BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–10

SUPPLEMENTARY ARTICLE - CalMaTe: A method and software to improve allele-specific copy number of SNP arrays for downstream segmentation.

Maria Ortiz-Estevez [1,2], Ander Aramburu [1], Henrik Bengtsson [3,4], Pierre Neuvial [3,5] and Angel Rubio [1*]

[1]Group of Bioinformatics, CEIT and TECNUN, University of Navarra, San Sebastian, Spain. [2]Computational Biology Group, Celgene Institute for Translational Research (CITRE), Sevilla, Spain. [3]Department of Statistics, University of California, Berkeley, USA. [4]Department of Epidemiology and Biostatistics, University of California, San Francisco, USA. [5]Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 - USC INRA, France.

S1 FEATURES OF ASCN PROCESSING METHODS

Table S1 illustrates the characteristics and differences among the cited ASCN estimation methods and CalMaTe.

	"Illumina"	ACNE	TumorBoost	CalMaTe
Open-source implementation	-	x	x	x
Cross-platform	-	-	x	x
Applicable w/o matched normals	x	x	-	x

Table S1. CalMaTe is the only ASCN processing method that is open source, cross-platform, and that does not require matched normals.

S2 AVAILABILITY OF CALMATE IN R

The CalMaTe algorithm takes allele-specific SNP signals as input, or more generally, total copy-number signals (TCNs) and B-allele fractions (BAFs) from any microarray technology.

The CalMaTe method is implemented in *calmate*, which is an open-source R package available on CRAN (http://cran.r-project.org/). The *calmate* package provides a high-level application programming interface (API) for CalMaTe, which plugs into the Aroma Project framework (http://www.aroma-project.org/) and uses its data types, as well as a low-level API that uses basic R data types. One advantage of using the Aroma Project framework is that all its methods are designed to be bounded in memory, meaning that any number of arrays can be processed even with limited amount of RAM (as low as 0.5-1.0 GB). Another advantage is that all results are persistent so that they are readily available also after restarting R. The low-level API is made available so that CalMaTe can be incorporated and used elsewhere, e.g. Bioconductor. Internally, the high-level API utilizes the low-level API.

*to whom correspondence should be addressed

S2.1 High-level API for the Aroma Project framework

Here we use the public Affymetrix Mapping250K_Sty dataset GSE12702 (Castro *et al.*, 2009) from GEO to illustrate how to execute CalMaTe using the R package *calmate*. To obtain TCN and BAF signals from the Affymetrix dataset, we utilize an allele-specific version of the CRMA v2 preprocessing method (Bengtsson *et al.*, 2009a) available in the *aroma.affymetrix* R package (Bengtsson *et al.*, 2008). To apply AS-CRMA v2 (or short just CRMA v2) on this dataset, do

```
library("aroma.affymetrix");
csR <- AffymetrixCelSet$byName("GSE12702",
                chipType="Mapping250K_Nsp");
dsN <- doASCRMAv2(csR, plm="RmaCnPlm");
```

Because this will take several minutes per array processed, we recommend adding argument `verbose=TRUE` to see progress. More details on `doASCRMAv2()` can be found on the Aroma Project website. The `dsN` object returned by `doASCRMAv2()` contains both TCN and BAF estimates. Contrary to the `csR` AffymetrixCelSet object, the `dsN` object is not specific to a particular technology, i.e. what follows would be same for, say, Illumina microarray data.

To calibrate the TCN and BAF estimates obtained above using CalMaTe do

```
library("calmate");
cmt <- CalMaTeCalibration(dsN);
dsNC <- process(cmt);
```

Note that the second line of code only sets up the method, whereas it is at the last line that CalMaTe is actually performed. Likewise, the above takes several minutes per array, so adding `verbose=TRUE` is useful. The `dsNC` object returned by `process()` contains both TCN and BAF estimates in the same format as input `dsN` object.

For a thorough example with detailed illustrations on how to extract and plot the calibrated TCN and BAF estimates, or the corresponding calibrated ASCN estimates, see the vignettes available online at the Aroma Project website.

S2.2 Low-level API

If TCN and BAF estimates for a set of samples already exist, for instance by using another SNP microarray technology or another preprocessing pipeline, then the low-level API of the *calmate* package can be used.

Assume that the TCN and BAF estimates are available in R as a $J\times 2\times I$ array named `data`, where the first dimension specifies loci $j = 1, \ldots, J$, the second dimension TCN and BAF, and the third dimension arrays $i = 1, \ldots, I$. For non-polymorphic loci, BAFs are not defined, which is represented as a missing value (`NA`/`NaN` in R).

With this setup, the $J\times I$ TCN matrix for all loci across all arrays can be obtained as `TCN <- data[,1,]`. Analogously, the corresponding BAF matrix is `BAF <- data[,2,]`. Moreover, the $2\times I$ TCN and BAF matrix for, say, the 54^{th} locus across all arrays can be obtained as `locusData <- data[54,,]`. If this locus is non-polymorphic, then all of the BAF signals in `locusData[2,]`, which is a vector of length I, are missing values.

To calibrate these TCN and BAF estimates using CalMaTe do

```
library("calmate");
dataC <- calmateByTotalAndFracB(data);
```

The `dataC` object returned by `calmateByTotalAndFracB()` has the same data type and dimension ($J\times 2\times I$) as `data`, making it easy to use in place of `data`.

If the estimates are available in the ASCN space, these should be transformed to the TCN and BAF space. Alternatively, `calmateByThetaAB()`, which takes ASCN estimates, may be used. Note that this only works for SNPs, because ASCNs are not defined for non-polymorphic loci. In other words, it is up to the user to make sure this function is only used for SNPs and not for non-polymorphic loci.

As for the high-level API, there are also low-level examples in the vignettes available online at the Aroma Project website.

S3 CALMATE METHOD

S3.1 SNP-specific crosstalk model

The main assumption of CalMaTe is that cross-hybridization between alleles is linear and possibly different between SNPs but preserved across samples. Consider a SNP $j = 1, \ldots, J$, and let $\mathbf{H}_j^c$ be the $2\times I$ matrix with column vectors $(C_{Aij}, C_{Bij})^T$ of the unobserved *true ASCNs* across all samples $i = 1, \ldots, I$. The corresponding *observed ASCNs* $\mathbf{H}_j$ can then be modeled as

$$\mathbf{H}_j = \mathbf{W}_j \mathbf{H}_j^c + \boldsymbol{\varepsilon}_j, \tag{S1}$$

where $\mathbf{W}_j$ is an unknown 2×2 *crosstalk* matrix shared by all samples, and $\boldsymbol{\varepsilon}_j$ is a $2\times I$ error matrix. In turn, assuming that $\mathbf{W}_j$ is invertible, then

$$\mathbf{H}_j^c = \mathbf{T}_j \mathbf{H}_j + \boldsymbol{\xi}_j, \tag{S2}$$

where $\boldsymbol{\xi}_j = -\mathbf{W}_j^{-1}\boldsymbol{\varepsilon}_j$ and $\mathbf{T}_j = \mathbf{W}_j^{-1}$ is a 2×2 matrix that backtransforms the observed ASCNs ($\mathbf{H}_j$) into true ASCNs ($\mathbf{H}_j^c$) plus noise. $\mathbf{W}_j$ is expected to be diagonally dominant because the affinity of a probe is larger for its perfect-match DNA than for a sequence with one mismatch (the other allele). Since a diagonally dominant matrix is always invertible (Levy-Desplanques Theorem), $\mathbf{W}_j$ can be inverted.

In what follows, we will for simplicity drop SNP index j, i.e. we will write $\mathbf{W}$ instead of $\mathbf{W}_j$, T instead of T_j, $\mathbf{H}^c$ instead of $\mathbf{H}_j^c$, and so on. This also illustrates that CalMaTe is a method applied to each SNP independently.

S3.2 Fitting the model

CalMaTe uses a specific subset of S normal samples (R) as a reference. For the reference set, we have from Equation (S1) that

$$\mathbf{H}_\mathrm{R} = \mathbf{W}\mathbf{H}_\mathrm{R}^c + \boldsymbol{\varepsilon}, \tag{S3}$$

where we add notation R to denote that this is for the reference samples. Since the reference samples are normal, we expect the ASCNs of $\mathbf{H}_\mathrm{R}^c$ to be either $(2,0)^T$, $(1,1)^T$ or $(0,2)^T$, corresponding to genotypes AA, AB and BB, respectively. For example,

$$\mathbf{H}_R^c = \begin{bmatrix} 2 & 1 & \ldots & 0 & 1 \\ 0 & 1 & \ldots & 2 & 1 \end{bmatrix} \tag{S4}$$

with rows representing the two alleles (A and B) and the columns the S normal samples within the reference set (R). The set of possible states is *small and discrete*. For this reason it is feasible to estimate $\mathbf{H}_\mathrm{R}^c$ from data (Section S3.2.1) and hence solve for $\mathbf{W}$ (Section S3.2.2). Note that in general there is no such constraint on $\mathbf{H}^c$, which is key when analyzing non-homogeneous samples such as tumors. With an estimate of $\mathbf{W}$, and hence $\mathbf{T}$, it is possible to calibrate observed ASCNs $\mathbf{H}^c$ for *all* samples (Section S3.3). The more reference samples used, the more stable and precise the parameter estimates will be, cf. Section S5.4. We recommend to use $S \geq 6$ reference samples.

If no normal samples are specified, all the samples are used as references, and as explained below we will rely on robustness of the estimators to obtain a good estimate of $\mathbf{W}$.

S3.2.1 A very simple genotyping algorithm. Consider a particular SNP in a set ('R') of S normal samples. The genotype of sample $s = 1, \ldots, S$ can be called from the observed BAF ($\{\beta_s\}$) as

$$\hat{\mathbf{H}}_{s,R}^c = \begin{cases} (2,0)^T & \text{if } \beta_s \leq 1/3 \\ (0,2)^T & \text{if } \beta_s \geq 2/3 \\ (1,1)^T & \text{otherwise,} \end{cases} \tag{S5}$$

where $\hat{\mathbf{H}}_{s,R}^c$ is column s of matrix $\hat{\mathbf{H}}_R^c$. If there was no genotyping errors, then the called $\hat{\mathbf{H}}_R^c$ would be identical to the true $\mathbf{H}_R^c$. However, since the genotypes are only used for estimating the crosstalk parameters, having a few genotyping errors is not critical.

S3.2.2 Estimating crosstalk parameters. Substituting $\mathbf{H}_R^c$ with $\hat{\mathbf{H}}_R^c$ and $\mathbf{W}^{-1} = \mathbf{T}$ with $\hat{\mathbf{T}}$, we can invert Equation (S3) as

$$\hat{\mathbf{T}}\mathbf{H}_R = \hat{\mathbf{H}}_R^c + \boldsymbol{\zeta}, \tag{S6}$$

where $\boldsymbol{\zeta}$ is an error term. This matrix equation can be solved for $\hat{\mathbf{T}}$ by multiplying it by the pseudoinverse of $\mathbf{H}_R$. However, since there may be genotyping errors, we use a robust solver. The solution is obtained in two steps. Firstly, two subproblems (using constraints on the sum and the difference of the CNs of the two alleles) are robustly solved and secondly, the entries of the $\hat{\mathbf{T}}$ matrix are found.

In what follows, let

$$\hat{\mathbf{T}} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}, \quad \text{(S7)}$$

where we, for sake of clarity, also drop the circumflex ("ˆ") of the individual entries. Using the example genotypes (assuming no errors for simplicity), Equation (S6) expands to

$$\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \mathbf{H}_R = \begin{bmatrix} 2 & 1 & \ldots & 0 & 1 \\ 0 & 1 & \ldots & 2 & 1 \end{bmatrix} + \boldsymbol{\zeta}. \quad \text{(S8)}$$

Constraint on allele sums: The sum of the CNs of the two alleles for the references is expected to be 2. Ignoring the random errors ($\boldsymbol{\zeta}$), we therefore have

$$[1 \;\; 1]\hat{\mathbf{T}}\mathbf{H}_R = [2 \;\ldots\; 2]. \quad \text{(S9)}$$

Focusing on the product of the vector of ones with the $\hat{\mathbf{T}}$ matrix,

$$[1 \;\; 1] \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = [t_{11} + t_{21} \; t_{12} + t_{22}] = [D \; E], \quad \text{(S10)}$$

we can write Equation (S9) as

$$[D \; E]\, \mathbf{H}_R = [2 \;\ldots\; 2]. \quad \text{(S11)}$$

Equation (S11) is a linear system of equations in D and E which can be solved using robust methods. In particular, CalMaTe uses an *iteratively reweighted least squares* (IWLS) method, which is implemented in the `rlm()` function of the *MASS* package in R.

Constraint on allele differences: The difference between the CNs of the two alleles is expected to be either 2, 0, or -2, if the predicted genotype is AA, AB, and BB, respectively. Because of this and using the previous example, we have that

$$[1 \;\; -1]\hat{\mathbf{T}}\mathbf{H}_R = [2\,0 \;\ldots -2\;0]. \quad \text{(S12)}$$

Expanding the $\hat{\mathbf{T}}$ matrix we get

$$[1 \;\; -1] \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = [t_{11} - t_{21} \; t_{12} - t_{22}] = [F \; G], \quad \text{(S13)}$$

and therefore we can write Equation (S12) as

$$[F \; G]\, \mathbf{H}_R = [2\,0 \;\ldots -2 \;\; 0]. \quad \text{(S14)}$$

This system of equations can also be solved using IWLS.

At the end, when both linear systems have been solved for D, E, F and G, we combine Equations (S10) and (S13) as

$$\begin{bmatrix} D & E \\ F & G \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}, \quad \text{(S15)}$$

from which $\hat{\mathbf{T}}$ follows by inversion

$$\hat{\mathbf{T}} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} D & E \\ F & G \end{bmatrix}. \quad \text{(S16)}$$

To conclude, the above procedure provides a robust estimate of $\mathbf{T}$.

S3.2.3 Non-identifiable cases. If, for a particular SNP, all the normal samples happen to have the same homozygous (AA or BB) genotype, then it is not possible to obtain affinities of the non-present allele. In order to handle this special case we constrain the model by assuming that the crosstalk is symmetric for such SNPs. In practice, this is done by swapping half of the values of the initial matrix (assigning the values of the A allele to the B allele and vice versa) and then calculating the calibration matrix as above.

In addition to the non-identifiable cases, in some cases the robust linear solver does not converge in a reasonable number of iterations. In these cases, the software computes the centroid of all the SNPs that have the same genotype by using the medians of the signals for each allele. Once the location of the three points (for AA, BB and AB genotypes) is known, the regression line that links them using minimum squares is computed. The errors are weighted by the number of samples within each genotype so that, the regression line gets closer to the genotype that represents more samples. If one of the genotypes is missing, the regression line is computed based on the other two points. If there exists a single point (all the samples have the same genotype and hence, a non-identifiable case), the regression line is computed by assuming that the calibration matrix is symmetric.

S3.3 Calibration

Finally, with an estimate $\hat{\mathbf{T}}$, *all* samples can be calibrated. The *calibrated ASCNs* $\tilde{\mathbf{H}}^c$ can be calculated as

$$\tilde{\mathbf{H}}^c = \hat{\mathbf{T}}\mathbf{H}, \quad \text{(S17)}$$

which is a backtransformation that follows from Equation (S2) by substituting $\mathbf{T}$ with $\hat{\mathbf{T}}$ and dropping the error term. We denote $\hat{\mathbf{T}}$ the *calibration matrix*, because it calibrates the observed ASCNs. Contrary to the notation of the above *parameter estimates*, we use the tilde ("˜") notation to denote that $\tilde{\mathbf{H}}^c$ contains *calibrated* ASCNs.

S4 PERFORMANCE

S4.1 ROC analysis

In order to formally evaluate the influence of CalMaTe on signal to noise ratio, we have used receiver operating characteristic (ROC) analysis on several known change points. ROC analysis was performed as described in Bengtsson *et al.* (2009a, 2010). We refer to these papers for a more comprehensive description of this ROC evaluation. We chose the same change points as those used for the evaluation of the TumorBoost method (Bengtsson *et al.*, 2010) in order to facilitate comparison between CalMaTe and TumorBoost, and interpretation of the results. These change points are taken from a specific tumor-normal sample: TCGA-23-1027. They correspond to four common copy number state transitions, and at one region with no change point (negative control) as summarized in Bengtsson *et al.* (2010, Table 1). This evaluation was performed on hybridization data from the Affymetrix GenomeWideSNP_6 platform as well as the Illumina Human1M-Duo platform. See the two dedicated ROC Supplementary Notes for comprehensive results.

Here, we explain in detail how this ROC evaluation has been carried out for a change point at ∼124Mb on Chr. 2 between a

normal region, and a region of gain of one DNA copy (Fig. 1 and Fig. S4). This change point corresponds to a change in both TCN and BAF: we label the normal state (left of the change point) as "negative", and the gained state (right of the change point) as "positive". We focus on a genomic region surrounding the change point. For BAF signals, we calculate for each heterozygous SNP j (the definition of a heterozygous SNP is given and discussed in the next section) the Decrease in Heterozygosity: $\mathrm{DH}_j = 2|\mathrm{BAF}_j - 1/2|$. DH is appropriate for ROC evaluation purposes because contrary to BAF its distribution only has one mode, as discussed e.g. in Staaf *et al.* (2008); Bengtsson *et al.* (2010).

Consider a threshold value τ. A SNP j is classified as "positive" if $\mathrm{DH}_j \geq \tau$, and as "negative" otherwise, and we report the true-positive rate (TPR) and the false-positive rate (FPR) in the region. We build a ROC curve for DH by plotting TPR against FPR in the genomic region for each possible value of τ. A ROC curve can be built along similar lines for TCN, where a locus j is classified as "positive" if $\mathrm{TCN}_j \geq \tau$. Using this strategy, we estimate a TCN and a DH ROC curve for each method to be compared, e.g. genomic signals before CalMaTe, and after CalMaTe.

S4.2 Genotypes

Because our ROC evaluation is based on DH (in addition to TCN), which are only defined for heterozygous SNPs, the evaluation itself requires that we call genotypes and identify heterozygous SNPs.

First of all, following Bengtsson *et al.* (2010), we only talk about genotypes for normal cells. In particular, we don't define or discuss the notion of genotype in tumor cells. We define the "genotype" of a SNP in a sample as the genotype of this SNP in the germline. A SNP can only have three genotypes (AA, AB or BB), if we exclude trisomy and rare cases of copy number polymorphisms in a SNP. A SNP is said to be heterozygous if its (germline) genotype is AB, and homozygous otherwise.

The choice of genotyping algorithm is not critical for the assessement, that is, even a very basic caller will do, because the assessment itself, as well as the segmentation methods it is imitating, is rather insensitive to a few genotyping errors (as long as they are randomly scattered along the genome). More importantly, in order to make the comparison fair, we keep as much as possible similar by using the same naive genotyping algorithm (Bengtsson *et al.*, 2010) for all sets of ASCNs evaluated.

Second, CalMaTe will adjust ASCNs such that the genotype calls for some SNPs will not be the same before and after calibration. For the objective of identifying genomic aberrations, potential genotyping discrepancies are not a concern. We wish to emphasize that CalMaTe was designed to improve the SNRs along the genome for the purpose of identifying genomic aberrations and not per SNP. Because of this, we advice against using CalMaTe ASCNs for genotype studies per se, e.g. GWAS. Continuing, because of the above discrepancies in genotype calls, there will be *different sets (and thus different numbers) of heterozygous SNPs* and hence a different set of DH signals available for the ROC analysis. This means that it is not possible to directly compare the ROC curves for CalMaTe on the one hand with those of the raw signals and TumorBoost on the other hand. To overcome this limitation, we choose to reestimate DH signals from the existing ones at a set of common loci, which is an idea borrowed from Bengtsson *et al.* (2009b). Specifically, we split the genome into non-overlapping bins of a certain size (e.g. $h = 25$kb in the example below) and calculate the mean DH in each bin. This approach asserts that the resulting ROC curves are objectively comparible across methods, while still making use of all available data.

Note that the evaluation of TCN does not require genotype calls, as it is defined for any SNP and any CN locus. In order to make the ROC curves for TCN comparable with the ROC curves for DH, we also smoothed the TCN signals using the same bins ("same resolution") as for DH.

S4.3 Results of the ROC analysis

Figure S1 illustrates the results of the ROC analysis described above for a change point between a normal and a gained region on Chromosome 2 of sample TCGA-23-1027, for a bin size of $h = 25kb$. Results for other bin sizes and change points, as well as for both Affymetrix and Illumina, are given in the dedicated Supplementary Notes. The general conclusion is that, with CalMaTe there is more power to detect a given change point than without, but also than with TumorBoost.

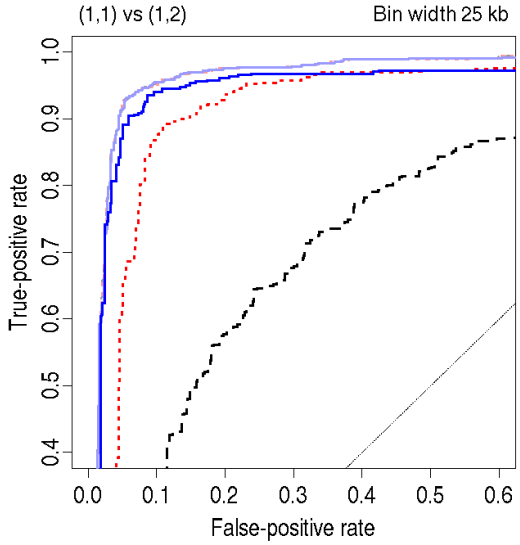


Fig. S1. Results of ROC analysis for a change point between a normal and a gained region on Chromosome 2 of sample TCGA-23-1027, for a bin size of $h = 25kb$. The TCN and DH curves for the same method are depicted with the same line type and color with the difference that the TCN curves use a lighter version of color for "CRMAv2" (dashed black), "CRMAv2,TumorBoost" (dotted red), and "CRMAv2,CalMaTe" (solid blue). The three TCN curves overlap closely making them appear as one curve.

S5 EXAMPLE RESULTS OF CALMATE

S5.1 Allele-specific copy numbers

Figure S2(a) shows the different genotyping clouds of four SNPs (SNP_A-2010640, SNP_A-2010642, SNP_A-2010643 and SNP_A-2010648) in 59 HapMap samples (The International HapMap Consortium, 2003) hybridized on the Affymetrix GenomeWideSNP_6 chip type. These ASCN have been obtained using CRMA v2. Each SNP is plotted in a different color. In Figure S2(a) it can be seen that the clouds are *not* centered at

their theoretical locations at $(2,0)$, $(1,1)$ and $(0,2)$, although the samples are normal. Figure S2(b) shows the ASCNs after applying CalMaTe. In this case the clouds are centered closer to their expected location ("accuracy") and they are clustered tighter ("precision") than before.

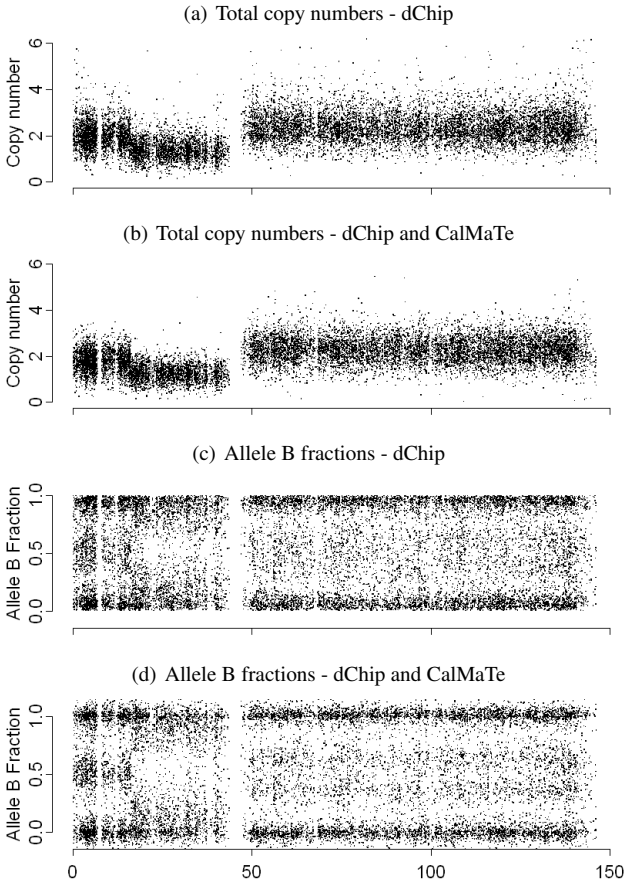


Fig. S3. TCNs and BAFs along Chr. 8 of tumor sample GSM318736 with and without CalMaTe. Data are from dChip-processed Affymetrix Mapping250K_Nsp arrays. Panels (a) and (b) show the TCNs before and after CalMaTe, respectively. Panels (c) and (d) show the corresponding BAFs. There are three main regions, a normal at 0-20Mb with 2 copies and BAFs near 0, 1/2 and 1 (AA, AB and BB), a deletion at 20-44Mb with 1 copy and BAFs close to 0 and 1 (A and B), and a gain at 49-145Mb with 3 copies and BAFs near 0, 1/3, 2/3, and 1 (AAA, AAB, ABB and BBB).

S5.2 Effect on signals summarized by dChip

In the Application Note, we applied CalMaTe to Affymetrix GenomeWideSNP_6 data preprocessed by CRMA v2. However, it can be also applied to other chips or preprocessing methods such as dChip (Lin *et al.*, 2004). Figure S3 shows the TCNs and BAFs obtained by dChip with and without CalMaTe. These data were obtained from NCBI-GEO GSE12702 prostate cancer dataset (Castro *et al.*, 2009) consisting of 20 tumors and 20 normals hybridized to Affymetrix Mapping250K_Nsp arrays. These arrays were also included in order to show that CalMaTe can be applied to any SNP CN microarray technologies (not only the most recent ones). Moreover, these types of arrays are currently among the most commonly available chiptypes on Gene Expression Omnibus (GEO), as shownin Table S2. All 40 samples were used as a reference to illustrate the robustness of the estimator (Section S2). After CalMaTe, the BAFs (Fig. S3(d)) better distinguish the expected number of genotype tracks and their positions than the BAFs given directly by dChip (Fig. S3(c)). Similar improvements are seen if dChip would be replaced by CRMA v2 (not shown for this dataset).

S5.3 Effect on Illumina signals

To show that CalMaTe also works with other microarray platforms than Affymetrix, we also present the result of applying CalMaTe to Illumina data. More specifically, from The Cancer Genome Atlas (TCGA) project (TCGA, 2011) we downloaded the Illumina Human1M-Duo data for the same ovarian tumor sample (TCGA-23-1027) that was used in the main paper to illustrate CalMaTe on Affymetrix GenomeWideSNP_6 (Fig. 1) as well as in the ROC analysis. Figures S4(a) and (c) show TCNs and BAFs as obtained by "XY method" available in Illumina's BeadStudio software (Illumina, 2007), whereas Figures S4(b) and (d) show the same data calibrated by CalMaTe. As was the case for the Affymetrix data, the BAFs after CalMaTe calibration distinguish the different genomic segments better.

S5.4 Effect of the number of reference samples

CalMaTe uses all or a subset of the arrays as references to compute the calibration matrix. Naturally, using a larger number

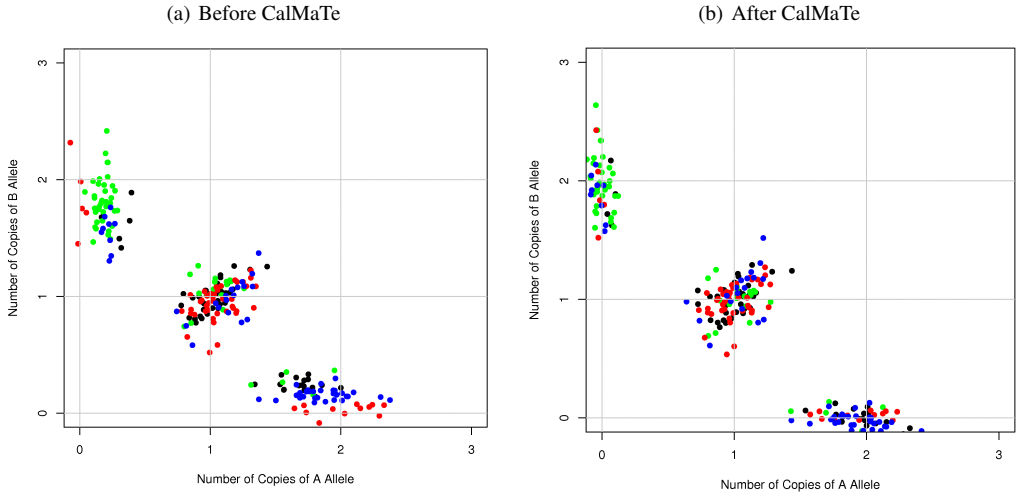


Fig. S2. ASCNs (C_A, C_B) for 59 HapMap samples at four SNPs (SNP_A-2010640, SNP_A-2010642, SNP_A-2010643 and SNP_A-2010648) before (Panel (a)) and after (Panel (b)) CalMaTe. The different SNPs are plotted in different colors. The samples were hybridized on Affymetrix GenomeWideSNP_6 arrays and preprocessed using CRMA v2. **(a)** Each SNP forms tight clouds that correspond to the three possible genotypes in a normal sample. However, the location of these clouds is different for each of the SNPs and not centered at the expected locations $(2, 0)$, $(1, 1)$ and $(0, 2)$ corresponding to genotypes AA, AB and BB. This location bias translates into noise when studying BAFs along the genome for a particular sample. **(b)** After CalMaTe, the ASCN estimates are more accurate as well as more precise.

Array generation	Chip type	Number of samples	GEO Platform(s)	URL(s)
GenomeWideSNP_6	GenomeWideSNP_6	**6026**	GPL6801	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6801
GenomeWideSNP_5	GenomeWideSNP_5	**279**	GPL6804, GPL9704	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6804 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL9704
Mapping250K		**16233**		
	Mapping250K_Nsp	8574	GPL3718, GPL3811	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3718 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3811
	Mapping250K_Sty	7659	GPL3720, GPL3812	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3720 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3812
Mapping50K		**7418**		
	Mapping50K_Hind240	3467	GPL2004, GPL2014	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2004 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2014
	Mapping50K_Xba240	3951	GPL2005, GPL2015	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2005 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2015
Mapping10K		**8604**		
	Mapping10K_Xba142	8185	GPL2641	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2641
	Mapping10K_Xba131	356	GPL1266	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL1266
	Mapping10K Early Access	63	GPL1855, GPL3400	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL1855 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3400

Table S2. Summary of Affymetrix SNP & CN data sets on GEO as of February 27, 2012

of references improves the estimates both of the TCNs and the BAFs. Figures S5 and S6 illustrate this fact. Figure S5 is an "after versus before" plot that compares the estimated BAFs of a normal sample using CRMA v2 alone against also using CalMaTe for different number of references. Since CRMA v2 is a single-sample summarization method, the CRMA v2 BAFs estimates are identical for any reference. These data were obtained from NCBI-GEO GSE19539 ovarian tumor dataset (Affymetrix GenomeWideSNP_6). The sample whose TCNs and BAFs were estimated is GSM492495_IC022N. It can be seen that for larger numbers of references the clouds are more tightly clustered. The density plots of the BAF estimates using CalMaTe are included for convenience. In Figure S6, a similar effect occurs in the TCN estimates: the interquantile range (distance between the first and the third quartile) decreases as the number of references increases.

S5.5 Segment calling using simulated data

We have tested CalMaTe using simulated data. These data were generated as follows. We ran CalMaTe on 25,000 SNPs in Chr. 22 from 30 CEU HapMap samples (Affymetrix GenomeWideSNP_6). We modified the code from CalMaTe to get the calibration matrix and the residuals of the fit for each SNP. Multiplying these calibration matrices by the simulated copy numbers and adding the computed residual (from a randomly selected different sample), we generated artificial signals for each of the alleles. The simulated copy numbers were obtained assuming the proportion of heterozygous SNPs to be 27%. Within the chromosome, some regions of different lengths were assumed to have 3 copies in the tumoral tissue. For the SNPs within these regions, the allele-specific copy numbers of the tumor (number of copies of allele A and B) are (0,3), (1,2) (2,1), (0,3). We assumed that the tumor purity was 25%.

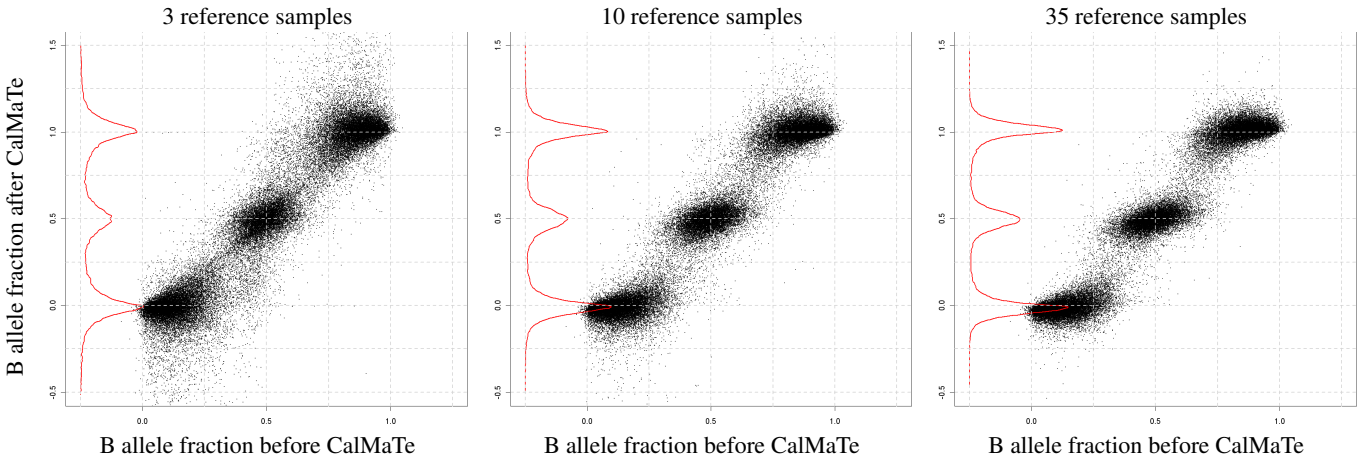


Fig. S5. Comparison of the BAF estimates with and without CalMaTe along chromosome 7 for the normal sample GSM492495_IC022N from NCBI-GEO GSE19539 ovarian tumor dataset. The calibration with CalMaTe is performed using 3, 10 and 35 references, respectively. In addition, the density plot of the calibrated BAF with CalMaTe is represented in red, in the Y axis. As it is shown, the BAF estimates using CalMaTe improve, as the number of references increases. In none of the five simulations, the normal sample GSM492495_IC022N has been included in the set of references. The samples were hybridized on Affymetrix GenomeWideSNP_6 arrays and preprocessed using CRMA v2.

For this contamination percentage, the simulated true copy numbers are (0,2.25), (1,1.25), (1.25,1), (0,2.25). Specifically, for each SNP, the true copy numbers in the tumor and in the normal fall in one of the four categories shown in Table S3. The lengths of the segments

Tumor	**Matched Normal**	**25% Tumor + 75% Normal**
(0,3)	(0,2)	(0,2.25)
(1,2)	(1,1)	(1,1.25)
(2,1)	(1,1)	(1.25,1)
(3,0)	(2,0)	(2.25,0)

Table S3. Correspondence between allele-specific copy numbers in a pure tumor, a matched normal, and a tumor with 75% normal contamination.

are 2000, 1000, 500, 300, 200, 100, 60 and 30 SNPs. Each of them are flanked by normal regions of 2000 SNPs. Non-polymorphic loci (aka "copy-number probes") were not included in the simulation.

As it can be seen from Figures S7 and S8, the use of CalMaTe allows the segmentation methods (PSCN (Chen *et al.*, 2011) and PSCBS (Olshen *et al.*, 2011)) to find more segments than using CRMA v2 with or without TumorBoost. These findings are in line with the ones of the ROC analyses, which assess the power to detect individual change points (not segments).

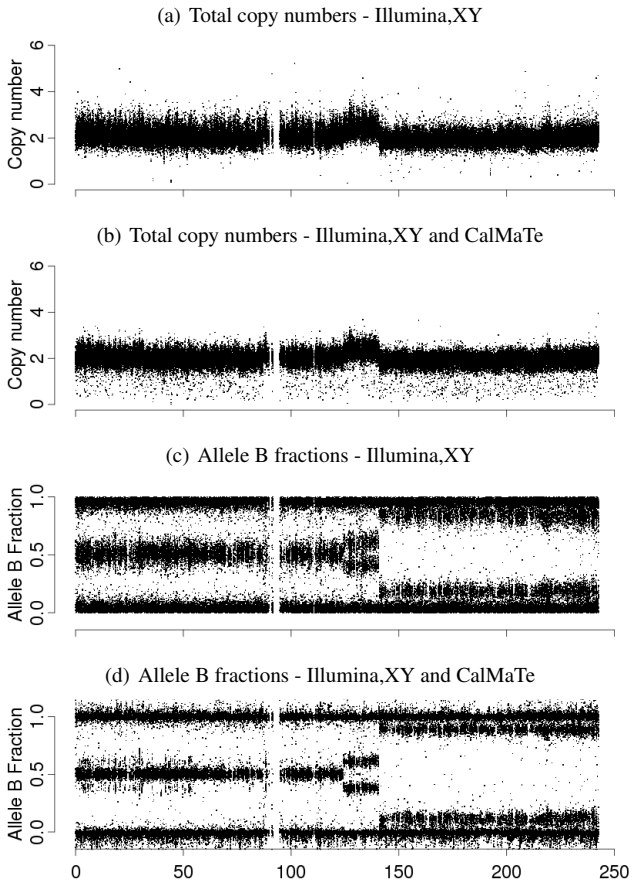


Fig. S4. TCNs and BAFs along Chr. 2 of TCGA ovarian cancer TCGA-23-1027 with and without CalMaTe. Data are from Illumina-processed Illumina Human1M-Duo arrays. Panels (a) and (b) show the TCNs before and after CalMaTe. Panels (c) and (d) show the corresponding BAFs. In this particular chromosome three different regions can be distinguished: a normal (i) region (in 0-120Mb), (ii) a gain at 3 copies (in 120-140Mb), and (iii) a neutral copy number LOH (140Mb till the end of the chromosome). The reason for observing four tracks in the LOH region instead of two is because of normal contamination. After applying CalMaTe, the tracks corresponding to different ASCNs are also more tightly packed, e.g. the four tracks of the gain are better distinguished using CalMaTe. If the BAF signals for CalMaTe were truncated at 0 and 1 the difference would be even more apparent. CalMaTe's TCN estimates have less dispersion but the algorithm has also introduced some outliers with low copy numbers.

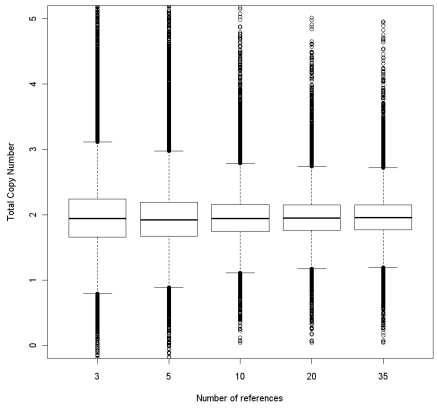
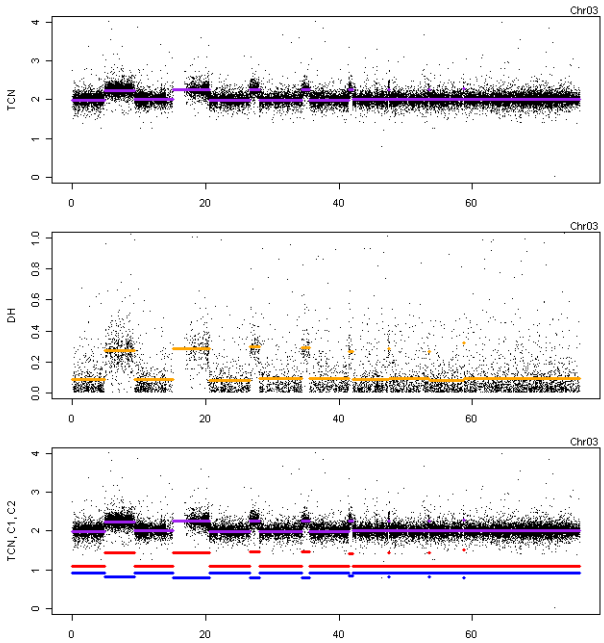


Fig. S6. Boxplot of the TCN estimates with CalMaTe using different number of references (same sample as in Figure S5). The TCN estimates improve, as the number of references increases. In none of the five studies, the normal sample GSM492495_IC022N has been included in the set of references. The samples were hybridized on Affymetrix GenomeWideSNP_6 arrays and preprocessed using CRMA v2.

(a) PSCN after CRMA v2 and CalMaTe



(b) PSCN after CRMA v2

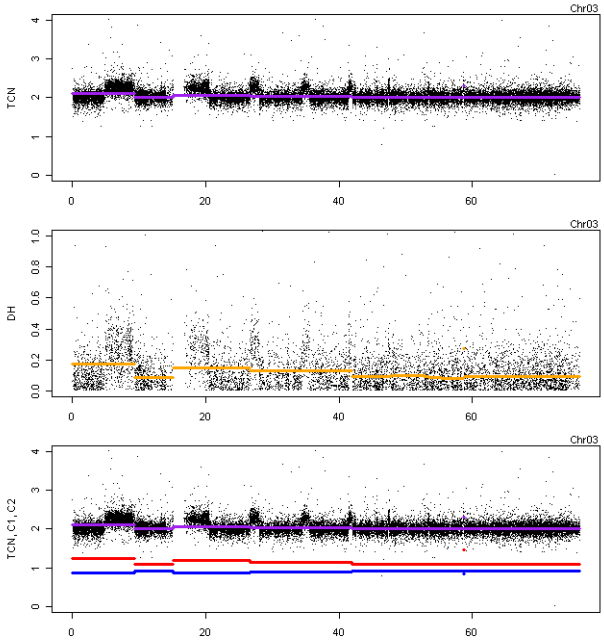
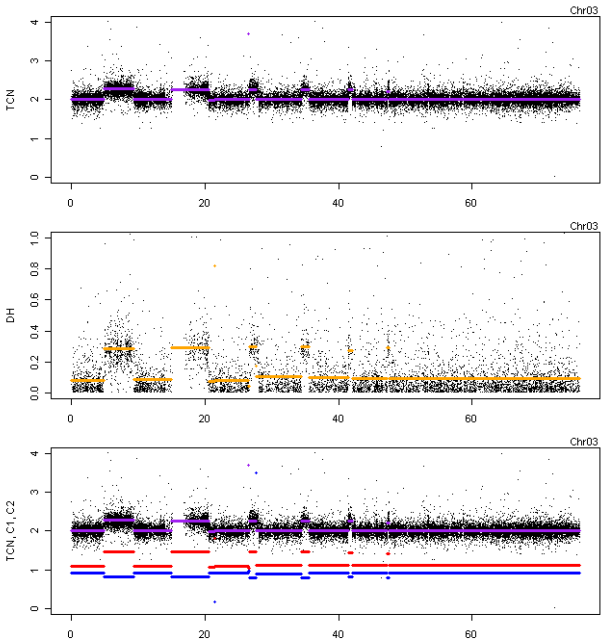


Fig. S7. Total copy numbers, Decrease of heterozygosity and major and minor allele copy numbers using PSCN as segmentation method applied with CRMA v2 & CalMaTe (upper) and CRMA v2 (lower). If CalMaTe is used the PSCN segmentation algorithm detects the eight segments. In other simulations the smallest segment (that spans 30 different probes) was missed. Without CalMaTe, the detected segments merge zones with different copy numbers.

(a) PSCBS after CRMA v2 and CalMaTe (without TumorBoost)



(b) PSCBS after CRMA v2 and TumorBoost

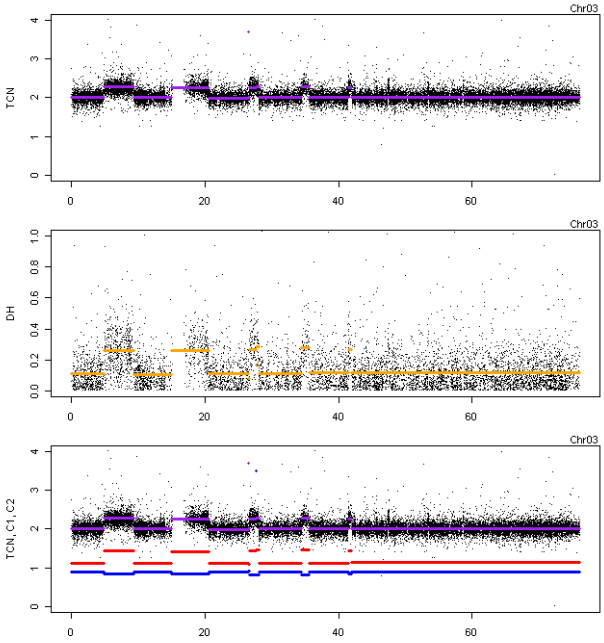


Fig. S8. Total copy numbers, Decrease of heterozygosity and major and minor allele copy numbers using PSCBS segmentation applied using CRMA v2 & CalMaTe (upper) and CRMA v2 & TumorBoost (lower). PSCBS was applied using default parameters. Using CalMaTe, it identifies six out of the eight segments. If CalMaTe is not used, only five segments are identified.

REFERENCES

Bengtsson, H., Simpson, K., Bullard, J., and Hansen, K. (2008). 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., Wirapati, P., and Speed, T. (2009a). A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**(17), 2149–2156.

Bengtsson, H., Ray, A., Spellman, P., and Speed, T. P. (2009b). A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics*, **25**(7), 861–867.

Bengtsson, H., Neuvial, P., and Speed, T. (2010). TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, **11**(1), 245.

Castro, P., Creighton, C., Ozen, M., Berel, D., Mims, M., and Ittmann, M. (2009). Genomic profiling of prostate cancers from African American men. *Neoplasia*, **11**(3), 305–12.

Chen, H., Xing, H., and Zhang, N. (2011). Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Computational Biology*, **7**(1), e1001060.

Illumina (2007). *BeadStudio Genotyping Module v3.2 - User Guide*. Illumina Inc. Part no: 11284301.

Lin, M., Wei, L., Sellers, W., Lieberfarb, M., Wong, W., and Li, C. (2004). dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**(8), 1233.

Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P., Olshen, R. A., and Seshan, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, **27**(15), 2038–2046.

Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Göansson, H., Juliusson, G., Rosenquist, R., Höglund, M., Borg, A., and Ringnér, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol*, **9**(9), R136.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.

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