared ACNE with CRMA v2 (Bengtsson *et al.*, 2009), Affymetrix's CN5 (Affymetrix Inc., 2008) and dChip (Li and Hung Wong, 2001). We divided the evaluation into two parts. First total CNs are assessed, then ASCNs. For ASCN, we compare ACNE with allele-specific versions of CRMA v2 and dChip. CN5 has not been included in this comparison because it needs paired samples in order to estimate ASCNs.

3.1 Total DNA CN results

The evaluation of the total CN results have been done by selecting a part of the genome in a tumor sample where there is a change in CN and comparing total CN for SNPs located on each side of the change point. We selected three different regions with a CN change. Figure 2a shows one such region near 150 Mb on chromosome 2 in the sample GSM182834 of the breast cancer study. This region presents a change from one to two copies. The other two regions are presented in the Supplementary Material, of which one is based on GWS6 data from Liu *et al.* (2009).

The SNPs located before the change point are considered to have a certain total CN (deleted region) and SNPs after the jump a larger total CN. For any threshold, we have true positives, TP, (SNPs in the normal region that are above the threshold), false positives, FP, (SNPs in the deleted region that are above the threshold), true

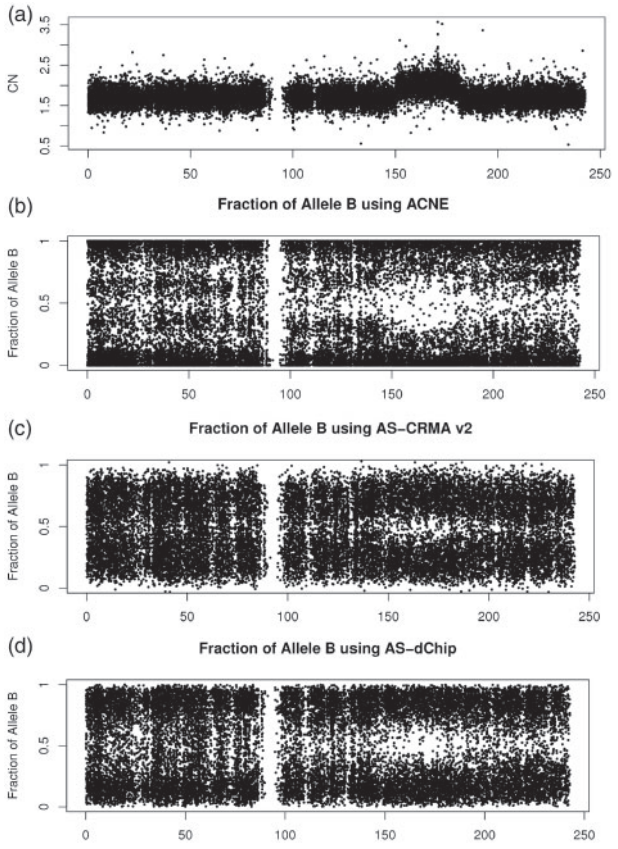

Fig. 2. Total CNs (**a**) along chromosome 2 in sample GSM182834 of the breast cancer study together with allele B fractions (β) using ACNE (**b**), AS-CRMA v2 (**c**) and AS-dChip (**d**). β is the ratio between allele B signal and the sum of the signals of both alleles. If a SNP has genotypes AA, BB and AB, β will be close to 0, 1 and 1/2, respectively. If the number of copies is larger than two (for example AAB), β will be close to 0.33. (a) Shows a deletion of almost all the chromosome. Comparing the CN and β plots, we infer that there is a loss of one copy of the chromosome and, at region 150 to 180 Mb, the cell replicated the remaining copy showing LOH with neutral CN. Moreover, this sample is not pure tumor, but a mixture of tumor and normal tissue. This explains that the deletion found through almost all the chromosome in (a) has four different clouds in the β plot (normal impurity provokes the additional clouds) and its fractional CN. The region with LOH with neutral CN (150–180 Mb) shows two broad clouds in (b) that correspond to four clouds hardly distinguishable for this level of noise. Data is for the Affymetrix Mapping250K_Nsp platform.

negatives, TN, (SNPs in the deleted region below the threshold) and false negatives, FN, (SNPs in the normal region that are below the threshold). The FP rate FPR = FP/(FP + FN) and the TP rate TPR = TP/(TP + TN) are evaluated at different thresholds constituting an ROC curve. This evaluation method was also used in Bengtsson *et al.* (2009) and is deeply explained in the corresponding Supplementary Material.

For this particular sample, the selected region is located from 140 to 160 Mb. We have also set a 'safety' zone between 148 and 152 Mb excluded in the analysis since it is difficult to know the exact location of the jump. Figure 3 shows that ACNE outperforms other methods (CRMA v2, dChip and CN5). CRMA v2 and dChip have

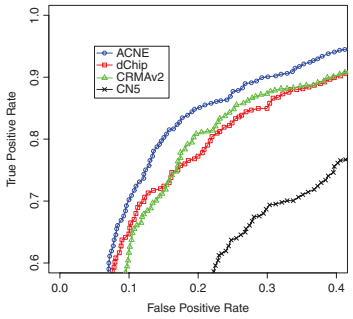

Fig. 3. This figure shows the ROC curve to distinguish a total CN alteration in sample GSM182834 shown in Figure 2a. Here, we have used the data within positions 140 and 160 Mb, omitting from the position at 148 to 152 Mb, where there is a jump from one to two copies. The number of SNPs that belong to the region under study is 1874. To have a ground truth, CN to the left of the change (excluding a 'safety' zone to avoid problems with the specific location of the jump) are considered to belong to one class, and CN to the right of the change is considered to belong to the other class. In turn, setting a threshold, it is possible to assign a state for each CN. Comparing the estimated state with the 'ground truth', it is possible to create a contingency table and, therefore, draw ROC curves. It can be seen that the ROC curve for ACNE is better (smaller FP and greater TP rates) than dChip or CRMA v2 for any threshold. Data is for the Affymetrix Mapping250K_Nsp platform.

similar performance. In this study, CN5 results are behind the other methods.

3.2 Allele-specific CN results

We have compared our method with the allele-specific data obtained by using dChip ('AS-dChip') and an allele-specific version of CRMA v2 ('AS-CRMA v2').

Validation of ASCN is harder, because there exist no 'ground truth' to compare with. We have taken two approaches to validate our method: the analysis of the plot of the fraction of the B allele signal (β) and the analysis of heterozygous and homozygous calls using genotyping data from HapMap.

3.2.1 More distinct allele B fractions Regarding β, its value for a SNP in a sample i is

$$\beta_i = \frac{C_{Bi}}{C_{Ai} + C_{Bi}}. \quad (15)$$

In a normal sample, the plot of β shows three clouds of points (close to 0, 1/2 and 1) as shown in Supplementary Figure S8.

Figure 2 shows the CN and the β plots of a tumor sample difficult to analyze. The CN plot (Fig. 2a) displays three different regions: from the beginning of the chromosome to 150 Mb (whose CN is smaller than two), from position 150 to 180 Mb (with neutral CN) and from 180 Mb to the end of the chromosome (CN is again below two). A likely reason why there is a fractional CN is that this sample is not pure tumor, but a mix of normal and tumor tissues. The presence of normal tissue provokes the existence of new clouds of points in the β plot. In a deletion, the tumor sample show only one allele A (B), and if in the normal tissue the same SNP has both alleles AB, then the allele B (A) will give some signal due to the mixture of both tissues. In the second region there is LOH with neutral CN from 150 to 180 Mb. In this case, the tumor tissue will have either AA or BB call. In this case, the signal of β for SNPs with

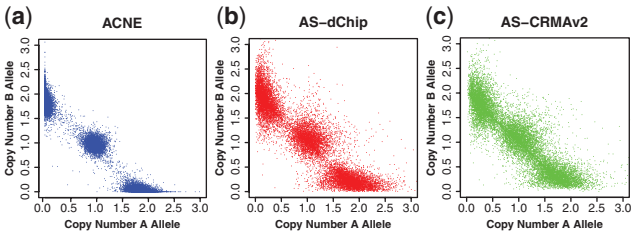

Fig. 4. Allele-specific CNs (C_A, C_B) using ACNE (**a**), AS-dChip (**b**) and AS-CRMA v2 (**c**). This figure compares the CNs for both alleles using all the SNPs of chromosome 8 from the HapMap sample NA12264. Ideally, there should be three clouds located around $(2,0)$, $(1,1)$ and $(0,2)$. ACNE is able to discern better the three clouds. Data is for the Affymetrix Mapping250K_Nsp platform.

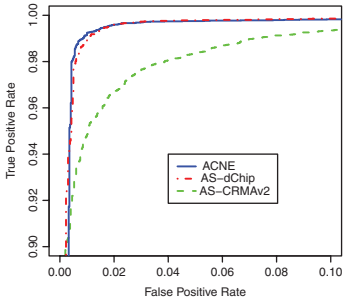

Fig. 5. ROC curve to distinguish heterozygous from homozygous genotypes using ACNE, AS-dChip and AS-CRMA v2. Previous figures shows intuitively that ACNE provides a better estimation of CN for each of the alleles. HapMap provides the genotypes for a number of samples. This ROC curve compares the genotypes of the chromosome 8 provided by HapMap for sample NA12264 with the ASCN provided by ACNE, AS-CRMA v2 and AS-dChip. We have considered the LH quantity to discern whether a SNP is homozygous or not. Setting different thresholds for LH and using the genotype information from HapMap it is possible to create a contingency table for each threshold and, hence, an ROC curve. The number of SNPs used to create the ROC is 14 189. Data is for the Affymetrix Mapping250K_Nsp platform.

heterozygous calls in normal tissue will also be shifted towards 1/2, but this effect is weaker than in the previous case because the signal for the tumor samples is two times larger.

Figure 2b, c and d shows the different clouds for β in the three regions using ACNE, AS-dChip and AS-CRMA v2. We conclude that in this particular sample there is a deletion in the first and the third segments of the chromosome and LOH with neutral CN in the middle region. This example illustrates how the estimation of ASCN clarifies the understanding of a particular sample.

ACNE separates better the different allele-specific states than AS-dChip or AS-CRMA v2 and the plot of β shows tighter clouds of points located in the expected positions.

3.2.2 Improved detection of heterozygous calls In addition to the study of the β plots, we have used the genotyping information provided by HapMap in order to evaluate ACNE. In a normal sample, CN for each of the alleles can be $(2,0)$, $(1,1)$ or $(0,2)$. We have analyzed some samples of the HapMap Mapping250K_Nsp dataset, which have the genotypes of the SNPs available. The genotype gives the ASCN in a normal sample.

The estimated ASCN for ACNE, AS-dChip and AS-CRMA v2 are shown in Figure 4. This figure shows intuitively that ACNE performs better than the other two methods. Combining ASCN with genotype information, it is possible to quantify the performance of each of the methods.

To do this, we calculate the *level of heterozygosity* (LH) defined as

$$\mathrm{LH}_i = 2 \cdot \frac{\min(C_{Ai}, C_{Bi})}{C_{Ai} + C_{Bi}}. \qquad (16)$$

Using ACNE, LH is within $[0,1]$. For AS-CRMA v2, this value can be slightly outside this range since negatives estimates of C_A and C_B are allowed. If a SNP is heterozygous its LH is close to one and if it is homozygous LH is close to zero. Therefore, LH can be used to discern whether a SNP is homozygous or not. Setting different thresholds for LH and using the genotype information from HapMap as the gold standard, it is possible to create a contingency table for each threshold and hence, an ROC curve. This curve informs us about the quality of different algorithms for computing LH and, therefore, the quality of the ASCN. Figure 5 shows the ROC curves using ACNE, AS-dChip and AS-CRMA v2 methods. In addition to this curve, we have included the density plots of allele B fractions for a normal sample (Supplementary Fig. S9). We have truncated the plots for ACNE above, because there are a large number SNPs with β very close to zero and one. These density plots show that the clouds of points using ACNE are tighter and their centroids are closer to the theoretical values than for AS-dChip and AS-CRMA v2.

4 DISCUSSION

The greatest improvement of ACNE summarization method, if compared with previous state of the art summarization methods, is its ability to estimate more accurately the CN values of each allele. We also observe an improvement for total CNs compared with CRMA v2, dChip and CN5. One of the reasons why ACNE outperforms the others is that it is able to estimate the cross hybridization of each probe included in a SNP.

In the case of the GWS6 arrays, as explained in Section 2.2.4, there are only three or four replications of one SNP probe pair. This problem can be solved reasonably well using other algorithms instead of ACNE. As shown in the Supplementary Material, the improvement in total CN for these arrays is not so strong as in the previous chip types of Affymetrix, although the ASCN results are better using ACNE than CRMA v2 (Supplementary Figs S10 and S11).

The complexity of ACNE is linear with the number of probes (K) and samples (I). Computing time in simulations for up to 5000 samples confirms, almost perfectly, a linear relationship. In other words, ACNE scales well with the number of samples.

Recently Staaf *et al.* (2008) suggested new segmentation algorithms that use as input data not only the total CN, but also the fraction of the allele B (β). These algorithms have been applied mainly to Illumina data since the β plot is less noisy. An accurate estimation of the β will help these and other algorithms to discern the different segments of the genome using Affymetrix arrays.

5 CONCLUSION

This article describes a new algorithm for estimating ASCNs from any of Affymetrix genotyping arrays. The initial preprocessing steps

are borrowed from CRMA v2. Specifically, the algorithm is focused in the summarization method. Using NMF, it provides an estimation for the CN of each of the alleles. ROC analysis shows that ACNE outperforms other state of the art methods, not only in ASCN but also in total CN estimations. These improvements, especially in ASCN, make it possible to augment the information of segmentation algorithms to discover different aberrations in the genome.

Funding: University of Navarra; Fundación para la Investigación Médica Aplicada; NCI (grant U24 CA126551).

Conflict of Interest: none declared.

REFERENCES

Affymetrix Inc. (2008) *Affymetrix Genotyping Console 3.0 - User Manual.* Affymetrix.

Affymetrix Inc. (2009) http://www.affymetrix.com/.

Altshuler,D. *et al*. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Bengtsson,H. *et al*. (2008a) aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Technical Report 745*. Department of Statistics, University of California, Berkeley.

Bengtsson,H. *et al*. (2008b) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.

Bengtsson,H. *et al*. (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**, 2149–2156

Duffy,K. *et al*. (2008) A novel procedure for genotyping of single nucleotide polymorphisms in trisomy with genomic DNA and the invader assay. *Nucleic Acids Res.*, **36**, e145.

Haverty,P. *et al*. (2008) High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer*, **47**, 530–542.

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

LaFramboise,T. *et al*. (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics*, **8**, 323–336.

Langdon,W. *et al*. (2009) Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. *Brief. Bioinform.*, **10**, 259–277.

Lee,D. and Seung,H. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Li,C. and Hung Wong,W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, **2**8, RESEARCH0032.

Liu,W. *et al*. (2009) Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat. Med.*, **15**, 559–565.

Nannya,Y. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.*, **65**, 6071–6079.

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

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pollack,J. *et al*. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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.

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

Zdunek,R. and Cichocki,A. (2008) Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems. *Comput. Intell. Neurosci.*, **2008**, 939567.