# Statistical Estimation of Genomic Tumoral Alterations

Yi Liu, Christine Keribin, Tatiana Popova, Yves Rozenholc

# Statistical Estimation of Genomic Tumoral Alterations

Yi Liu [1] & Christine Keribin [2] & Tatiana Popova [3] & Yves Rozenholc [4]

[1] *Université Paris Descartes, UMRS-1147 (Médecine Personnalisée, Pharmacogénomique, Optimisation Thérapeutique)& INRIA Saclay Ile-de-France équipe Select, yi.liu0717@gmail.com,*

[2] *Université Paris Sud, Laboratoire de Mathématiques, Equipe Probabilités et Statistiques - UMR 8628 & INRIA Saclay Ile-de-France, équipe* Select, *christine.keribin@math.u-psud.fr*

[3] *Institut Curie, INSERM U830, tatiana.popova@curie.fr*

[4] *Université Paris Descartes, MAP5 - UMR CNRS 8145 & INRIA Saclay Ile-de-France, équipe* Select, *yves.rozenholc@parisdescartes.fr*

**Résumé.** La caractérisation des altérations génomiques tumorales est une étape importante dans le développement de la médecine personnalisée en cancérologie. Parmi les méthodes de traitement des données de micro-array, la méthode GAP (Genome Alteration Print) de Popova et al. (2009) caractérise les mutations à partir de la segmentation des signaux du nombre de copies et de la fréquence de l'allèle majoritaire obtenus en chaque site de SNP. Elle utilise un critère déterministe que nous proposons de remplacer par une modélisation probabiliste paramétrique. Nous définissons ainsi un modèle de mélange gaussien dont les classes caractérisent les types de mutations. Ce modèle est estimé par maximum de vraisemblance grâce à l'algorithme EM, permettant d'obtenir l'estimation des paramètres et la caractérisation de l'altération tumorale de chaque segment. Dans notre approche, la ploïdie de la tumeur est déduite de l'utilisation d'un critère pénalisé de sélection de modèle. Notre modèle est testé avec des données simulées et expérimentales.

**Mots-clés.** Modèle de mélange, algorithme EM, critère BIC, GAP, mutations tumorales, SNP micro-array.

**Abstract.** Characterization of the tumoral genomic alterations is an important step in the development of personalized medicine in cancerology. Among the methods for treating micro-array data, the GAP (Genome Alteration Print) method of Popova et al. (2009) characterizes the mutations based on the segmentation of copy number and B-allele frequency signals obtained on each SNP. It uses a deterministic criterion that we propose to replace by a parametric probabilistic model. In this way, we define a Gaussian mixture model whose classes characterize the mutation types. This model is estimated by maximum likelihood through the EM algorithm, allowing us to obtain the estimation of the parameters and the characterization of tumoral alterations on each segment. In our

approach, the tumoral ploidy is deduced from a penalized model selection criterion. Our model is tested on simulated data and real data.

**Keywords.** Mixture model, EM algorithm, BIC criterion, GAP, tumoral mutations, SNP micro-array.

# 1 Introduction

Recent research reveals that personalized medicine is arguably the best way to treat cancer because of, for example, the immense diversity of underlying genomic alterations. In order to develop personalized medicine, characterizing the genomic alterations is a vital component. One way to characterize this alteration is to use a Single Nucleotide Polymorphism (SNP) microarray. A SNP is a nucleotide showing variability in the population. In theory, there are four possible variations, however, in practice, only two variations are observed which are called A-allele and B-allele, one being common in a large part of the population. Since the chromosomes in human come in pairs, it is possible for a SNP to have the genotype AA, BB, AB, or BA. The two former cases are called homozygous SNP, and the two latter, which are indistinguishable, are called heterozygous SNP.

Using microarrays one can detect genomic alterations such as copy-number variation and allele-imbalance. Having at hand two microarrays, one for the tumor, the other for the normal tissue, one can get rid of the unknown proportion $p$ of normal tissues in the tumor sample which acts as misleading parameter in the tumoral genotype estimation. However, clinicians are expecting to retrieve this information from only one microarray of tumor sample. Several methods have been developed for this goal for Illumina platforms. They basically fall into two groups. The first group simultaneously integrates the segmentation and characterization into one single step by using a hidden Markov model (HMM). GenoCNA by Sun et al. (2009), OncoSNP by Yau et al. (2010), and GPHMM by Sun et al. (2011) are several examples of this group. The other group separates the segmentation and characterization into two separate steps. Examples of this approach include GAP by Popova et al. (2009) and ASCAT by Loo et al. (2010). Mosén-Ansorena et al. (2012) compared between these two groups and found that in general, the two-step approaches have better performance. The method GAP uses an empiric method to estimate the mutation types and other parameters such as the proportion of normal tissues. Here we develop a probabilistic model to estimate statistically the parameters and the mutation types of each segment. Note that Liu et al. (2014) have recently developed an HMM method called TAFFYS for the other platform Affymetrix.

# 2 Model

For a SNP, microarray measures the intensities $I_A$ and $I_B$ of the two alleles which are proportional to their effective number of copies $n_A^g$ and $n_B^g$: $I_A = \gamma n_A^g$ and $I_B = \gamma n_B^g$. From

the two intensities it is possible to derive two variables characterizing the copy-number and the allele imbalance of the SNP

$$\text{lrr} = \log_2\left(\frac{I_A + I_B}{I_{Ref}}\right)^{\alpha} = \alpha \log_2(CN) + \beta,$$
$$\text{baf} = \frac{I_B}{I_A + I_B} = \frac{n_B^g}{n_A^g + n_B^g},$$

where $CN = n_A^g + n_B^g$ is the copy-number, $\alpha$ the contraction factor due to experimental techniques, $\beta = \alpha \log_2(\gamma / I_{Ref})$ dependent on the sample ploidy, and $I_{Ref}$ is a reference intensity. By definition, baf is bounded between 0 and 1. Assume that the proportion of normal tissues is $p > 0$ in the biopsy and that the tumor cells have mutation type with copy-number $(n_A, n_B)$ for the two alleles, then for heterozygous SNP we have $n_A^g = p + (1-p)n_A$ and $n_B^g = p + (1-p)n_B$. It follows that

$$\text{lrr} = \alpha \log_2(2p + (1-p)(n_A + n_B)) + \beta, \quad \text{baf} = \frac{p + n_B(1-p)}{2p + (n_A + n_B)(1-p)}.$$

Measurement values are noisy and we observe on the SNP $m$

$$\text{LRR}_m = \text{lrr}_m + \eta\xi_m, \quad \text{BAF}_m = \text{baf}_m + \sigma\varepsilon_m$$

where $\xi_m$ and $\varepsilon_m$ are independent random variables with zero mean and unit variance, and $\eta$ and $\sigma$ two positive real numbers. We assume that $\sigma$ and $\eta$ do not depend on the SNP.

Structural genomic alterations occurring on intervals of the genome, the values $\text{lrr}_m$ and $\text{baf}_m$ are fixed on one homogeneous interval with respect to the mutation. Hence the two distributions of $\text{BAF}_m$ and $\text{LRR}_m$ can be considered piecewise constant as $m$ varies. It follows that the mutation characterization can be realized from a proper segmentation of the two distributions along the genome. Following Popova et al. (2009), we assume these segmentations have already been realized and we focus on the characterization part. On each segment, the BAF distributions are symmetric around 1/2, so we confine ourselves to the range of $[0.5, 1]$ by symmetry.

On the homogeneous segment $i$ of length $N_i$, are computed averaged values for the B-allele frequency and the LRR, namely, $\text{BAF}_i^0$ (resp. $\text{BAF}_i^1$) for the averaged BAF of the $N_i^0$ heterozygous (resp. $N_i^1$ homozygous) SNP, and $\text{LRR}_i$ as the average of the $N_i^0 + N_i^1 = N_i$ values of the LRR on the whole segment.

We assume these observations are independent and follow:

$$\begin{aligned}
\mathrm{BAF}_i^0 &= \mathrm{baf}_k^0 + \varepsilon_i^0 \frac{\sigma}{\sqrt{N_i^0}},\\
\mathrm{BAF}_i^1 &= \mathrm{baf}_k^1 + \varepsilon_i^1 \frac{\sigma}{\sqrt{N_i^1}},\\
\mathrm{LRR}_i &= \mathrm{lrr}_k + \xi_i \frac{\eta}{\sqrt{N_i}}
\end{aligned}$$

where $k$ is the class label indicating the underlying mutation type of the segment, which is characterized by $\mathrm{baf}_k^1$ and $\mathrm{lrr}_k$ as $\mathrm{baf}_k^0$ is always 1. Moreover, $\varepsilon_i^0$, $\varepsilon_i^1$ and $\xi_i$ are assumed to be independent standardized Gaussian variables thanks to the Central Limit Theorem. Notice that this is a weak assumption, as $N_i^j$ is generally much larger than tens. This leads to two bivariate observations $C_k^j := (\mathrm{BAF}_i^j, \mathrm{LRR}_i)$ around the theoretical positions $c_k^j := (\mathrm{baf}_k^j, \mathrm{lrr}_k)$, $j = 0, 1$, which are associated with the underlying mutation and which only depend on the unknown parameters $p$, $\alpha$, $\beta$.

Therefore the segmented observations follow a Gaussian mixture model with centers at fixed positions (see Figure 1) when $p$, $\alpha$ and $\beta$ are known. In this model, the likelihood of one observation is given by

$$f(C_i^j, N_i^j, N_i) = \sum_{\ell=1}^{L} \pi_\ell \, \phi(C_i^j; c_\ell, \Sigma),$$

where $\Sigma$ is the diagonal matrix with diagonal $(\sigma^2/N_i^j, \eta^2/N_i)$, $\phi$ is the bi-dimensional Gaussian density and $\pi_\ell$, $\ell = 1, \ldots, L$ is the mixing proportions, that is the probability for the observation to be emitted from the center $c_\ell$ which corresponds to one of the $c_k^j$, when $k$ varies and $j = 0, 1$ (while the notation $c_k^j$ keeps a trace of the connection between centers associated with one single mutation, $c_\ell$ represents an arbitrary order of all centers. One can think for example to the lexical order).

# 3 Maximum likelihood estimation

We denote by $n$ be the number of interval of the segmentation. We use a maximum likelihood approach to estimate the parameters of our model, namely $p$ the proportion of normal tissues, $\alpha$ and $\beta$ linking the copy number to the LRR, $\eta$ and $\sigma$ the standard deviations, together with the mixing proportions $\pi_\ell$, $\ell = 1, \ldots, L$.

To this aim, we introduce the membership component indicator $z_{i\ell}$, $i = 1, 2, \ldots, n$, whose value is 1 if the observation $i$ is emitted from the underlying center $c_\ell$, 0 otherwise. The parameter $p$ is tricky to infer simultaneously with the other parameters, hence we use a two-level strategy: for a given $p$, we implement an EM algorithm (Dempster et al. 1977) to maximize the likelihood when the parameter $p$ is fixed and estimate the other parameters, and then, use gradient descent method to find the optimal value of $p$.

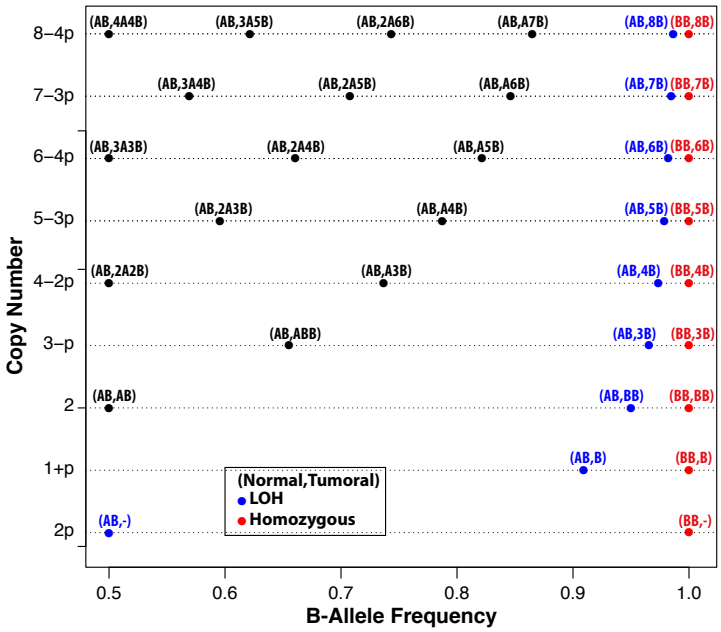


Figure 1: Schematic illustration of the correspondence between the tumoral mutation and the $(\mathrm{baf}, \mathrm{lrr})$ values. Mutations of germ line homozygous are in red. Mutations of germ line heterozygous are in black and blue. The mutation of the latter characterizes Loss of Heterozygosity (LOH).

Our model depends on the $L$ considered mutation centers, which need to be selected. To this end, considering all possible mutations with lrr in the interval $[\mathrm{lrr}_{\min}, \mathrm{lrr}_{\max}]$, we select the value $\mathrm{lrr}_{\max}$ using a penalized maximum likelihood approach with a BIC criterion (Schwarz 1978, Keribin 2000), while $\mathrm{lrr}_{\min}$ is determined by the parameters $p$, $\alpha$, and $\beta$ through $\mathrm{lrr}_{min} = \alpha \log_2(2p) + \beta$. The tumoral ploidy is then computed as the weighted average of the copy numbers which can be deduced from our estimation.

This performs well and results are presented on simulated and real datasets.

**Acknowledgements** C. Keribin is supported for this work by the *Paris-Saclay Center for Data Science* funded by the *IDEX Paris-Saclay*, ANR-11-IDEX-0003-02 and Yi Liu by *Programme Label Cancéropôle Ile de France*.

# Bibliographie

[1] Sun, W.; Wright, F.A.; Tang, Z.Z.; Nordgard, S.H.; Loo, P.V.; Yu, T.W.; Kristnesen, V.N. et Perou, C.M. (2009), Integrated study of copy number states and genotype calls using high-density SNP arrays, Nucleic Acids Research, 37, 16, 5365-5377.

[2] Yau, C.; Mouradov, D.; Jorrissen, R.N.; Colella, S.; Mirza, G.; Steers, G.; Harris, A.; Ragoussis, J.; Sieber, O et Holmes, C. (2010), A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data, Genome Biology, 11:R92.

[3] Li, A.; Liu, Z.Z.; Lezon-Geyda, K.; Sarkar, S.; Lannin, D.; Schulz, V.; Krop, I.; Winer, E.; Harris, L. et Tuck, D. (2011), GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays, Nucleic Acids Research, 39, 12, 4928-4941.

[4] Popova, T.; Manié, E.; Stoppa-Lyonnet, D.; Rigaill, G.; Barillot, E. et Stern, M.H. (2009), Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays, Genome Biology, 10:R128

[5] Loo, P. et al. (2010), Allele-specific copy number analysis of tumors, PNAS, 107, 39.

[6] Mosén-Ansorena, D.; Aransay, A.M. et Rodríguez-Ezpeleta, N. (2012), Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data, BMC Bioinformatics, 13:192.

[7] Liu, Y.N.; Li, A.; Feng, H.Q. et Wang, M.H. (2014), TAFFYS: an integrated tool for comprehensive analysis of genomic aberrations in tumor samples (under review), private communication.

[8] Dempster, A.P.; Laird, N.M. et Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1), 1–38.

[9] Schwarz, G.E. (1978), Estimating the dimension of a model, *Annals of Statistics* 6 (2): 461464

[10] Keribin, C. (2000), Consistent Estimation of the Order of Mixture Models, *Sankhya* Series A, volume 62, Part. 1, pp 49-66, 2000