TumorBoost: A single-sample method for calibrating allele-specific tumor copy numbers in paired tumor-normal designs

Henrik Bengtsson*[1], Pierre Neuvial[1] and Terence P. Speed[1,2]

[1] Department of Statistics, University of California, Berkeley, USA
[2] Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia

Email: Henrik Bengtsson*- hb@stat.berkeley.edu; Pierre Neuvial -pierre@stat.berkeley.edu; Terence P. Speed - terry@stat.berkeley.edu;

*Corresponding author

This version: 2009-04-30 13:00

Abstract

Background: High-throughput genotyping microarrays can be used not only to assess changes in total DNA copy number but also changes in allele-specific copy numbers (ASCNs). Even after state of the art preprocessing methods, ASCN estimates for Affymetrix genotyping arrays still suffer from systematic effects that make them difficult to use effectively for downstream studies in cancers.

Results: We propose a single-sample method for calibrating ASCN estimates of a tumor based on ASCN estimates of a paired normal. The method applies to any paired tumor-normal estimates regardless of technology and generation. We demonstrate that our method leads to a much clearer separation between different ASCN states, including *copy number neutral events* that cannot be detected using total copy numbers only.

Conclusions: Combined with single-array preprocessing methods, such as CRMA v2, we conclude that ASCN estimates with high precisions can be obtained from a single pair of tumor-normal hybridizations, and recommend using paired tumor-normal DNA microrarray experiments when applicable.

Availability: A single-sample bounded-memory implementation is available in *aroma.cn*.

PN:

- single "sample" or single "individual" method ?

TODO:

- test the method on Illumina data.

Background

The development of microarray technologies to assess DNA copy number changes over the past few years has been triggered by the fact that genomic alterations are hallmarks of gene deregulation and genome instability in cancers [1, 2].

Recent technologies include genotyping microarrays that quantify not only total copy numbers (TCN) but also allele-specific copy numbers (AS-

CNs), that is, allelic contributions of each allele to TCN. ASCNs estimates are crucial as they can pinpoint genomic alterations that are TCN neutral, such as uniparental disomy (UPD), or allele-specific amplifications.

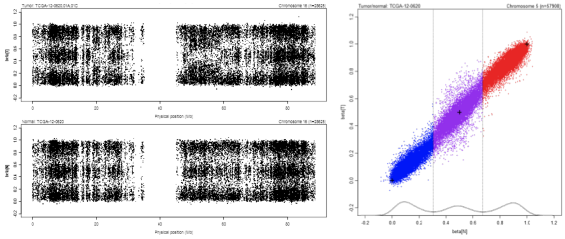
TODO: Technologies: Illumina and Affy.

Several preprocessing methods have been developped for these technologies [3–7]. **TODO:** REFs for Illumina preprocessing ? The typical output of genotyping microarray preprocessing method for a given SNP unit is an estimation of the total copy number — the sum of allele A and allele B intensities divided by a reference — and the fraction β of allele B intensity relative to the total intensity. Together, these two quantities determine the raw ASCNs at each SNP locus.

SNP effects after preprocessing

The main motivation of this paper is that existing preprocessing methodsdo not correct for SNP-specific effects, that result in SNP-specific distributions of genotype clusters. As a consequence, β estimates are not comparable across probes, as illustrated in Figure XXX.

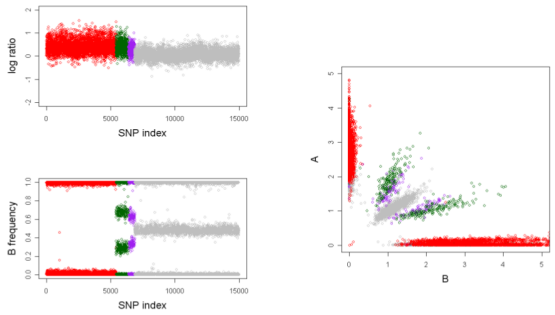
TODO: Update these figures



Fraction of allele B (β) in a tumor and its matched normal on chrosomosome XXX. Each point corresponds to a SNP. Left: β along chromosome XXX for the tumor (top) and the normal (bottom). Right: Scatter plot of β in the tumor vs β in the normal. Black '+' indicate the expected location of normal genotype clusters: $(0,0)$, $(1/2,1/2)$ and $(1,1)$. Data were preprocessed using the CRMA v2 method [4].

There is considerable deviation from the expected location of genotype clusters, even in the normal when no (or very few) genomic alterations are expected (Figure XXX, right). Interestingly, this deviation is quite reproducible between the tumor and the normal: most points are scattered along the diagonal. As a consequence of this deviation, genomic profiles of allele B fraction are quite noisy both in the tumor and the normal (Figure XXX, right).

PN: Shall we add a second illustration to illustrate the motivation for downstream analyses, ie (θ_A, θ_B) for a given sample across SNPs in a region as in Nancy's talk ?



Proposed method

In this paper we present a method calibrating the allele-specific copy-number estimates (ASCNs) of a tumor tissue given ASCNs of matched normal tissue or blood extract. The method does not require external references and it is only the relative ASCNs that are calibrated: total CN estimates are neither used as an input nor adjusted.

We show that our method leads to a much clearer separation between different ASCN states, including *copy number neutral events*. Such events cannot be detected using total copy numbers only; however, detecting them is crucial in cancer studies as they can help finding new tumor suppressor genes.

Genotypes *of the normal* hybridization are used for the calibration, hence the performance of our method depends on genotype quality. We compare the results obtained using genotypes inferred using a set of samples to those of a naive single-array genotyping **PN:** add "based on local minima of the density of β across SNPs" ? . We argue that the proposed naive genotyping algorithm yields calibrated β values that are good enough for typical downstream analyses such as the search of regions of Loss of Heterozygosity (LOH) or ASCN segmentation, as all these analyses involve smoothing or segmentation of β along the genome. The calibration method obtained using this genotyping algorithm is therefore a purely *single-sample method*.

A single sample method

The realization of a single-sample method has several implications: (i) Each tumor-normal pair can be analyzed immediately without needing reference samples. (ii) Samples can be processed in parallel on different hosts/processors making it possible to decrease the processing time of large data sets. (iii) There is no need to reprocess a sample when new samples are produced, which further saves time and computational resources. Furthermore, (iv) the decision to filter out poor samples can be made later, because a poor sample will not affect the processing of other samples. More importantly, a single-sample method is (v) more practical for applied medical diagnostics, because individual patients can be analyzed at once, even when they come singly rather than in batches. This may otherwise be a limiting factor in projects with a larger number of samples.

Outline

The outline of this paper is as follows. In Methods, we describe the underlying model, its estimation, and an interpretation in terms of allelic crosstalk. In Results, we show that the signal-to-noise ratios (SNRs) of the calibrated ASCNs are significantly larger than corresponding non-calibrated estimates. In Discussion, we conclude the study, discuss potential limitations, ..., and give future research directions.

Methods

Let $(\theta_{N,j,A}, \theta_{N,j,B})$ be the signal intensities after preprocessing for SNP j in the Normal (N) tissue, and $(\theta_{T,j,A}, \theta_{t,j,B})$ the corresponding intensitites in the tumor (T) tissue. The allele B fraction [8] for SNP $j = 1, \ldots, J$ in the nnormal (N) tissue is defined as:

$$\beta_{N,j} = \frac{\theta_{N,j,A}}{\theta_{N,j}},$$

where $\theta_{N,j} = \theta_{N,j,A} + \theta_{N,j,B}$ is the non-polymorphic signal at locus j. For a diploid SNP j, for which the genotype is either AA, AB or BB, we expect the allele B fraction $\beta_{N,j}$ to be fall close to 0, 1/2, or 1, respectively. The allele B fraction $\beta_{T,j}$ for the tumor (T) tissue/blood extract is defined analoguously. Because we cannot expected a SNP in a tumor to be diploid at any random SNP, we cannot make any prediction on where $\beta_{T,j}$ falls. For a tumor-normal pair, we observe $(\beta_{N,j}, \beta_{T,j})$ at each SNP j.

Model and estimation

Consider a SNP j that is diploid in the normal tissue. Let the true genotype be denoted by the true allele B fraction $\mu_{N,j}$ with possible states $\{0, 1/2, 1\}$ corresponding to genotypes $\{\mathrm{AA}, \mathrm{AB}, \mathrm{BB}\}$.

Based on this, for SNP j we model the observed allele B fraction for the tumor and the normal pair as:

$$\begin{aligned} \beta_{T,j} &= \mu_{T,j} + \delta_j + \varepsilon_{T,j} \\ \beta_{N,j} &= \mu_{N,j} + \delta_j + \varepsilon_{N,j} \end{aligned}$$

where δ_j is a SNP-specific effect and $\varepsilon_{T,j}$ and $\varepsilon_{N,j}$ are independent zero-mean error terms. The difference $\mu_{T,j} - \mu_{N,j}$ can then be estimated straightforwardly from the data.

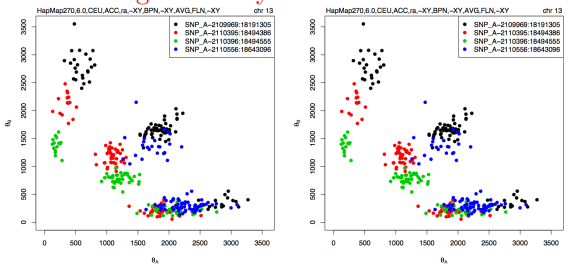
$$\widehat{\mu_{T,j} - \mu_{N,j}} = \beta_{T,j} - \beta_{N,j} \,.$$

Model interpretation in terms of allelic crosstalk

In [5] and [4] it is argued that there exist crosstalk between the allele signals. For instance, for a diploid SNP that is truly AA we will not only observe a great signal in the $\mathrm{PM_A}$ probes, but also some signal in the $\mathrm{PM_B}$ probes. One explanation for this is crosshybridization due to the close similarity of probe sequences. In the afformentioned studies, the authors propose an offset and crosstalk correction that is applied globally (to the six different heterozygous groups). In addition, [4] suggest to apply an additional nucleotide-position probe sequence normalization, which further corrects for imbalances between the two alleles. It is shown that these corrections significantly improve the ability to detect total CN changes. However, when looking at the distribution of the allele-specific summaries for a particular SNP j across a set of samples, that is $\{(\theta_{ijA}, \theta_{ijA})\}_i$, it is clear that for some SNPs there exists a remaining crosstalk, which is likely to be SNP specific. In Figure XXX, the allele-specific summaries for SNP_A-XXXXXX in data set XXXXXX are shown, which clearly shows that the two homozygous genotype groups AA and BB are not located along the axes, that is, they are not orthogonal. For this particular SNP the heterozygous group is located along the diagonal as expected.

TODO: Illustration: something like the following (θ_A, θ_B) plots across samples for two or three consecutive SNPs (maybe with CNA, CNB instead

to avoid the discussion about the copy number scale ?). Left: normal and tumor before calibration. Right: same with the calibrated tumor. Argue that the angle is really different from SNP to SNP.



PN: I can draw a couple of such plots for a few units of interest. We propose the following crosstalk model for estimated allele-specific summaries $\{(\theta_{ijA}, \theta_{ijA})\}_i$:

$$\boldsymbol{\theta}_{ij} = \mathbf{S}_i \mathbf{x}_{ij} + \boldsymbol{\varepsilon}_{ij},$$

where the crosstalk matrix

$$\mathbf{S}_i = \begin{bmatrix} S_{iAA} & S_{iAB} \\ S_{iBA} & S_{iBB} \end{bmatrix}$$

is shared by all SNPs of sample i.

Evaluation methods

In order to assess the performance of TumorBoost, we compare the ability of the fraction of allele B to detect allele-specific genomic alterations before and after calibration. More specifically, we selected one tumor-normal pair for which we identified a region of uni-parental disomy (UPD, a.k.a. copy number neutral LOH) in the tumor, that is, a region in which the total copy number is 2 and there are homozygous and heterozygous loci in the normal, but there are only homozygous loci in the tumor. Importantly, such a region cannot be identified by analyzing total copy number only as it is not accompanied by any change in total copy number.

Quantifying UPD

In order to identify the start and end breakpoints of this region, we focus on SNPs that are *heterozygous in the normal* and use a transformed version of the fraction of allele B, denoted by ρ and defined by

$$\rho = \left| \beta - \frac{1}{2} \right| .$$

This kind of transformation is widely used for downstream analyses **TODO:** REFS: Peiffer ? Attiyeh ? probably all Illumina SNP papers. and can be motivated as follows for the particular case of a region of UPD. In such a region, β is expected to have only two bands, corresponding to the two homozygous states, whereas it has three bands in a normal region. As β is symmetric around $a/2$, ρ has only one band near 0 in an UPD region, and two bands in a normal region. After excluding SNPs that are homozygous in the normal, ρ also only has one band in a normal region, which is expected to be near $1/2$. **PN:** Painful explanation... draw a picture !

PN: <sidetracking> In order to interpret ρ it is useful to note that it can be written as the minimum allele fraction:

$$\rho = \beta \wedge (1 - \beta) ,$$

or, equivalently, as the minimum ASCN, rescaled by the total copy number

$$\rho = (CN_A \wedge CN_B) \times (CN_A + CN_B) .$$

PN: ASCNs and TCN have not been formally defined yet. </sidetracking>

Data points in the two XXXkb regions centered around the start and the end position of the UPD region were excluded. The remaining data points are annotated to belong to either the normal state or the UPD state. We use a Receiver Operator Characteristics (ROC) analysis to assess how well the fraction of allele B separates between the neutral and the UPD data points.

Resolution

This evaluation is done on the full-resolution β as well as on a smoothed version, where βs are smoothed by using non-overlapping bins for which the average β is calculated. This approach is motivated by the fact that downstream analyses of ASCN use smoothing or segmentation and are therefore concerned with the influence of the amount of smoothing on the output of the analysis. It was inspired by total copy number studies [9,10]. A similar approach has been used more recently in [4,5]

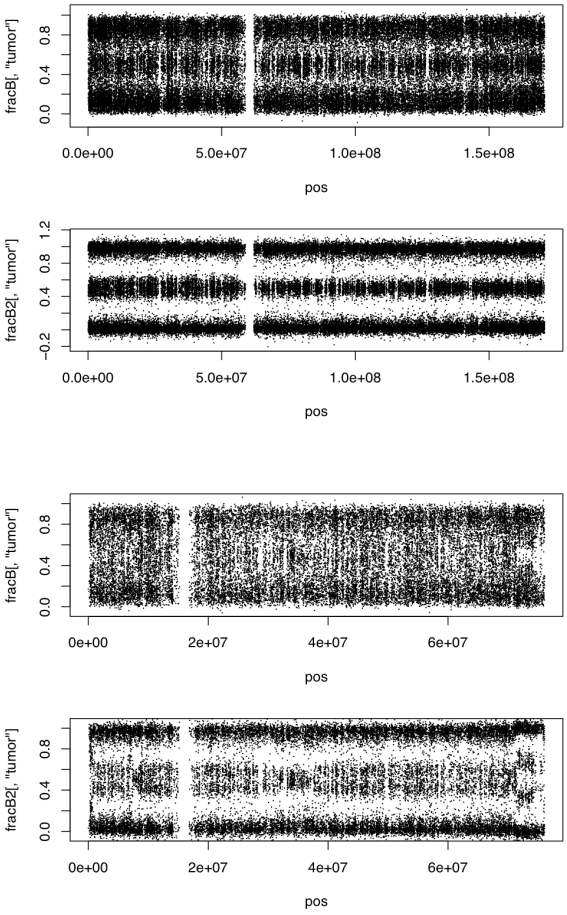
Show sensitivity to genotype calls by comparing results with naive genotyping, TGCA genotypes, and "random" genotypes.

Results

Data set

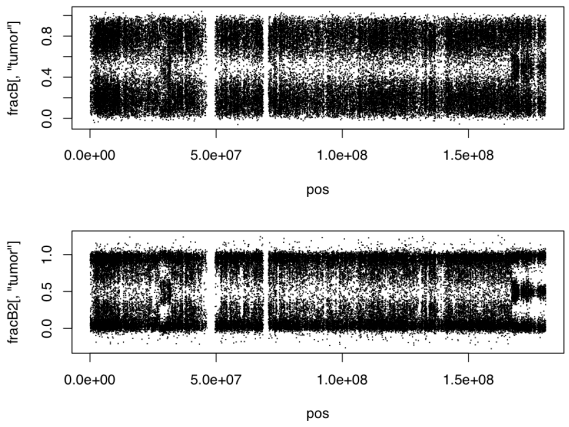
We used data from the Cancer Genome Atlas (TCGA) project [11,12], a collaborative initiative to better understand cancer using existing large-scale whole-genome technologies. Several tumor types are or are planned to be studied, including brain cancer (glioblastoma multiforme; GBM), ovarian cancer and lung cancer.

From the Data Coordinating Center (http://tcga-data.nci.nih.gov/tcga/homepage.htm), we downloaded (Jan 2009) raw data (Level 1; CEL files) for a set of GBM tumor-normal pairs. For the purpose of illustrating our method, we will focus mainly on sample TCGA-02-0104 (vials 01A vs 10A), because it has a large number of CN aberrations on Chr 3 at different mean levels.



Allele B fraction before and after calibration

PN: Add (θ_A, θ_B) plots for each sample in a right panel (colored by region ?) and/or (β_T^*, β_N) plots to each of the next 3 plots in order to get another sense of the improvement ?



Detecting copy number neutral LOH

We use one pair of tumor/normal hybridizations to assess how well ASCN estimates (or ρ estimates) can differentiate between normal and UPD states. Given a global threshold t, a locus with ρ below t is considered to belong to the UPD state, otherwise the normal state. By calculating the fraction of correctly called UPD loci, we obtain an estimate of the true positive rate, and by calculating the fraction of incorrectly called normal loci, we obtain an estimate of the false positive rate. By adjusting the threshold, we can estimate a ROC curve.

Full resolution β

The true positive rate of calling a UPD (among normal loci) as a function of false-positive rate is depicted in Figure XXX. The ROC curves show that...

Smoothed β
Using a windowing technique similar to that in [4,13] for a fixed false-positive rate we can estimate the true positive rate as a function of amount of smoothing. Since a given amount of smoothing corresponds to a given distance between loci this provides us with a first approximation to the effective resolution of a method. In the upper panel of Figure XXX the true-positive rate (for normal v. UPD) as a function of resolution is shown before and after calibration.

Discussion

Downstream analysis methods

= increasing power to detect copy number events, and making it possible to detect copy number neutral events from Affymetrix genotyping arrays, e.g.:

- Segmentation of ASCN when normal genotypes are available
- Iterative segmentation of ASCN when normal genotypes are not available
- Quantile segmentation of ASCN when normal genotypes are not available

When genotype calls for the normal sample are available they can be used to segment tumor ASCNs using any copy number segmentation algorithm. When these genotype calls are not avaiable we propose two algorithms that achieve joint segmentation of normal and tumor ASCNs, both leading to genotype calls for the normal sample as a by-product.

Genotyping errors

In the Results section we have shown that our calibration methods leads to an improved signal ratio at the chromosome or at the genome scale. However the correction factor is genotype-specific, so if the normal genotype call for a given SNP is wrong, the correction factor will actually add bias to the estimated fraction of allele B. We argue that this is not a serious issue, for two main reasons.

First, we can take genotype *confidence scores* into account, either by discarding SNPs for which we are not confident in the genotype or by using the confidence scores in whatever downstream analyses. **TODO:** in suppl. mat.: a section with beta along the genome after calibration for different levels of stringency for genotype calls (for the naive genotyping)

Second, although our method does improve allele-specific copy number calls at the single locus level whenever normal genotype calls are correct, our goal is to improve the SNR of the fraction of allele B *along the genome* in order to facilitate downstream analyses. In the results section we showed that our calibration method leads to a significant improvement of this SNR regardless of the genotype calling algorithm chosen.

Extensions of the method

- applicability / usefulness for Illumina data. paragraph on why Affy is more noisy than Illumina ? Sequences are all the same for Illumina, hence no probe sequence specific effects.
- multi source version
- multi sample (unpaired) version ?

Conclusions

Text for this section ...

Algorithm and implementation

The TumorBoost calibration method is available in R [14] package *aroma.cn* part of the *aroma.affymetrix* framework [15]. The method is designed and implemented to have bounded-memory usage, regardless of the number of samples/arrays processed. Furthermore, the complexity of the algorithm is linear in the number of loci (J). Since it is a single-sample method, the tumor-normal pairs can be calibrated in parallel on multiple hosts/processors. The method applies to estimates obtained by any SNP technology.

Authors contributions

Text for this section ...

Acknowledgements

We gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing samples, tissues, data processing, and making data and results available.

Funding: NCI grant U24 CA126551.
Conflict of interest: none declared.

References

1. Albertson DG, Collins C, McCormick F, Gray JW: **Chromosome aberrations in solid tumors**. *Nat. Genet.* 2003, **34**:369–76.

2. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57–70.

3. Affymetrix Inc: *Affymetrix Genotyping Console 3.0 - User Manual*. Affymetrix Inc. 2008.

4. Bengtsson H, Wirapati P, Speed TP: **A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6**. *resubmitted* 2009. [(resubmitted)].

5. Bengtsson H, Irizarry RA, Carvalho B, Speed TP: **Estimation and assessment of raw copy numbers at the single locus level.** *Bioinformatics* 2008, **24**(6):759–767, [http://dx.doi.org/10.1093/bioinformatics/btn016].

6. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.** *Nature Genet.* 2008, **40**(10):1253–1260, [http://dx.doi.org/10.1038/ng.237].

7. Li C, Wong W: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc. Natl. Acad. Sci. USA* 2001, **98**:31–6.

8. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL: **High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.** *Genome Res.* 2006, **16**(9):1136–1148, [http://dx.doi.org/10.1101/gr.5402306].

9. Lai WR, Johnson MD, Kucherlapati R, Park PJ: **Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data.** *Bioinformatics* 2005, **21**(19):3763–3770, [http://dx.doi.org/611].

10. Willenbrock H, Fridlyand J: **A comparison study: applying segmentation to array CGH data for downstream analyses.** *Bioinformatics* 2005, **21**(22):4084–4091, [http://dx.doi.org/10.1093/bioinformatics/bti677].

11. Collins FS, Barker AD: **Mapping the Cancer Genome**. *Scientific American* 2007, **296**(3):50–57, [http://www.sciam.com/article.cfm?id=mapping-the-cancer-genome&print=true].

12. TCGA Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061–1068, [http://dx.doi.org/10.1038/nature07385].

13. Bengtsson H, Irizarry RA, Carvalho B, Speed TP: **Estimation and assessment of raw copy numbers at the single locus level**. *Bioinformatics* 2008, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn016v1].

14. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2008, [http://www.R-project.org]. [ISBN 3-900051-07-0].

15. Bengtsson H, Simpson K, Bullard J, Hansen K: **aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory**. Tech. Rep. 745, Department of Statistics, University of California, Berkeley 2008.