Mixture model with multiple allocations for clustering spatially correlated observations in the analysis of ChIP-Seq data

Ranciati, Saverio; Viroli, Cinzia; Wit, Ernst C.

Published in:
BIOMETRICAL JOURNAL

DOI:
10.1002/bimj.201600131

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Biometrical Journal – (—-) –, zzz–zzz / DOI: 10.1002/bimj.200100000

Mixture model with multiple allocations for clustering spatially correlated observations in the analysis of ChIP-Seq data

Saverio Ranciati[*1], **Cinzia Viroli** [1], and **Ernst C. Wit** [2]

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy

[2] Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands

Received zzz, revised zzz, accepted zzz

Model-based clustering is a technique widely used to group a collection of units into mutually exclusive groups. There are, however, situations in which an observation could in principle belong to more than one cluster. In the context of Next-Generation Sequencing (NGS) experiments, for example, the signal observed in the data might be produced by two (or more) different biological processes operating together and a gene could participate in both (or all) of them. We propose a novel approach to cluster NGS discrete data, coming from a ChIP-Seq experiment, with a mixture model, allowing each unit to belong potentially to more than one group: these multiple allocation clusters can be flexibly defined via a function combining the features of the original groups without introducing new parameters. The formulation naturally gives rise to a 'zero-inflation group' in which values close to zero can be allocated, acting as a correction for the abundance of zeros that manifest in this type of data. We take into account the spatial dependency between observations, which is described through a latent Conditional Auto-Regressive process that can reflect different dependency patterns. We assess the performance of our model within a simulation environment and then we apply it to ChIP-seq real data.

Key words: ChIP-Seq; Mixture model; Model-based clustering; Multiple allocations; Spatial dependency;

Supporting Information for this article is available from the author.

1 Introduction

In the last 15 years, the development of parallel massively sequencing platforms for mapping the genome has completely revolutionized the way of studying genetic information. These recent technologies called Next Generation Sequencing (NGS) allow to simultaneously investigate thousands of features within a single reliable and cost-effective experiment, thus representing a valid alternative to microarray experiments in enhancing our understanding of how genetic differences affect health and disease (Nagalakshmi *and others*, 2008). Indeed, these innovative platforms have been quickly applied to many genomic contexts giving rise to a large amount of available data in complex form. Data coming from such experiments are highly structured and their analysis has raised an imperative need for specific methodologies: technical or biological replicates can be observed in several experimental conditions (Bao *and others*, 2014), and the single observational units, such as genes or exons, are very likely characterized by spatial dependencies and relationships (Mo, 2012). Moreover, the abundance of a particular transcript is measured as a count and so the single data point is a discrete measurement. In the context of Chromatin ImmunoPrecipitation and sequencing (ChIP-Seq) data, the authors of Thomas *and others* (2016) survey and review some established algorithms and some new ones, in order to compile a set of characteristics and properties that could define

*Corresponding author: e-mail: saverio.ranciati2@unibo.it

a common benchmark platform. They evaluate methods that aim at decoupling background and signal in the data: these algorithms may rely on test statistics (Binomial or Poisson, for example) or normalized scores that employ simultaneously input samples (that is, a 'control condition') and ChIP samples.

We consider in this paper the same context, ChIP-Seq experiments, and in particular data on proteins p300 and CBP analysed by Ramos *and others* (2010). In this experiment, two technical replicates are observed after 30 minutes from the initial interaction of the proteins with the human genome, with the aim of discovering the binding profile of these transcription factors. Figure 1 displays summarized counts for 1000 base pairs contiguous windows along chromosome 21. The plot shows segments with high read counts and segments where there is a uniformly low level of signal, thus suggesting a potential spatial effect.

In ChIP-Seq data like the one investigated (see Kuan *and others* (2011) for a review), researchers usually associate the counts to two specific components: a background level, which accounts for the noise in the process and the inactivity of the regions with respect to protein binding; a signal level, that is described by a higher counts of sequenced DNA fragments, indicating that the protein is actually interacting with those specific genomic regions. From a statistical point of view, the problem of the detection of such biological processes can be addressed by introducing a mixture model with the aim of identifying groups of genomic regions that exhibit similar preferential binding by a protein. Typically, conventional model-based clustering methods perform classification of units into mutually exclusive partitions. However, looking at Figure 1, it could be interesting to uncover components that may arise from the multiple overlapping action of the main aforementioned group processes. Multiple partitions can be obtained in (at least) two ways: (a) by fuzzy or 'soft' clustering that is the assignment of a unit to the group with some posterior probabilities (see, for instance, Bezdek, 1981; Heller *and others*, 2008) or (b) by accounting for multiple allocations directly within a generative model. Our proposal employs this second perspective that explicitly assumes that genomic units are fully co-actors of multiple processes in a model based framework. The idea stems from earlier contributions aimed at discovering multiple clusters. Battle *and others* (2005) introduced a probabilistic-based method to discover overlapping cellular processes and the associated gene regulation scheme. Given the complexity of a cellular system, they propose a decomposition of the observed continuous data matrix into layers representing biological processes and groups of co-regulated genes, allowing every unit to express itself in more than one activity layer and belong to multiple clusters. In Banerjee *and others* (2005) and Fu and Banerjee (2008), the problem of multiple allocation is solved within a model based clustering strategy, where the distribution of the groups is extended generalizing the Gaussian probability distribution used in Battle *and others* (2005) to the case of exponential family distributions. The main idea of such approach known as 'Model-Based Overlapping Clustering' is to re-parameterize the density of the overlapped clusters as the product of some primary densities that, being members of the exponential family, still result in the same parametric family. Heller and Ghahramani (2007) extended these models by employing a nonparametric Bayesian technique to infer the number of groups in their overlapping clusters model, while maintaining the characterization of the mixture densities as members of the exponential family. More recently, Zhang (2013) proposed the 'epistatic model based clustering' for the analysis of microarray data. In this approach, a more explicit description of the mixed component densities in terms of Gaussians is given; different interactions between the parameters of the primary groups are investigated but the order of the interactions between these original clusters and the overlapped counterparts is practically limited to the second order.

The aim of this work is to define a general Multiple Allocation Mixture (MAM) model for analyzing the ChIP-Seq data. The peculiar features of these experimental data demand for specific treatment. First, their discrete nature and a marked overdispersion require a flexible count distribution such as the Negative Binomial, that however does not generally belong to the exponential family, unless its dispersion parameter is known and fixed. To this aim we generalize the model-based overlapping clustering to arbitrary parametric probabilistic functions. In addition, as shown in Figure 1, ChIP-Seq data are characterized by the inflation of non-structural zeros. These aspects are naturally taken into account by the proposed model, where each component of the multiple mixture corresponds to a primary or to an overlapping cluster distributed as

Negative Binomials with parameters that are function of the primary parameters only. A further important aspect that emerges from Figure 1 is that the protein interactions with DNA are spatially correlated. We will show that the model can be easily extended in order to account for the spatial linkage among the genes, via a Conditional Auto-Regressive (CAR) process.

In what follows, we will present our proposal in three gradual steps in order to sequentially address these issues. First, in Section 2, the general MAM model will be presented and then we will adapt the model to the NGS data. We will illustrate how to extend this approach in order to model spatial dependent observations. In Section 3, a simulation study aimed at investigating the flexibility and effectiveness of the proposal is presented. Furthermore, we use the newly developed model to study the genome wide binding of the protein p300 in order to investigate its transcriptional regulatory function. In Section 4, conclusions are discussed.

2 Methods

2.1 Model-based clustering with mixture model

Finite mixture models have been receiving a wide interest in the statistical literature as a tool for performing model based clustering and density estimation (see Banfield and Raftery, 1993; McLachlan and Peel, 2000; Fraley and Raftery, 2002).

Let $\boldsymbol{y}_j$ ($j = 1, \ldots, p$) be an observed vector of values that we want to classify in some unknown groups. The conventional model based clustering model assumes

$$f(\boldsymbol{y}_j|\Theta) = \sum_{i=1}^{k} \pi_i f(\boldsymbol{y}_j|\boldsymbol{\theta}_i), \tag{1}$$

where $f(\boldsymbol{y}_j|\boldsymbol{\theta}_i)$ are the component densities with parameters θ_i and π_i are the prior probabilities to belong to each component, satisfying $\pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$. According to this model, a sample of p observations arises from some underlying populations of unknown proportions and the purpose is to decompose the sample into its mixture components, where each component corresponds to a cluster. In doing so, the data are partitioned into mutually exclusive k groups. This is achieved by introducing a latent variable, say z_{ji}, which allocate each observation j to the component i. More precisely, $\boldsymbol{z}_j$ is a vector of length k that takes the value 1 in correspondence of the cluster assignment, and 0 elsewhere, so that $\sum_{i=1}^{k} z_{ji} = 1$. According to the maximum a posteriori probability (MAP) rule, the partition is then obtained by assigning subjects to their most likely class according to the posterior probabilities of z given the data:

$$f(z_{ji}|\boldsymbol{y}_j; \Theta) = \frac{\pi_i f(\boldsymbol{y}_j|z_{ji}; \boldsymbol{\theta}_i)}{\sum_{i'=1}^{k} \pi_{i'} f(\boldsymbol{y}_j|z_{ji'}; \boldsymbol{\theta}_{i'})}.$$

In this sense the classification produced by a mixture model is 'hard' (because a unit is allocated to the mixture component with the maximum posterior probability of belonging to) but in principles it could be 'soft' by assigning each cluster a weight that equals its posterior probability as in the partial membership model (Heller *and others*, 2008). However, a soft assignment perspective does not mitigate the limitation of the model based classification, that results when data points may simultaneously belong to multiple clusters. In such situations, a change in the generative model is required, by explicitly assuming that the allocation vector $\boldsymbol{z}_j$ may contain several - and not just a single - ones.

2.2 Multiple Allocation Mixture model

In order to construct a new generative model for multiple components, we define k in Eq. 1 as the number of primary groups which are not mutually exclusive. For such a reason, we assume the prior probabilities of each primary group satisfying the constraint $\pi_i > 0$ but not necessarily summing up to one, $\sum_{i=1}^{k} \pi_i \neq 1$.

The total number of possible single and multiple allocation clusters is $k^* = 2^k$ and $\sum_{i=1}^{k} z_{ji} = n_j$ is the multiplicity of the cluster membership for the unit jth. When $n_j = 1$ we have a single group allocation, otherwise we have multiple group allocations. More precisely, if $n_j = 2$ the unit jth belongs to two groups simultaneously. These two groups altogether may be thought of as a new *secondary* group. If $n_j = 3$ the unit belongs to three groups that jointly define a *tertiary* group and so forth. When $n_j = 0$, the unit is assigned to what we call a 'outward cluster': this group collects observations that ideally do not belong to any clusters, and for this reason their distribution might be described by peculiar parameters depending on the empirical context. For instance, in many applications it could represent a group of outliers or noisy observations, characterized by high variance. The definition and existence of the 'outward cluster' is particularly relevant for the analysis of ChIP-Seq data, where the clusters are interpretable as biological processes. A gene that does not take part to any biological processes will have extremely low values (close to zero or zero). Thus, the outward cluster has the purpose to describe the group of 'inactive genes' and, in so doing, it acts as a zero-inflation adjustment for the model. For k fixed, let U be a connection matrix of dimension $2^k \times k$, with elements $u_{hi} \in \{0, 1\}$, containing all the possible assignment configurations. For instance, for $k = 3$:

$$U = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \tag{2}$$

In this case we define the prior probability to belong to a general (single or multiple allocation) group

$$\pi_h^* = \prod_{i=1}^{k} \pi_i^{u_{hi}} (1 - \pi_i)^{1 - u_{hi}} \tag{3}$$

with $h = 1, \ldots, k^*$. Notice that, by definition, these weights satisfy $\sum_{h=1}^{k^*} \pi_h^* = 1$. Each $\pi_h^\star$ comes naturally from the representation of clusters as sets in a sample space. For instance, if $k = 2$, $\sum_{h=1}^{k^*} \pi_h^* = (1 - \pi_1)(1 - \pi_2) + \pi_1(1 - \pi_2) + (1 - \pi_1)\pi_2 + \pi_1\pi_2 = 1$.

The Multiple Allocation Mixture model (MAM) can be defined as a reparameterization of Eq. 1 in the new formulation

$$f(\boldsymbol{y}_j|\Theta) = \sum_{h=1}^{k^*} \pi_h^* f(\boldsymbol{y}_j|\psi(\boldsymbol{u}_h, \boldsymbol{\theta}), \phi_h) \tag{4}$$

where $\psi(\boldsymbol{u}_h, \boldsymbol{\theta})$ is a location parameter and ϕ_h is a nuisance parameter. More specifically, $\psi(\boldsymbol{u}_h, \boldsymbol{\theta})$ is a function that depends on $\boldsymbol{u}_h$, which is the h-th row of U, and transforms the primary parameters $\boldsymbol{\theta}_i$ into the parameters of the multiple allocation components according to several possible schemes:

(i) *additive model*: the parameters of the mixed groups could be the sum of the original parameters, that is

$$\psi(\boldsymbol{u}_h, \boldsymbol{\theta}) = \begin{cases} \sum_{i=1}^{k} u_{hi}\boldsymbol{\theta}_i & \text{if } h > 1 \\ \boldsymbol{\theta}_b & \text{if } h = 1; \end{cases}$$

(ii) *co-dominance of order 1*:

$$\psi(\boldsymbol{u}_h, \boldsymbol{\theta}) = \begin{cases} \frac{\sum_{i=1}^{k} u_{hi}\boldsymbol{\theta}_i}{\sum_{i=1}^{k} u_{hi}} & \text{if } h > 1 \\ \boldsymbol{\theta}_b & \text{if } h = 1; \end{cases}$$

(iii) *co-dominance of order 0*:

$$\psi(\boldsymbol{u}_h, \boldsymbol{\theta}) = \begin{cases} \left(\prod_{i=1}^{k} \boldsymbol{\theta}_i^{u_{hi}}\right)^{1/\sum_{i=1}^{k} u_{hi}} & \text{if } h > 1 \\ \boldsymbol{\theta}_b & \text{if } h = 1. \end{cases}$$

How suitable is each scheme depends on the task at hand and the nature of the phenomenon being studied: an *additive* framework assumes that the effect of two main groups sums up to produce the overall activity of the multiple allocation group they originate; a *co-dominance* scheme reflect the idea that a multiple allocation group has a mean effect that is the average of the mean parameters of the main groups that originated it. The order of *co-dominance* simply describes the type of mean we are using, with order 1 being the arithmetic mean (AM) and order 0 the geometric mean (GM). Given the inequality of arithmetic and geometric means (AM-GM inequality), that is, $AM \geq GM$ when computed on the same values, the latter might be particularly useful to alleviate the effect of primary groups having 'extreme' mean parameters. Finally, the nuisance parameters ϕ_h are taken h-specific because we aim to add more flexibility to the mixture model.

2.2.1 Model selection

The choice of combining scheme to be used and number of clusters K are two aspects addressed via the adoption of a selection criterion, weighing both fitness and complexity of the resulting model. While choosing among different combining functions $\psi(\cdot)$ has the advantage to exploit (prior) information from the expert, how many clusters are needed to represent the heterogeneity in the data is a question that we answer by computing different quantities. The first option is the BICM (Raftery *and others*, 2007), that is a posterior simulation-based version of the canonical Bayesian Information Criterion

$$\text{BICM} = 2\ln(\mathcal{L}) - 2\tilde{s}^2 \ln(n),$$

where $\mathcal{L}$ is the likelihood of the model and $\tilde{s}^2$ a sample variance of the log-likelihood values, computed at each iteration of the MCMC. We select the number of clusters K producing the highest BICM. A second option is the BIC-MCMC criterion (Frühwirth-Schnatter, 2011)

$$\text{BIC-MCMC} = 2\ln(\tilde{l}) - \kappa \ln(n)$$

with $\tilde{l}$ being the largest computed log-likelihood value among the iterations of the MCMC and κ the number of parameters in the model. Again, the selected K is the one with the corresponding largest value of the BIC-MCMC. A third option is represented by the Deviance Information Criterion (DIC) (Spiegelhalter *and others*, 2002) and the use of an associated prior probability distribution for K, i.e. a Uniform prior in the range $(2, K_{\text{max}})$ for some reasonable and context-specific K_{max}.

2.3 Multiple Allocation Mixture model for ChIP-Seq Data

Suppose we observe the ChIP-Seq counts of p genes in D biological conditions or replicates. We denote Y_{jd} the random variable that expresses the read counts, say y_{jd}, mapped to gene j (j=1, ..., p), in sample d with $d = 1, \ldots, D$. Let $\boldsymbol{Y}_j$ be the random vector of length D denoting the counts for a gene across the different conditions. Let $\boldsymbol{y}_j$ be the observed value. We assume that Y_{jd} is distributed according to the Negative Binomial ($\mathcal{NB}$) distribution, given both the discrete nature of the observations and the flexibility of having a specific parameter used to model the overdispersion, which makes this distribution preferable over the Poisson. We further assume that, conditional on the group, the replicates are independent draws so that the mixture model in Eq. 4 becomes:

$$f(\boldsymbol{y}_j|\Theta) = \sum_{h=1}^{k^*} \pi_h^* \prod_{d=1}^{D} \mathcal{NB}\left(y_{jd}|\psi(\boldsymbol{u}_h, \boldsymbol{\mu}_d), \phi_{dh}\right),$$

where ϕ_{dh} are specific dispersion parameters of the negative binomials and $\mu^*_{dh} = \psi(\boldsymbol{u}_h, \boldsymbol{\mu}_d)$ are the means defined in the extended space of multiple components. We allow dispersion parameters ϕ_{hd} to vary for each of the k^* possible assignment and replicates because it is not directly clear a priori what is the most reasonable variance structure to define through a function for the multiple allocation clusters. This allows, nevertheless, for a certain degree of flexibility in describing the context-specific variability in the data. With reference to the means μ^*_{dh} they are modelled as function of primary k means through the combination scheme ψ. More specifically, with reference to the connection matrix U in Eq. 2, μ^*_{d1} is the mean parameter in condition d of the outward distribution; μ^*_{d2} is the mean parameter of the units that belong to the first primary group only and not to mixed groups, and so on, till μ^*_{d8} that is the mean of the units that belong simultaneously to the three components under condition d. Given the nature of the data we consider the 'outward' group as the component devoted to describe the inactive genes and for this reason we fix its mean to a very low value, such as $\mu^*_{d1} = 0.01$. The other mean parameters are obtained as combination scheme through ψ of k primary values that represent the means of the units that belong to the primary groups or related multiple groups. In order to construct an hierarchical form of the model, we define with $\boldsymbol{z}^*$ the allocation matrix in the augmented space given by k^* as a $p \times k^*$ 'multinomial' matrix. While each row of $\boldsymbol{z}$ can have multiple ones according to the assignment of the unit to multiple clusters, the rows of $\boldsymbol{z}^*$ only have a single one and the matrix U allows for a connection between original parametrization allocation and the re-parametrized version. Given the whole set of parameters, say Θ, the joint complete likelihood of our model is the following:

$$\mathrm{P}(\boldsymbol{y}, \boldsymbol{z}^*|\Theta) = \mathrm{P}(\boldsymbol{y}|\boldsymbol{z}^*, \Theta)\mathrm{P}(\boldsymbol{z}^*|\Theta). \tag{5}$$

Inference is carried out within a Bayesian framework. The posterior distribution of the parameters and the latent variables is:

$$\mathrm{P}(\boldsymbol{z}^*, \Theta, \boldsymbol{\xi}|\boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y}|\boldsymbol{z}^*, \Theta)\mathrm{P}(\boldsymbol{z}^*|\Theta)\mathrm{P}(\Theta|\boldsymbol{\xi}) \tag{6}$$

with $\mathrm{P}(\Theta|\boldsymbol{\xi})$ specifies priors for the mixture parameters and $\boldsymbol{\xi}$ as the vector containing the hyper-parameters. The posterior in Eq. 6 can then be rewritten as:

$$\begin{aligned}
&\mathrm{P}(\boldsymbol{z}^*, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\xi}|\boldsymbol{y}) \propto \mathrm{P}(\boldsymbol{y}|\boldsymbol{z}^*, \boldsymbol{\mu}, \boldsymbol{\phi})\mathrm{P}(\boldsymbol{z}^*|\boldsymbol{\pi})\mathrm{P}(\boldsymbol{\pi}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\mu}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\phi}|\boldsymbol{\xi}) = \\
&= \prod_{j=1}^{p}\prod_{d=1}^{D}\prod_{h=1}^{k^*}\left\{\frac{\Gamma(\phi_{hd}+y_{jd})}{\Gamma(\phi_{hd})\Gamma(y_{jd}+1)}\left[\frac{\phi_{hd}}{\phi_{hd}+\psi(\boldsymbol{u}_h,\boldsymbol{\mu}_d)}\right]^{\phi_{hd}}\left[\frac{\psi(\boldsymbol{u}_h,\boldsymbol{\mu}_d)}{\phi_{hd}+\psi(\boldsymbol{u}_h,\boldsymbol{\mu}_d)}\right]^{y_{jd}}\right\}^{z^*_{jh}} \times \\
&\times \prod_{j=1}^{p}\prod_{h=1}^{k^*}\left(\prod_{i=1}^{k}\pi_i^{u_{hi}}(1-\pi_i)^{1-u_{hi}}\right)^{z^*_{jh}} \mathrm{P}(\boldsymbol{\pi}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\mu}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\phi}|\boldsymbol{\xi}),
\end{aligned} \tag{7}$$

where $\mathrm{P}(\boldsymbol{\pi}|\boldsymbol{\xi}), \mathrm{P}(\boldsymbol{\mu}|\boldsymbol{\xi})$ and $\mathrm{P}(\boldsymbol{\phi}|\boldsymbol{\xi})$ are prior distributions for these quantities. As prior distribution for the weights we assume a Beta distribution so that we get the following hierarchical structure:

$$\begin{aligned}
\pi_i &\sim \text{Beta}\,(1,1) \\
\mathrm{P}(\boldsymbol{z}^*_{jh} = 1|\boldsymbol{\pi}^*) &= \pi^*_h \\
\boldsymbol{y}_{jd}|\boldsymbol{z}^*_{jh}; \boldsymbol{\Theta} &\sim \mathcal{NB}(\psi(\boldsymbol{u}_h, \boldsymbol{\mu_d}); \boldsymbol{\phi}).
\end{aligned} \tag{8}$$

For the other two parameters we select conjugate and flat priors. More precisely, for every i and h, $\mu_{id} \sim \text{Gamma}(a_\mu, b_\mu)$ and $\phi_{hd} \sim \text{Unif}(a_\phi, b_\phi)$, where $(a_\mu, b_\mu, a_\phi, b_\phi)$ are elements of the vector $\boldsymbol{\xi}$ of hyperparameters. Given these chosen priors, full conditionals can be derived and a Gibbs sampling MCMC algorithm applied to estimate the parameters. At the implementation step of the algorithm we exploit the Gamma-Poisson mixture representation of a Negative Binomial distribution. We introduce a further latent variabile s, specific to each unit j in each replicate/condition d, that has a Gamma density with shape and rate parameters equal to ϕ_{hd}. It follows that:

$$\begin{aligned}
f(s_{jd}) &= \text{Gamma}(\phi_{hd}, \phi_{hd}) \\
p(y_{jd}|s_{jd}) &= \text{Pois}(\psi(\boldsymbol{u}_h, \boldsymbol{\mu}_d)s_{jd}) \\
p(y_{jd}) &= \int p(y_{jd}|s_{jd})f(s_{jd})ds = \mathcal{NB}(\psi(\boldsymbol{u}_h; \boldsymbol{\mu}_d), \phi_{hd})
\end{aligned}$$

which allows us to use a Gibbs sampler for the mean parameters μ_i.

2.4 Modelling spatial correlation with a CAR structure

The units/observations (in our case, genes or genomic locations) are not independent but spatially correlated. We introduce the spatial correlation by allowing the primary weights π_i to be j-varying and we denote them as π_{ij}. The spatial relationship is taken into account by allowing the weights of the mixture in Eq. 4 to vary from one gene to another. The way we formulate it is inspired by the work proposed by Fernández and Green (2002). They first introduced k independent p-dimensional latent variables with a Markov random-field distribution. The weights are then a non-linear function of these latent variables and spatial relationships are expressed in terms of neighborhood relationships. This could be restrictive because the correlation between two genes should decrease as their distance increases. We extend the approach by considering a Gaussian conditional auto-regressive model (Pettitt *and others*, 2002), where the distances are directly used to model correlations instead of dummies denoting the neighborhood condition. In a Bayesian framework, this is accomplished by introducing additional layers to the hierarchy formulation shown in Eq. 8, with their own set of hyper-parameters. With reference to $\boldsymbol{\pi}$, we introduce the spatial latent vectors, denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_k$, with $i = 1, \ldots, k$: each $\boldsymbol{x}_i$ is a Gaussian conditional autoregressive model given by

$$f(\boldsymbol{x}_i) = \mathcal{N}(\mathbf{0}, Q^{-1}) \tag{9}$$

where Q is a precision matrix of order p and $\gamma_{jj'}$ is some non-linear function of $\delta_{jj'}$, which are the distances between all the units. More specifically,

$$Q = I_p + \Delta - \Gamma = \begin{cases} 1 + \sum_{j'=1}^{p} \gamma_{jj'} & \text{if } j\text{-th diagonal element} \\ -\gamma_{jj'} & \text{elsewhere} \end{cases}$$

After some algebraic steps it is possible to show that Eq. 9 is equivalent to

$$f(\boldsymbol{x}_i) = c \cdot \exp\left\{ -\frac{1}{2} \left[\sum_{j=1}^{p} \sum_{j'=1}^{p} \gamma_{jj'} (x_{ij} - x_{ij'})^2 + \sum_{j=1}^{p} x_{ij}^2 \right] \right\} \tag{10}$$

with c the normalization constant

$$c = (2\pi)^{-p/2} \prod_{j=1}^{p} (1 + v_j)^{1/2}.$$

In the previous expression, v_j's denote the eigenvalues of the spatial matrix $\Delta - \Gamma$. Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, the weights for location j can be obtained via logistic formulation

$$\pi_{ij} = \frac{\exp(x_{ij}/\eta)}{1 + \exp(x_{ij}/\eta)}$$

where η is a 'shrink-or-stretch' tuning parameter to be estimated that provides a way to exaggerate the differences in units allocation among the clusters.

2.5 Conditional Auto-Regressive Multiple Allocation Mixture (*CAR-MAM*)

In order to account for spatial correlation between the units/observations (in our case, genes or genetic locations), we introduce another layer in the hierarchical structure of the model. Starting from Eq. 5, the updated joint complete likelihood is:

$$\mathrm{P}(\boldsymbol{y}, \boldsymbol{z}^*, \boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\phi}, \eta, \boldsymbol{\xi}) = \mathrm{P}(\boldsymbol{y} | \boldsymbol{z}^*, \boldsymbol{\mu}, \boldsymbol{\phi}) \mathrm{P}(\boldsymbol{z}^* | \boldsymbol{x}, \eta) \mathrm{P}(\boldsymbol{x} | \boldsymbol{\xi}) \tag{11}$$

leading to the following posterior distribution

$$\mathrm{P}(\boldsymbol{z}^*, \boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\phi}, \eta | \boldsymbol{y}, \boldsymbol{\xi}) \propto \mathrm{P}(\boldsymbol{y}|\boldsymbol{z}^*, \boldsymbol{\mu}, \boldsymbol{\phi})\mathrm{P}(\boldsymbol{z}^*|\eta, \boldsymbol{x})\mathrm{P}(\boldsymbol{x}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\mu}|\boldsymbol{\xi})\mathrm{P}(\boldsymbol{\phi}|\boldsymbol{\xi}) \tag{12}$$

where the vector $\boldsymbol{\xi}$ now also contains the hyper-parameters for the new latent layer and $\mathrm{P}(\boldsymbol{y}|\boldsymbol{z}^*, \boldsymbol{\mu}, \boldsymbol{\phi})$ is defined as in Eq. 7. More precisely, the complete latent structure in Eq. 12 is equal to:

$$\begin{aligned} \mathrm{P}(\boldsymbol{z}^*, \boldsymbol{x}|\eta) \quad &= \prod_{j=1}^{p} \prod_{h=1}^{k^*} \left[\prod_{i=1}^{k} \left(\tfrac{\exp(x_{ij}/\eta)}{1+\exp(x_{ij}/\eta)} \right)^{u_{hi}} \left(1 - \tfrac{\exp(x_{ij}/\eta)}{1+\exp(x_{ij}/\eta)} \right)^{1-u_{hi}} \right]^{z^*_{jh}} \times \\ \times \quad & \prod_{i=1}^{k} c \exp\left\{ -\tfrac{1}{2} \left[\sum_{j=1}^{p} \sum_{j'=1}^{p} \gamma_{jj'} (x_{ij} - x_{ij'})^2 + \sum_{j=1}^{p} x_{ij}^2 \right] \right\} \end{aligned}$$

where c is the constant defined in Eq. 10. As proposed in Fernández and Green (2002), we integrate out the latent allocation variable $\boldsymbol{z}$ when implementing the Metropolis sampler for $\boldsymbol{x}$ in order to employ the information carried by the data $\boldsymbol{y}$ and bring the two layers of the hierarchical structure of the model closer together.

3 Results

3.1 Simulation study — Multiple Allocation Mixture (*MAM*) model

We assess the performance of our model (*MAM*) under different scenarios and we compare it with a classical non-overlapping components mixture (*NegBinMix*) and a variant (*OverlapPois*) of the algorithm proposed in Heller and Ghahramani (2007). The algorithm *OverlapPois* follows the same idea presented by the authors in their work: the density of an overlapping region is modeled by the product of the cluster-specific densities of the involved original groups. In particular, the authors exploit the properties of exponential families and the fact that the product of two (or more) distributions from the exponential family belongs to the same exponential family, with new parameters defined from the natural parameters of the main groups (see Heller and Ghahramani (2007) for further details, Section 2). We use the Poisson distribution as a reference for *OverlapPois* because the Negative Binomial distribution does not belong to the exponential family (unless the dispersion parameter ϕ is assumed known, which is not our case). Our implementation differs in one main aspect: the choice of k. In their work, the authors use a Dirichlet Process to infer the number of clusters: in our context, the number of clusters k is fixed and provided. Data are generated from two independent Negative Binomial distributions ($D = 2$) and $p = 2000$ units, allowing for overlapping clusters with a number of groups $k = \{2, 3\}$: in the augmented k^* space this equals to represent a situation, as in a classical model-based clustering framework, where the actual number of groups ranges from $k^* = 4$ to $k^* = 8$. We explore three degrees of clustering between primary and outward/non-primary groups by selecting three scenarios which we call low, medium and high activation: this is achieved by setting all the π equal to - respectively - 0.25, 0.50 and 0.75. We run the three algorithms (*MAM*, *NegBinMix* and *OverlapPois*) for 10000 MCMC iterations and a 5000 burn-in window is selected. For every μ_{id}, we choose hyperparameters of the Gamma prior distributions equal to $a_\mu = 1$ and $b_\mu = 0.001$; for every dispersion parameter ϕ_{hd}, we select the ranges of the prior uniform distributions to be $[a_\phi = 100, b_\phi = 2000]$. Convergence is checked for every chain and we assign to the clusters according to the maximum a posteriori rule: we compute the posterior probabilities of the allocation vectors $\boldsymbol{z}^*_j$ given the data and, for every unit, we allocate to the component with the highest probability value (see also Section 2.1). We choose as an overall performance indicator the misclassification error rate (see Table 1), that is, the average number of units not correctly allocated when compared to the known true membership. The posterior means of the parameters in the selected models for *MAM* and *NegBinMix* are consistent with the true values and do not show any substantial bias. The posterior distributions of the parameters for *OverlapPois* are moderately shifted with respect to the true values and this impacts the overall performance of the method in terms of classification. We report the misclassification error rates for the estimated models in Table 1: the

percentage of misclassified units is always lower in the low activation scenario because most of the observations are allocated in the outward component, which has small variance and near zero (fixed) mean, thus simplifying the clustering task. As we can see in Table 1, our model has comparable (sometimes better) classification rates, with respect to the conventional mixture of Negative Binomial, in simpler simulated dataset ($k = 2$). When $k = 3$, *MAM* model always outperforms the compared mixture with noticeable improvements on the misclassification error rate. Also, *OverlapPois* is performing constantly worse than the other two.

3.2 Simulation study — *MAM* with Conditional Autoregressive model (*CAR-MAM*)

We simulate data from a mixture of Negative Binomials with multiple allocation and spatial information; we choose $k = 2, 3, 4$ ($k^* = 4, 8, 16$) and for every number of groups we adopt two different spatial structures. In the former, the latent variable $\boldsymbol{x}$ is drawn from a CAR model using a reciprocal function $\gamma_{jj'}$ (see Pettitt *and others*, 2002), which is an inverse function of the distances between uniformly drawn positions (pos_j, $j = 1, \ldots, p$). In the latter, a sine function is employed in the data generating process, in order to have stronger spatial relationships. More precisely, we assume that the spatial latent vectors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are:

$$x_{ij} = \sin\left(i\pi \frac{pos_j}{max\left\{pos_j\right\}_{j=1,\ldots,p}} \right).$$

For each scenario we run three algorithms assuming k is known: *NegBinMix*, which is a mixture of Negative Binomials without any further specification, and our two proposed models *MAM* and *CAR-MAM* assuming that the conditional autoregressive structure is computed through a reciprocal function $\gamma_{jj'}$. We average the misclassification error rates across 100 independent datasets simulated for each scenario and we compare the results to assess the performance. Again, we cluster the units according to the maximum a posteriori rule and we measure the performance through the misclassification error, computed for each model. In Table 2 and Table 3 we summarize the results of 100 runs of both scenarios, where data were generated - respectively - with a reciprocal function for the *CAR* part of the model or a sine function. As clear from the tables, we achieve improved accuracy in the clustering task with *CAR-MAM* model over the simpler *MAM* model and both of them perform better with respect to *NegBinMix* (Table 2) even when the estimated conditional autoregressive structure is not the same as the one used in the data generating process (Table 3).

3.3 p300 protein binding ChIP-Seq experiment

We apply our model to the data already discussed in the introduction and previously analysed by Ramos *and others* (2010). p300 is a transcription coactivator associated with many genes that are involved in multiple processes (i.e., differentiation, apoptosis, proliferation); also, the protein serves as a bridge for other transcription factors and it is involved in the transcription machinery of cells related to the development of cancer and other diseases. The data from p300 ChIP-Seq experiments consist of results from multiple interactions of the transcription coactivator with the chromatin in quiescent and stimulated cells, at different time points. We select two technical replicates (T01, T02) collected 30 minutes after the initial interaction between the chromatin and the protein p300 on a sequence of 1000 base-pairs windows describing the pre-processed raw counts across 4000 regions of the chromosome 21 (see Figure 1). We analyze both replicates jointly and we summarize the data of this subsample by computing the average of the counts for the two different replicates T01 and T02. As we can see in the plot, the majority of the observations lie in a 'band' of counts lower than 5, aggregated into segments that are spanned by smaller batch of regions exhibiting a higher count level, thus suggesting a spatial effect with respect to the protein binding process. Since in ChIP-Seq data researchers usually associate the counts to a background group and a signal group, we run the algorithm with $k = 2$ ($k^* = 4$) in order to capture the two aforementioned expected clusters and potentially a better characterization of them through our multiple allocation cluster, while simultaneously

taking into account the spatial dependency between the genomic regions. We choose the geometric mean (*co-dominance* of order 0) as our combination scheme $\psi(\cdot)$ to lessen the effect (on the multiple allocation cluster mean values) produced by the highest counts (as previously said at the end of Section 2.2). Out of the 4000 genomic regions, 159 are allocated in the outward cluster, which accounts for the zero-inflation in the data, represented by the observations around positions 17200 and 18000 (see Figure 5). This group has fixed mean values equal to 0.01 for both replicates T01 and T02, while the dispersion parameters are estimated by the algorithm. The first primary cluster $i = 1$, whose units are indicated by green dots in Figure 5, has posterior means equal to $\boldsymbol{\mu}_1 = (1.61, 1.07)$ and represents the background process of the protein binding: 3478 possibly inactive genomic regions, with very low counts, are allocated in this group. The second primary cluster, $i = 2$, is representative of a signaling group of 48 genomic regions having a higher mean count level of $\boldsymbol{\mu}_2 = (19.27, 34.23)$ depicted with red dots in Figure 5. The multiple allocation cluster in our analysis is interpretable as a group of 315 units involved in both the background and signal clusters: in this case, given that we are observing these counts after 30 minutes from the initial interaction of the protein with the strand of chromosome 21, the genomic regions allocated in this cluster could be thought as either being locations that were active immediately at the beginning and now not signaling anymore or locations only starting to interact with the p300 protein after 30 minutes. The mean values for this multiple allocation cluster are equal to 5.57 and 6.06 in the two replicates T01 and T02: the units belonging to it are shown as blue dots in Figure 5. Finally, the posterior probability for each unit to be allocated in the signaling group is shown in Figure 5 as a magenta solid line. We can see from the plot that this probability is higher in those segments where genomic regions with higher counts are observed; moreover, the allocation weight of the signal component follows a spatial pattern, increasing and decreasing across the analyzed strand and thus accounting for the spatial effect occurring among the observations. For comparison, we also run the algorithm for a conventional four components Negative Binomial mixture model, fixing the location parameter of the first component to 0.01 as we do in our model. The result shows that only a total of three clusters are estimated, with one of them capturing the same background mean level shared by the fixed component, thus producing a blurred classification with respect to one of the two main processes of interest. For comparative purposes, we also applied to the data the Bayesian Change Point (*BCP*) method (Xing *and others*, 2012), recently reviewed among other approaches in Thomas *and others* (2016). For *BCP*, the two technical replicates T01 and T02 are analyzed separately and results for both are provided in Figure 6 and Figure 7. In the case of *CAR-MAM*, an enrichment statement about each bin is obtained as a by-product of the allocation procedure, wether a location is put in the 'pure signal' group or 'potentially signaling' one. As for *BCP* we only obtain a bin-specific posterior probability of changes in the observed counts sequence: in order to call for the analyzed locations to be enriched or not we select a cut-off that produces an estimated False Discovery Rate of 5%. For the dataset we analyze, this cut-off on *BCP* posterior probabilities is equal to 0.46. To assess the advantage of our approach with respect to another Bayesian method as *BCP*, we summarize both classifications in Table 4. *BCP* detects 210 changes in the sequence of bins out of 4000 total locations: as we can see from the second column, we are able to provide with *CAR-MAM* a more detailed representation of the noise observed in the data; of the 3790 bins not detected as signal by *BCP* we classify 159 of them to be zero-inflation, 3 as 'pure signal' and the rest as either noise (3416 bins) or counts produced by both mechanism (signal and noise, 212). In the third column, of the 210 locations with an associated changepoint detected by *BCP*, our approach *CAR-MAM* finds that 148 of them are signal or an overlap of signal and noise while 62 bins are allocated in to the noise group (in contrast with *BCP*).

4 Conclusions

Motivated by the analysis of data coming from ChIP-Seq experiments, we proposed an extension of the conventional mixture model that allows for units to simultaneously belong to more than one group. As a by product of the model specification, an outward cluster has been introduced that can be used to describe

specific features of the data such as zero-inflation, outliers and so forth. A spatial dependency layer among the units is encoded in the formulation by the means of a conditional autoregressive model, allowing for spatial information to aid the clustering task. We have compared our proposed model with a mixture of Negative Binomials to investigate its advantages with respect to the conventional approach: results on the simulated data show an increase in performance in terms of misclassification error rates. We applied our model to data previously analyzed by Ramos *and others* (2010): a promising richer description of the signal in the observations is found, calling for a potentially deeper biological investigation of those genomic regions associated with it. A delicate issue that deserves future investigation is the choice of k, which is a fundamental aspect because it identifies the number of primary clusters. Some well-known methods such as reversible jump (Green, 1995) and birth-death process (Stephens, 2000) do not allow an extensive exploration of the range of values for k without incurring in a dramatic increase in computational cost. However, in some applications, this choice could be suggested by the empirical context such as in our case, where a number of primary clusters equal to two was reasonable because they represented background and signal groups.

Conflict of Interest

The authors have declared no conflict of interest.

References

BANERJEE, A., KRUMPLEMAN, S., BASU, S., MOONEY, R. AND GHOSH, J. (2005). Model-based overlapping clustering. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, 532–537.

BATTLE, A. AND SEGAL, E. AND KOLLER, D. (2005). Probabilistic Discovery of Overlapping Cellular Processes and their regulation. *Journal of Computational Biology*, **12**, 909–927.

BANFIELD, J. P. AND RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

BAO, Y., VINCIOTTI, V., WIT, E. AND A. C. 'T HOEN, P. (2014). Joint modeling of ChIP-seq data via a Markov random field model. *Biostatistics*, **15**(2), 296–310.

BEZDEK, JAMES C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, Norwell, MA, USA.

FERNÁNDEZ, C. AND GREEN, P.J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, **64**, 805–826.

FRALEY, C. AND RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

FRÜHWIRTH-SCHNATTER, S. (2011). Label switching under model uncertainty. *Mixtures: Estimation and Application*, 213-239

FU, Q. AND BANERJEE, A. (2008). Multiplicative Mixture Models for Overlapping Clustering. *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE*, 791–796.

GREEN, P. J. (1995). Reversible jump Markov Chain, Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

HELLER, A.K. AND GHAHRAMANI, Z. (2007). A Nonparametric Bayesian Approach to Modeling Overlapping Clusters. *International Conference on Artificial Intelligence and Statistics*, 187–194.

HELLER, A.K. AND WILIAMSON, S. AND GHAHRAMANI, Z. (2008). Statistical Models for Partial Membership. *Proceedings of the 25th international conference on Machine learning*, ACM, 392–399.

KUAN, PEI FEN, ET AL. (2011). A statistical framework for the analysis of ChIP-Seq data. *Journal of the American Statistical Association*, 106.495, 891-903.

MCLACHLAN, G. J. AND PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

MO, Q. (2012). A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics*, **13**(1), 113–128.

NAGALAKSHMI, U., AND WANG, Z. AND WAERN, K. AND SHOU, C. AND RAHA, D. AND GERSTEIN, M. AND SNYDER, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*,

320, 1344:1349, 2008.

PETTITT, A.N., WEIR, I.S. AND HART, A.G., (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data, *Statistics and Computing*, **12**, 353–367.

RAFTERY, A., NEWTON, M., SATAGOPAN, J., AND KRIVITSKY, P. (2007). Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity, In *Bayesian Statistics 8*, 1-45.

RAMOS Y, HESTAND M, VERLAAN M, KRABBENDAM E, ARIYUREK Y, VAN DAM H, VAN OMMEN G, DEN DUNNEN J, ZANTEMA A, 'T HOEN P (2010). Genome-wide assessment of differential roles for p300 and CBP in transcription regulation, *Nucleic Acids Res*, **38(16)**, 5396-5408.

THOMAS, R., THOMAS, S., HOLLOWAY, A.K. AND POLLARD, K.S. (2016). Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, doi: 10.1093/bib/bbw035.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., AND VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64(4)**, 583-639.

STEPHENS, M. (2000). Bayesian analysis of mixtures models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, **20**, 40-74.

XING, H., MO, Y., LIAO, W., AND ZHANG, M. Q. (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS Comput Biol*, **8(7)**, e1002613.

ZHANG, J. (2013). Epistatic Clustering: A Model-Based Approach for Identifying Links Between Clusters. *Journal of the American Statistical Association*, **108**, 1366–1384.

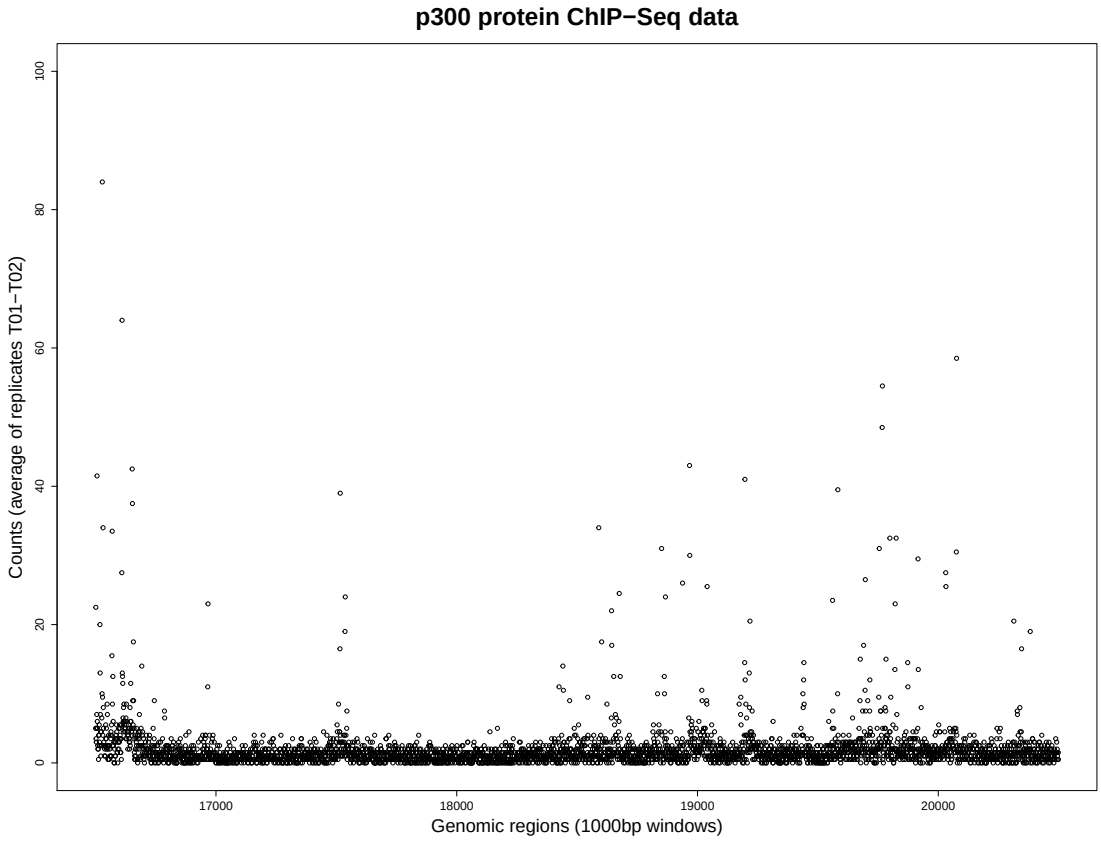

Figure 1 Summarized counts for 4000 bins from the chromosome 21. Regions on x-axis are 1000bp windows; averaged counts of the two replicates T01 and T02 reported on y-axis.

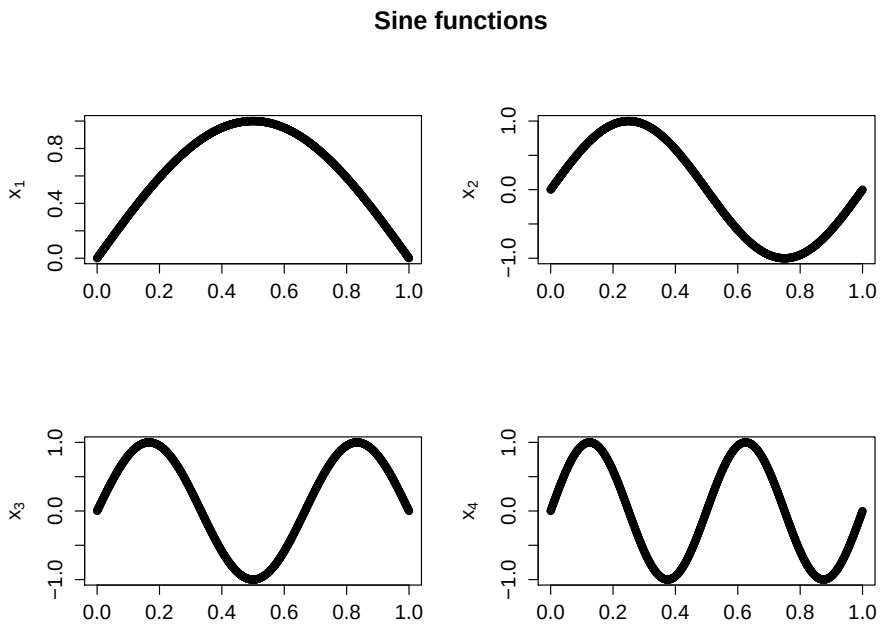

Figure 2 Graphical visualization of the sine functions for $i = 1$ up to $k = 4$.

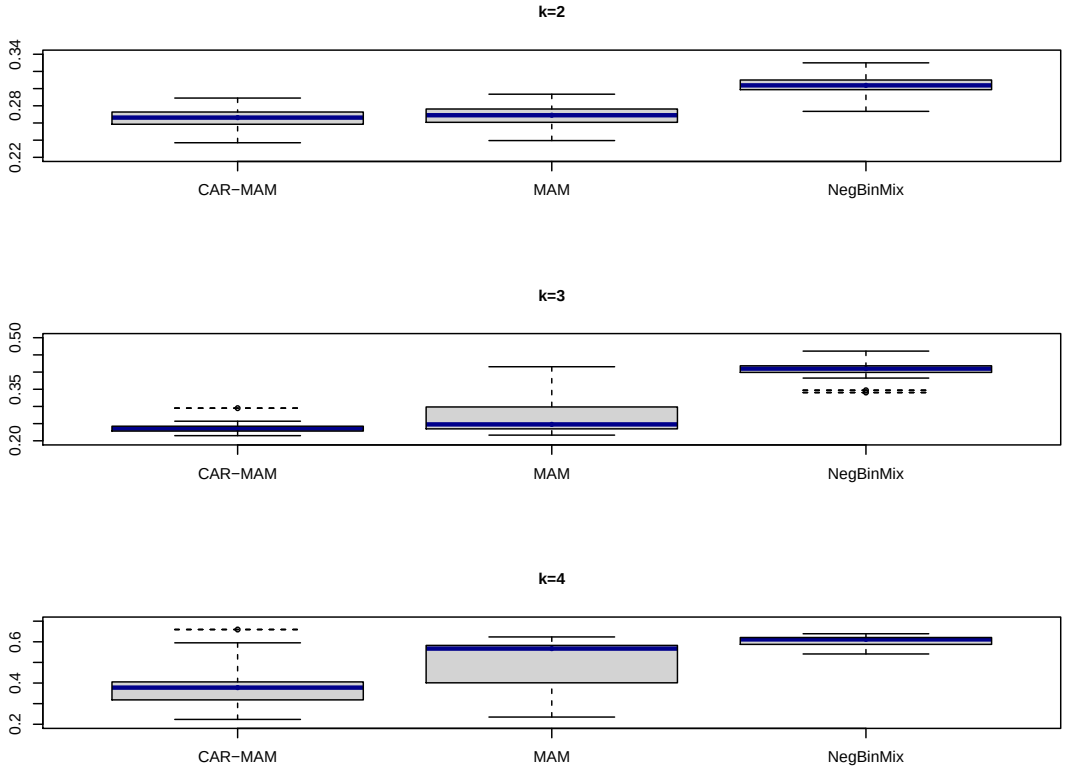

Figure 3 Boxplots of misclassification errors over 100 datasets generated from CAR model using reciprocal function.

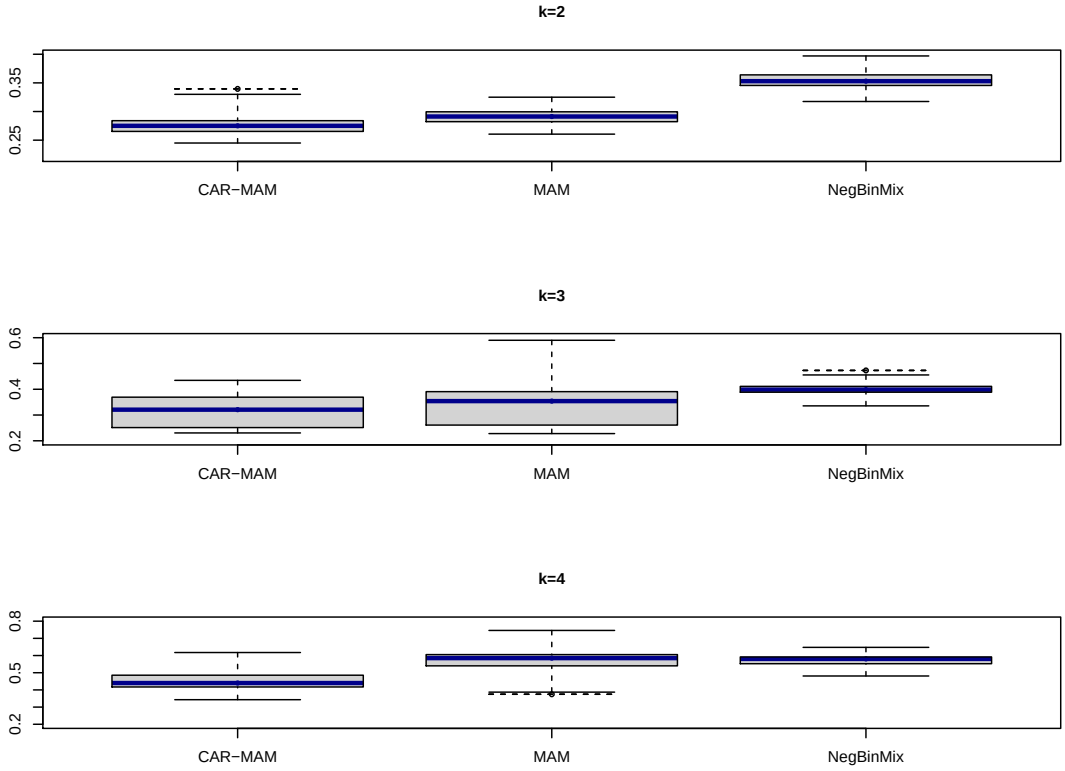

Figure 4 Boxplots of misclassification errors over 100 datasets generated from CAR model using sine function.

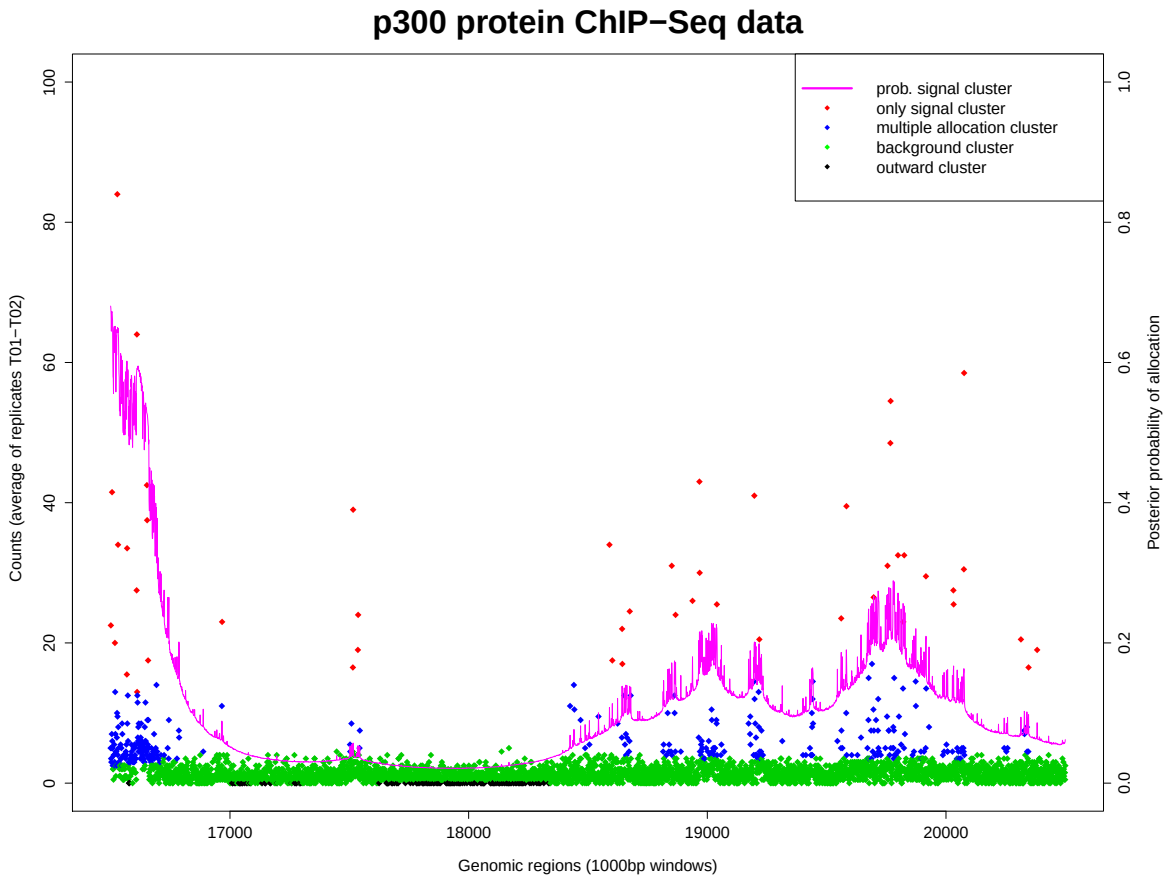

Figure 5 Clustering result for the analyzed segment. Counts are reported as average of the two technical replicates T01 and T02; solid line is the posterior probability to be allocated to the signal group.

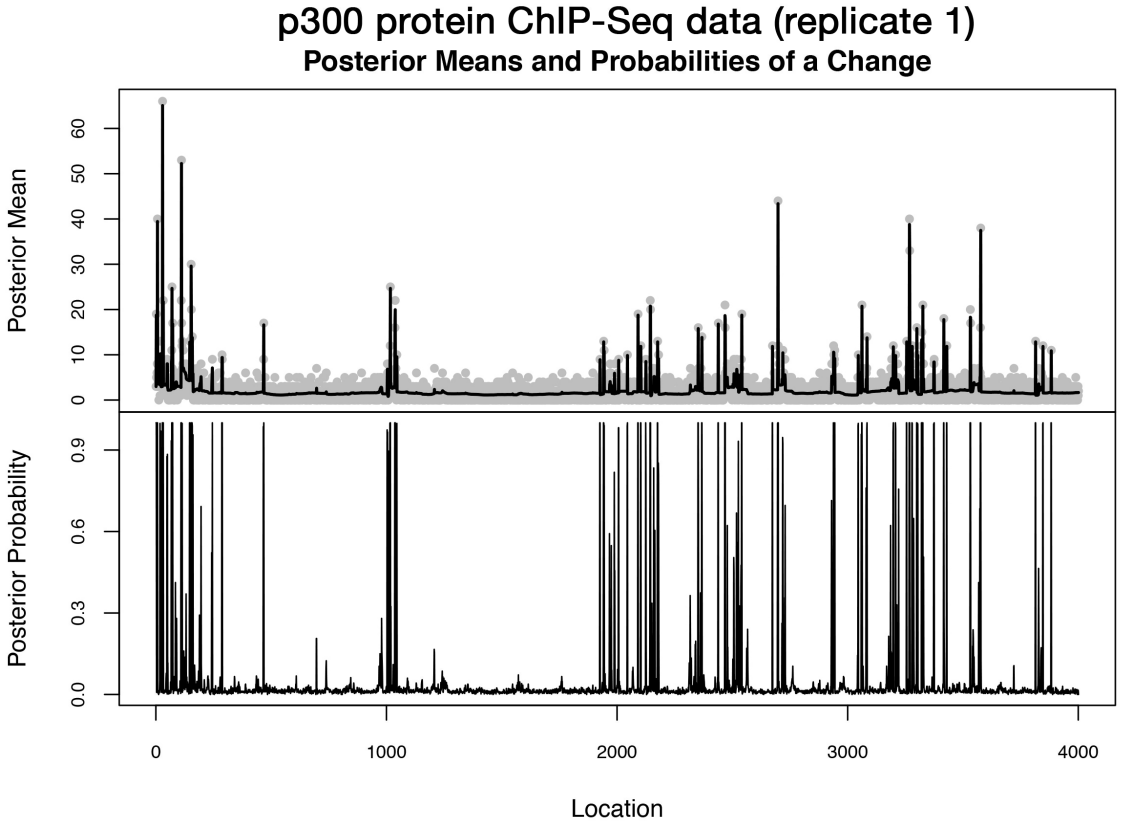

Figure 6 Bayesian Change Point analysis results for the first technical replicate T01. Posterior means are reported on the first plot (top); posterior probabilities of changepoint are provided in the second plot (bottom).

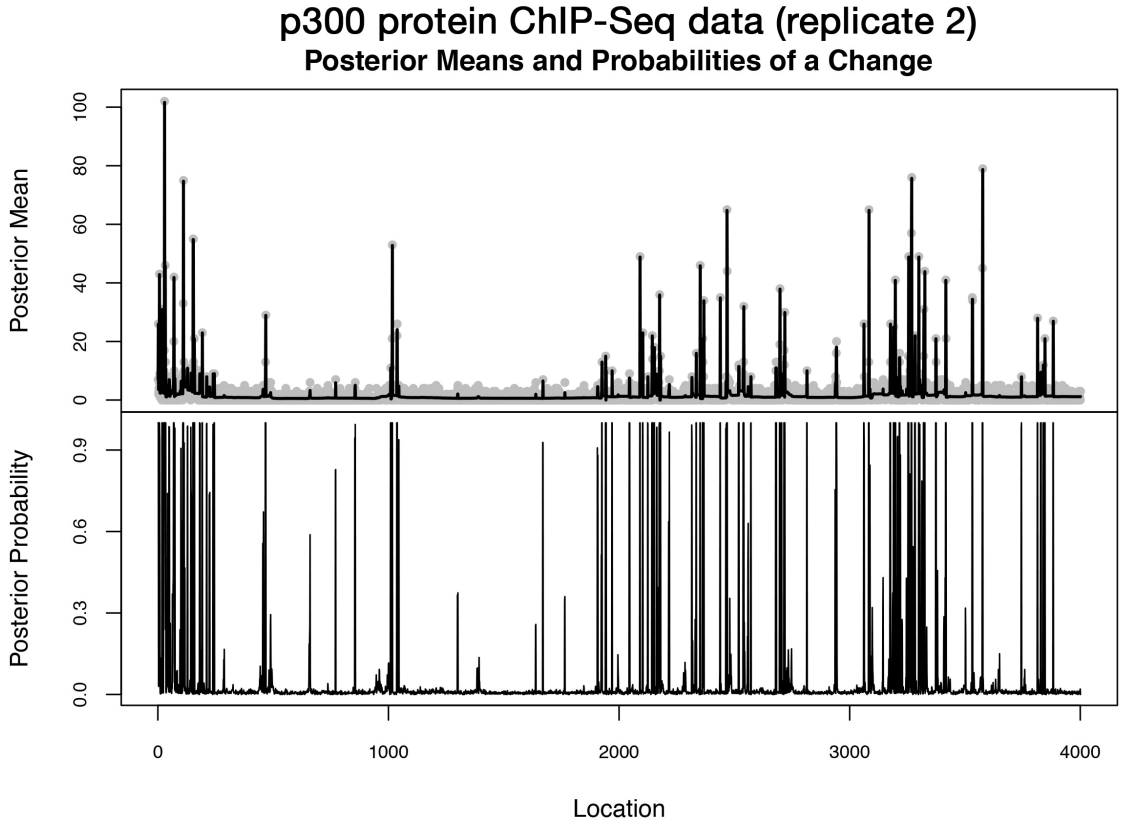

Figure 7 Bayesian Change Point analysis results for the first technical replicate T02. Posterior means are reported on the first plot (top); posterior probabilities of changepoint are provided in the second plot (bottom).

Table 1 Misclassification error rate (in percentage) for *NegBinMix*, *MAM* and *OverlapPois*, for the three scenarios of activation.

Number of clusters	Model	Degree of activation		
		$\pi_i = 0.25$	$\pi_i = 0.50$	$\pi_i = 0.75$
$k = 2$	*MAM*	**1.05**	2.70	2.65
(i.e. $k^* = 4$)	*NegBinMix*	1.20	**2.65**	**2.55**
	OverlapPois	7.25	24.40	19.30
$k = 3$	*MAM*	**0.95**	**3.05**	**5.10**
(i.e. $k^* = 8$)	*NegBinMix*	11.00	25.05	9.15
	OverlapPois	15.60	49.50	39.40

Table 2 Average misclassification error rate (in percentage) over 100 simulated datasets with standard errors in brackets; CAR structure with reciprocal function.

Number of clusters	*CAR-MAM*	*MAM*	*NegBinMix*
$k = 2$ $(k^* = 4)$	26.58 (*0.001*)	26.87 (*0.001*)	30.40 (*0.001*)
$k = 3$ $(k^* = 8)$	23.60 (*0.001*)	28.65 (*0.009*)	41.00 (*0.002*)
$k = 4$ $(k^* = 16)$	36.32 (*0.009*)	49.84 (*0.013*)	60.39 (*0.002*)

Table 3 Average misclassification error rate (in percentage) over 100 simulated datasets with standard errors in brackets; CAR structure with sine function.

Number of clusters	*CAR-MAM*	*MAM*	*NegBinMix*
$k = 2$ $(k^* = 4)$	27.71 (*0.002*)	29.13 (*0.001*)	35.42 (*0.001*)
$k = 3$ $(k^* = 8)$	31.31 (*0.006*)	35.06 (*0.011*)	40.11 (*0.002*)
$k = 4$ $(k^* = 16)$	44.93 (*0.005*)	56.12 (*0.007*)	57.18 (*0.003*)

Table 4 Comparison between the two classifications (*CAR-MAM* and *BCP*) on the number of enriched bins. For *BCP* method, the classification is conditional on an estimated False Discovery Rate equal to 5%.

CAR-MAM	*BCP*		Totals
	not enriched (noise)	enriched (signal)	
zero-inflation	159	0	159
noise	3416	62	3478
enriched (signal only)	3	45	48
enriched (mixed signal)	212	103	315
Totals	3790	210	4000

