CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation

Maria Ortiz-Estevez, Ander Aramburu, Henrik Bengtsson, Pierre Neuvial, Angel Rubio

To cite this version:


HAL Id: hal-01200761
https://hal.archives-ouvertes.fr/hal-01200761
Submitted on 18 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Genome analysis

CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation

Maria Ortiz-Estevez\textsuperscript{1,2,†}, Ander Aramburu\textsuperscript{1,†}, Henrik Bengtsson\textsuperscript{3,4,†}, Pierre Neuvial\textsuperscript{3,5,†} and Angel Rubio\textsuperscript{1,*,†}

\textsuperscript{1}CEIT and TECNUN, University of Navarra, 20018 San Sebastian, Spain, \textsuperscript{2}Biology Group, Celgene Institute for Translational Research (CITRE), 41092 Sevilla, Spain, \textsuperscript{3}Department of Statistics, University of California, Berkeley, CA 94720, USA, \textsuperscript{4}Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, 94107 USA and \textsuperscript{5}Laboratoire Statistique et Génome, Université d’Evry Val d’Essonne, UMR CNRS 8071 - USC INRA, 91037 Evry, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: CalMaTe calibrates preprocessed allele-specific copy number estimates (ASCNs) from DNA microarrays by controlling for single-nucleotide polymorphism-specific allelic crosstalk. The resulting ASCNs are on average more accurate, which increases the power of segmentation methods for detecting changes between copy number states in tumor studies including copy neutral loss of heterozygosity. CalMaTe applies to any ASCNs regardless of preprocessing method and microarray technology, e.g. Affymetrix and Illumina.

Availability: The method is available on CRAN (http://cran.r-project.org/) in the open-source R package calmate, which also includes an add-on to the Aroma Project framework (http://www.aroma-project.org/).

Contact: arubio@ceit.es

Supplementary information: Supplementary data are available at Bioinformatics online.

Version: 2012-03-23 16:28; Received on November 23, 2010; Revised on March 21, 2012; Accepted on April 23, 2012

\section{INTRODUCTION}

Several analytical pipelines for identifying total copy number (TCN) events from DNA microarrays are available. However, certain types of genomic alterations cannot be detected from TCNs, e.g. copy neutral LOH events. The identification of such events can be key to our biological understanding of cancer development and our ability to set up a personalized treatment plan (Albertson et al., 2003). Genotyping microarrays (Affymetrix Inc., 2007; Peiffer et al., 2006) quantify not only TCNs but also allele-specific copy numbers (ASCNs), which are necessary to identify CN states such as copy neutral heterozygosity LOH. ASCNs are the CN estimates of each allele variant (here A and B) at a particular (bi-allelic) single-nucleotide polymorphism (SNP). For the purpose of displaying ASCNs along the genome and also for detecting CN changes, ASCNs are often represented by their TCNs and B-allele fractions (BAFs) (Bengtsson et al., 2010). For instance, for diploid SNPs in a normal region, the expected TCN is 2 and expected BAFs are 0 (AA), 1/2 (AB) or 1 (BB). In a region of copy neutral LOH, the expected TCN is 2 and the expected BAFs are 0 or 1. For a single-copy gain, expected TCN is 3 and expected BAFs are 0, 1/3, 2/3 and 1. In Figure 1, observed TCNs and BAFs are displayed along the genome for a normal region, a gain and a region of copy neutral LOH.

Based on these type of data, segmentation methods (Chen et al., 2011; Olshen et al., 2011; Staaf et al., 2008; van Loo et al., 2010) identify regions of constant CN state. Their performances depend greatly on the signal-to-noise ratios (SNRs) of TCN and BAF (Bengtsson et al., 2011), which in turn depend on the preprocessing method used, e.g. dChip (Lin et al., 2004), CN5 (Affymetrix Inc., 2007), CRMA v2 (Bengtsson et al., 2004), ACNE (Ortiz-Estevez et al., 2011) and Illumina (Peiffer et al., 2006). Some of these...
methods perform better than others. It would be favorable to borrow strength between them, but in practice it is impossible because the preprocessing methods are developed specifically for a given SNP platform or chip generation. In addition, methods may also be proprietary, making it infeasible for researchers to improve upon them. A more sustainable solution for improving SNRs is to instead develop normalization methods that operate on the ASCN output of the aforementioned methods. For matched tumor-normal SNP microarray experiments, Bengtsson et al. (2010) provide the platform-independent TumorBoost method, which significantly improves the BAFs of the tumor. Here, we propose CalMaTe (for Calibration Matrix T), which to our knowledge is the only ASCN processing pipeline that is open source, cross-platform and that does not require matched normals.

2 METHOD AND RESULTS

CalMaTe is a platform-independent multi-array method that controls for SNP-specific systematic variation by modeling the crosstalk between alleles in bi-allelic SNPs as explained next. Non-polymorphic loci (on recent SNP array platforms) are normalized by a robust average across samples.

CalMaTe SNP model. The main assumption of CalMaTe is that crosstalk between alleles is linear and possibly different between SNPs but preserved across samples. Consider a SNP $j = 1, \ldots, J$ and let $H_j$ be the $2 \times 1$ matrix with column vectors $(c_{0j}, c_{1j})^T$ of the unobserved true ASCNs across all samples $i = 1, \ldots, I$. The corresponding observed ASCNs $\tilde{H}_j$ can then be modeled as

$$\tilde{H}_j = W_j H_j + \epsilon_j,$$

where $W_j$ is an unknown $2 \times 2$ crosstalk matrix shared by all samples, and $\epsilon_j$ is a $2 \times 1$ error matrix. This model and its estimation are outlined below and explained in more detail in the Supplementary Material.

Estimating the crosstalk. $W_j$ can be estimated from a set of normal samples (e.g., $W_j^{\text{ref}}$ for which the ASCNs (genotypes) are known, e.g.,

$$W_j^{\text{ref}} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 1 & 2 \\ 0 & 1 \end{bmatrix},$$

where $(2,0,0,1,1,2,0,2)$ correspond to genotypes AA, AB and BB. The set of possible states in $W_j^{\text{ref}}$ is discrete and small, which is why it is possible to estimate $W_j$. There is no such constraint on $H_j$, which is key when analyzing non-homogeneous samples such as tumors. Given genotypes $H_j^{\text{ref}}$ and observed $H_j$, an estimate $\tilde{W}_j$ is obtained by robustly solving Equation (1) for $W_j$. CalMaTe calls the genotypes from $H_j^{\text{ref}}$ using a naive genotyping algorithm. To minimize the impact of batch effects, CalMaTe normalizes ASCNs from any technology and across all samples

$$(I - C_{12}) (H_j - \tilde{H}_j) = \epsilon_j,$$

where $C_{12}$ is the $2 \times 2$ calibration matrix shared by all samples, and $\epsilon_j$ is an unknown $2 \times 2$ error matrix. The references samples are ideally in the same batch as the other samples. If normal samples are unavailable, all samples can be used as a reference. As long as the majority of the reference samples are normal at any given SNP (not necessarily the same samples for all SNPs), the robustness of the estimator warrants a good $\tilde{W}_j$ estimate. We recommend to use 6 or more reference samples (see also Supplementary Material).

Calibration of ASCNs. Given an estimate $\tilde{W}_j$, the calibrated ASCNs, $\tilde{H}_j$, are obtained from Equation (1) as the back-transformation

$$\tilde{H}_j = \tilde{W}_j H_j,$$

where $\tilde{W}_j = W_j^{\text{ref}}$. The $2 \times 2$ calibration matrix $C_{12}$ improves the SNRs, as confirmed by extensive ROC analysis (Fig. 2B and Supplementary Material), which makes it easier for segmentation methods to distinguish between different CN states. Similar improvements were also achieved for ASCNs from dChip as well as ASCNs originating from Illumina, as shown in the Supplementary Material.

3 CONCLUSIONS

CalMaTe normalizes ASCNs from any technology and preprocessing method, and without requiring matched normals. The normalized ASCNs are on average more accurate, resulting in greater SNRs along the genome. This enhances the power to detect alterations such as LOH using segmentation methods. The method is readily available in an open-source open-access R package, which provides a low-level API for incorporating CalMaTe in third-party solutions, as well as a high-level API for the Aroma Project framework (http://www.aroma-project.org/) making it possible to analyze an unlimited number of arrays.

Funding: HB & PN were supported by NCI grant U24 CA126551.

Conflict of Interest: none declared.

REFERENCES


