

1 INTRODUCTION

SMAP is based on *segmental a posteriori maximization*, a scheme where, in our context, the most plausible assignments of measurement points to DNA copy numbers are found by maximizing the joint posterior probability of the parameters of an Hidden Markov Model (HMM) and the HMM state sequence (copy number assignments). By integrating parameter optimization in the method we enable adaptability of the HMM to data, restricted by prior probabilities, and generate the most plausible HMM model that describes the data. The exploitation of prior knowledge about the data provides flexibility and enhance the ability to adapt the analysis to data sources with various characteristics.

2 PARAMETER SETTINGS USED FOR THE GBM AND SYNTHETIC DATA

Supplementary table 1 summarizes the parameter settings used for *SMAP* in the glioblastoma multiforme (GBM) study and on the synthetic data. There are two types of parameters, static and dynamic.

Supplementary table 1. Static parameters and dynamic parameters used as start solution for *SMAP* on the glioblastoma multiforme samples.

	Parameter	Value
static	τ	0.05
	σ_μ	0.05
	σ_{min}	0.05
	$\mu_i^{expected}$	{0.4, 0.7, 1.0, 1.3, 1.6, 4.0}
	L	5 Mb
dynamic	Ω	{(0.4, 0.1), (0.7, 0.1), (1.0, 0.1), (1.3, 0.1), (1.6, 0.1), (4.0, 0.1),
	$\eta_j (\forall \lambda_j \in \lambda)$	0.01
	$\pi_i (1 \leq i \leq 6)$	$\frac{1}{6}$
	$a_{ij} (i \neq j)$	0.001
	$a_{ii} (1 \leq i \leq 6)$	0.995

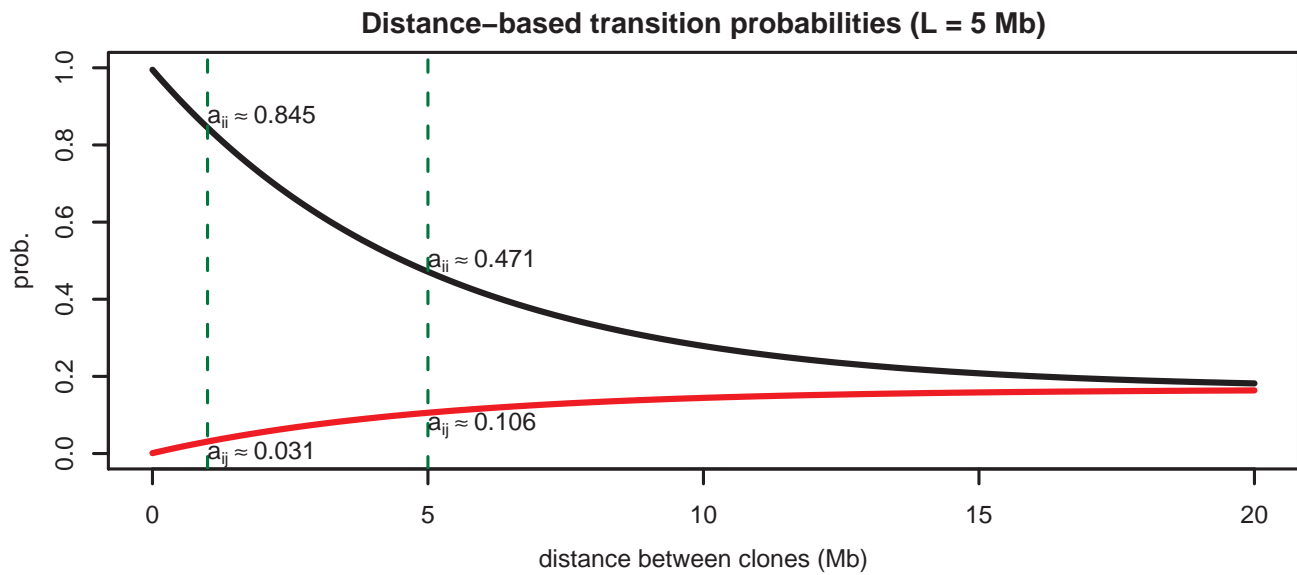
The static parameters are prior information about the expected means of Gaussian distributions associated with each copy number state ($\mu_i^{expected} (1 \leq i \leq N)$), the standard deviations of the expected means (σ_μ), the threshold (τ) which log joint posterior probability improvements must satisfy during *SMAP* and gradient descent iterations, and the length parameter (L) which controls the convergence of distance based transition probabilities towards $\frac{1}{N}$. Static parameters are fixed, hence not optimized during the parameter optimization of *SMAP*. The expected means are extracted from the dynamic parameters Ω , which specifies the initial parameters for the Gaussian distributions associated with each copy number state.

The dynamic parameters are optimized according to the data analyzed during the parameter optimization of *SMAP*. Apart from Ω , these include the transition probabilities (A), the initial probabilities (Π), and the individual learning rates of gradient descent considered parameters ($\eta_j (\forall \lambda_j \in \lambda)$).

3 DISCUSSION

Prior information may make a big difference in the conclusions that we draw from a set of observations. For instance, knowledge about normal cell admixture in the test DNA and the presence of noise in the data, which may cause intensity ratios to deviate from expected values, is crucial information which can be integrated in the process through appropriate values of Ω . *SMAP* optimizes Ω with respect to the analyzed data, a process that is controlled by σ_μ and σ_{min} . A high value of σ_μ means less control of the ability to adjust means of the state distributions. A too low value of σ_μ may, however, restrict the ability to adjust the means which may cause worse predictions. A value between 0.05 and 0.1 for σ_μ is recommended. σ_{min} controls the minimum allowed standard deviation.

The start solution for the HMM parameters is important since different settings may yield different results. For instance, too high initial values of transition probabilities between non-equal HMM copy number states may cause a higher false discovery rate (FDR). For the 6-state model used in the GBM study we used a value of 0.001 for $a_{ij} (i \neq j)$, i.e., the probability of staying in the same state is 0.995. *SMAP* uses distance-based transition probabilities which are derived from A , the distance between clones and a length parameter L . It is problematic to give any suggestions regarding the length parameter L that controls the convergence of transition probabilities towards equality as discussed by Colella *et al.* (2007). Experimental validation has indicated that 5 Mb is a reasonable choice on the 32K BAC array. As an example, consider Supplementary figure 1 in which a 6-state model is used with the parameter settings as in the GBM study (Supplementary table 1). At a distance of 1 Mb between clones, the transition probabilities $a_{ii} (1 \leq i \leq 6)$ have been reduced to ≈ 0.845 while the transition probabilities $a_{ij} (i \neq j)$ have been increased to ≈ 0.031 . It is hard to give any *good* recommendations on what value to use for L . Experimental validation has indicated that 5 Mb is a reasonable choice on the 32K BAC array.



Supplementary figure 1. Convergence of distance-based transition probabilities towards $\frac{1}{6}$. Black line corresponds to the transition probability of staying in the same state. Red line corresponds to the transition probabilities of changing state. Green dashed vertical lines depict the distances 1 Mb and 5 Mb between clones.

REFERENCES

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, **35**(6), 2013–2025.