Bayesian Nonparametric Clustering and Association Studies for Candidate SNP Observations

Charlotte Wang[a], Fabrizio Ruggeri[b], Chuhsing K. Hsiao[c], Raffaele Argiento[b,d,*]

[a] *Department of Mathematics, Tamkang University, Tamsui District, New Taipei City 25137, Taiwan*
[b] *CNR-IMATI, Milano 20133, Italy*
[c] *Bioinformatics and Biostatistics Core, Division of Genomic Medicine, Research Center for Medical Excellence, National Taiwan University, Taipei 100, Taiwan*
[d] *School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, U.K.*

Abstract

Clustering is often considered as the first step in the analysis when dealing with an enormous amount of Single Nucleotide Polymorphism (SNP) genotype data. The lack of biological information could affect the outcome of such procedure. Even if a clustering procedure has been selected and performed, the impact of its uncertainty on the subsequent association analysis is rarely assessed. In this research we propose first a model to cluster SNPs data, then we assess the association between the cluster and a disease. In particular, we adopt a Dirichlet process mixture model with the advantages, with respect to the usual clustering methods, that the number of clusters needs not to be known and fixed in advance and the variation in the assignment of SNPs to clusters can be accounted. In addition, once a clustering of SNPs is obtained, we design an individualized genetic score quantifying the SNP composition in each cluster for every subject, so that we can set up a generalized linear model for association analysis able to incorporate the information from a large-scale SNP dataset, and yet with a much smaller number of explanatory variables. The inference on cluster allocation, the strength of association of each cluster (the collective effect on SNPs in the same cluster), and the susceptibility of each SNP are based on posterior samples from Markov chain Monte Carlo methods and the Binder loss information. We exemplify this Bayesian nonparametric strategy in a genome-wide association study of Crohn's

*Corresponding author.
Email address: raffaele@mi.imati.cnr.it (Raffaele Argiento)

disease in a case-control setting.

Keywords: Bayesian, clustering, Dirichlet process mixture model, exchangeable partition probability function, genetic score, GWAS, logistic regression, nonparametric, random partitions

---

1. Introduction

Researchers nowadays prefer to test the association between multiple markers and a disease of interest in genetic association studies because the tests with multiple markers are more powerful, efficient, and biologically meaningful than single marker tests. Many statistical methods have been proposed based on those considerations, such as regularized regression models like lasso or ridge regression [1, 2, 3, 4, 5], gene-set enrichment analysis [4, 6, 7, 8], pathway [9], and network analysis [10]. Those methods are helpful to analyze large-scale markers and their corresponding interactions in the same pathway or network, when the analytic genomic region is pre-defined. Such tools, however, may be limited when utilized on regions containing a great amount of genetic markers or at the genome-wide scale. When analyzing data with such size, there may be no complete information about the role of each gene and the interaction among them, so that figuring out the association between these markers and disease phenotypes can be challenging. Therefore, an important issue for scientists is how to cluster or categorize the genomic markers in advance, so that the dimension of the data can be reduced and the genetic markers are represented with several relatively small and manageable sets.

Most current clustering algorithms evaluate first the distance between objects and then group them according to certain criteria. The definition of distance can vary from Euclidean measure for continuous observations to counting measure for discrete data. The choice depends on the problem and also on the data characteristics. For discrete observations like SNP genotypes, similarity or dissimilarity measures can be employed. A common measure with a natural biological interpretation is the linkage disequilibrium, where allele frequencies per locus and haplotype phase need to be derived *a priori*, based on genotype data. This derivation involves the uncertainty in haplotype configuration, introducing even more parameters in either case-control or pedigree studies [11, 12]. Other algorithms use mathematical formulations of similarity between SNP genotypes, including principle component analysis [13], k-means, and Hamming distance metric [14]. These tools are flexible in the sense that no biological information is required in

advance. In most clustering algorithms, however, the decision on the number of clusters is a difficult task. Its choice as a stochastic parameter usually complicates modeling and increases the computational burden [15, 16, 17].

Clustering or partitioning can be easily dealt within a Bayesian nonparametrics framework through the Dirichlet process mixture models, which allocate data to clusters and determine their number [18, 19]. Previous Bayesian applications in association studies either assumed two fixed clusters, i.e., associated vs. non-associated genes, and used Bayes factors for hypothesis testing, or applied a mixture model for every single marker [20, 21, 22]. No clustering procedure or multiple-marker effects were considered, and markers were examined individually, assuming exchangeability of their parameters.

From the modeling point of view, Dirichlet process mixture (DPM) models do not require the specification of the number of mixture components and the clustering procedure can be viewed as a Chinese restaurant process (see [23, 24] for more details). Inference on the number of clusters and mixture model parameters estimation are unified and performed by a suitable Markov chain Monte Carlo (MCMC) algorithm, also integrating out the nonparametric component by a so called Polya urns Gibbs scheme (see, for instance the research by Neal [25]). For more details on model based cluster analysis in Bayesian nonparametric setting we refer to [26, 27].

Bayesian models for cluster analysis are becoming more and more popular even in the genetic epidemiological and biomedical literature. Among the other papers, DPM models with Gaussian kernels are used to cluster microarray gene expression data [28, 29, 30]. Our approach differs from the previous papers since SNP genotypes take only three possible values and thus we consider a multinomial mixture model [31, 32]. It is worth mentioning that our goal is very similar to the one in [31], although with a different approach. They clustered individuals in groups (e.g., high risk, average risk and low risk for a certain disease) and then identified the covariates which were influent in clustering with DPM. In our approach the procedure is reversed, since we first cluster the SNPs according to a DPM model with multinomial kernels, and then we investigate which groups of SNPs affect the disease risk of an individual. In [33] we presented a model similar to the one discussed in this paper, by considering a wide class of processes, namely the normalized generalized gamma processes (NGG), as mixing distribution in the hierarchical mixture model. We stress here that in the current paper the focus is on the application, i.e., the association study between groups of SNPs and a disease, while in the previous work we were more interested in proving the feasibility of normalized generalized mixture model in addressing real problems, in modelling

and, furthermore, in providing a review of the model and its current applications.

More in detail, in the current work SNP genotypes are categorical and thus the codings do not affect the inference. Following the allocation of SNPs in clusters, we compute the genetic score of each cluster to investigate the cluster effects under the generalized linear mixed effect model (GLMM). The risk of each cluster can be evaluated based on its corresponding posterior probability and, in addition, the effect of each single marker inside the cluster can be evaluated as the mean of a suitable posterior functional.

The rest of the paper is organized as follows. In Section 2 we present a Bayesian nonparametric approach which clusters SNPs based on the observed numbers of counts of minor alleles via a Dirichlet process mixture model. We also give some detail on the Gibbs sampler to perform posterior inference and to compute the posterior clustering based on the so called Binder loss information. In Section 3 we propose a genetic score to investigate the cluster effect through a link function in GLMM. Each cluster can be identified to be positively or negatively associated with the disease phenotype based on its corresponding posterior probabilities and single SNP effects. In Section 4 we apply the analyses to a study of Crohn's disease from Wellcome Trust Case Control Consortium [34]. Results from the proposed method are compared with other analyses to evaluate the performance. Concluding remarks are given in Section 5.

2. Bayesian Nonparametric Model-based Clustering Algorithm

The SNP data we are going to consider belong to M different chromosome regions, and in this work we are going to suppose that clustering of SNPs are independent across different regions; it is well recognized that different, non adjacent, chromosome region may not be passed together from parents to offspring due to the so called random crossover, so that independence among different regions may be assumed. From a modeling point of view, we are then going to fit M independent Dirichlet mixture models, one for each region.

However, we simplify the notation just describing the Bayesian nonparametric mixture, the clustering results and the generalized linear model within a single chromosome region, so that we can suppress the index denoting the chromosome region. We underline, however, that the analysis in the application section is performed by considering all the M chromosome regions at hand.

Suppose the genotypes of m SNPs have been collected from n subjects to form the data matrix $\mathbf{X}$, which contains the column vectors $\mathbf{X}_i$ for $i = 1, \ldots, m$, where $\mathbf{X}_i = (X_{1i}, \ldots, X_{ni})$ denotes the genotype coding of the $i-$th SNP for all

n subjects. Here $X_{pi} \in \{0, 1, 2\}$, for $i = 1, \ldots, m$ and $p = 1, \ldots, n$, indicates the number of minor alleles that the $p-$th subject carries at the $i-$th SNP. That is, $X_{pi} = 0$ for genotype AA, 1 for Aa, and 2 for aa if a is the minor allele of this SNP.

2.1. Dirichlet Process Mixture Model

Let $S_{ij} = \sum_{p=1}^{n} I(X_{pi} = j)$ be the total number of subjects whose genotypes on SNP i are coded with j, where $j \in \{0, 1, 2\}$. For $i = 1, \ldots, m$, we model $\mathcal{S}_i = (S_{i0}, S_{i1}, S_{i2})$ with conditionally independent multinomial distributions, given $\underline{\theta}_i = (\theta_{i0}, \theta_{i1}, \theta_{i2})$, i.e.

$$\mathcal{S}_i = (S_{i0}, S_{i1}, S_{i2}) \sim Mult(n, \underline{\theta}_i = (\theta_{i0}, \theta_{i1}, \theta_{i2})).$$

Instead of considering parametric prior distributions on $\underline{\theta}_i$'s, we suppose they are generated by a Dirichlet process (DP), and a hierarchical model is obtained:

$$\begin{aligned} &\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m | \underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_m \sim \prod_{i=1}^{m} Mult(n, \underline{\theta}_i) \\ &\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_m | P \text{ iid } P \\ &P \sim DP(\alpha P_0) \end{aligned} \tag{1}$$

where $\alpha \in \mathbb{R}^+$ and $P_0(\cdot)$ is a Dirichlet distribution $Dir(a_0, a_1, a_2)$ with $a_0, a_1, a_2 \in \mathbb{R}^+$.

Note that the relation among SNPs is modeled through the parameters $\underline{\theta}_i$'s. In other words, it is assumed that, if the distributions of SNP genotypes are very likely to be similar, then their $\underline{\theta}_i$'s are similar. Therefore, SNPs with similar genotype distributions may be clustered together based on the above Dirichlet process mixture model. If all cases carry the AA genotype and all controls carry the aa, the genotype probability $\underline{\theta}$ of this SNP becomes (0.5, 0, 0.5). This SNP can distinguish precisely patients from healthy subjects. When studying complex diseases, however, the components of genotype probability for most SNPs are often nonzero.

We provide a brief description of the Dirichlet process to explain more accurately the law of the random partition of the data implicitly defined by a DPM. For this purpose, we use the *stick breaking* representation of a DP [35]. A realization P of the Dirichlet process of parameter αP_0, is almost surely a discrete probability

measure obtained from

$$P(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{\tau_j}(\cdot)$$

where the (support) points $\tau_1, \tau_2, \ldots$ are an iid sequence from P_0, while the random jumps $w_1, w_2, \ldots$ are obtained by a *stick* of unit length broken into pieces in a sequential manner: $w_j = v_j \prod_{l=1}^{j-1}(1 - v_j)$, and v_j for $j = 1, 2, \ldots$ is an iid drawn from a Beta$(1, \alpha)$ distribution. Since P is almost surely a discrete probability measure, then we observe ties with positive probability in a sample of iid $\underline{\theta}_1, \ldots, \underline{\theta}_m$ from P. We denote by $\underline{\psi}_1, \ldots, \underline{\psi}_K$ the unique values among the $\underline{\theta}'s$ while $m_k, k = 1, \ldots, K$, is the number of times the value $\underline{\psi}_k$ appears among $\underline{\theta}_1, \ldots, \underline{\theta}_m$. Therefore, the discreteness of the Dirichlet process induces a partition $\rho = \{C_1, \ldots, C_K\}$ among the data, with the $i-$th SNP belonging to the cluster C_k if and only if $\underline{\theta}_i = \underline{\psi}_k$, where $i = \{1, \ldots, m\}$ and $k \in \{1, \ldots, K\}$. Clearly here the partition ρ is random, and it will be the quantity of main interest in our analysis. In particular, the distribution of the random partition ρ (our prior on the main parameter) can be expressed by the so called exchangeable partition probability function (eppf) given by

$$\begin{aligned} \pi(\rho) &= Pr(\sharp C_1, \sharp C_2, \ldots, \sharp C_K) \\ &= \text{eppf}(m_1, m_2, \ldots, m_K, \alpha) \\ &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+n)} \alpha^{K-1} \prod_{k=1} K(m_k - 1), \end{aligned} \tag{2}$$

where $\sharp C_k = m_k$, $k = 1, \ldots, K$, denotes the number of elements in the cluster C_k.

As a trivial consequence, in our modeling, the number of clusters is random, and it can be shown that (2) leads to the following probability mass function for K:

$$Pr(K = k) = S_m(k) \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}, k = 1, \ldots, n$$

where $S_m(k)$ is the absolute value of the Stirling number of the first kind, see [36]. From the modeling point of view, it is very important to observe that the joint marginal law of $\underline{\theta}_1, \ldots, \underline{\theta}_m$ is uniquely characterized in terms of the joint law of the random partition ρ and the unique values $\underline{\psi}_1, \ldots, \underline{\psi}_K$ [24]. In other words, it

is possible to write the DPM model (1) in an equivalent way as

$$\begin{aligned}
&\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m | \rho, \underline{\psi}_1, \ldots, \underline{\psi}_K \sim \prod_{k=1}^{K} \prod_{i \in C_k} Mult(\mathcal{S}_i, \underline{\psi}_k) \qquad (3)\\
&\underline{\psi}_1, \ldots, \underline{\psi}_K | K \sim P_0(\cdot)\\
&\rho \sim \text{eppf}(m_1, \ldots, m_K, \alpha)
\end{aligned}$$

where $P_0(\cdot)$ is $Dir(a_1, a_2, a_3)$, and the mathematical expression of the eppf has been given in (2). Hierarchical mixture models as (1) are frequently adopted in the Bayesian nonparametric framework for their mathematical tractability. However, in this paper we find more convenient to express our model as in (3) (i.e., by integrating out the infinite dimensional parameter P). In fact, from one hand this representation is quite expressive since the random parameter contains the partition ρ, which is the object of our statistical analysis; on the other, we will exploit (3) in order to set up an efficient Gibbs sampler.

From the Bayesian nonparametric literature it is well known that the posterior cluster estimation under DPM model is strongly affected by the choice of the *mass* parameter α of the Dirichlet process. To make the inference more robust with respect to this hyperparameter we added a level of hierarchy in our model by taking α random; we choose a $gamma(0.1, 0.1)$ non-informative prior on it. It is worth noting that the gamma prior has been chosen since it leads to a closed form of the full conditional distribution in the MCMC algorithm.

A key aspect of the DPM model, particularly relevant to understand the cluster structure we are inducing among the SNP through the DPM model, is the so called Chinese restaurant process. We can look at this process as a procedure to sample the random partition ρ from the $eppf$ in (2). The process is essentially similar to a sequential restaurant "seating arrangement", as described next. Suppose customers arrive sequentially at a Chinese restaurant and they are randomly assigned to a table chosen from an unlimited number of circular ones $C_1, C_2, \ldots$, each with an unlimited capacity to seat customers. We label the selected tables with the values $\underline{\psi}_1, \underline{\psi}_2, \ldots$, with the index denoting the order of arrival of the customers. By default, the first customer is always seated at the first table (C_1), which is labeled with $\underline{\theta}_1 = \underline{\psi}_1$. Subsequently, for $r \geq 1$, the customer $r+1$ is seated according to the following prediction rule (often called the Blackwell-MacQueen Polya urn rule), applied to the partition $\rho_r = \{C_{1,r}, ..., C_{k,r}\}$ of $1, \ldots, r$ corresponding to the seating arrangement of the first r customers. That is, if ρ_{r+1} is the event that

customer $r+1$ is seated at a previous table $C_{j,r}$, then ρ_{r+1} occurs with probability

$$Pr(\rho_{r+1}|\rho_r) = \frac{\pi(\rho_{r+1})}{\pi(\rho_r)} = \frac{\#C_{r,j}}{\alpha + r}. \tag{4}$$

On the other hand, if ρ_{r+1} is the event that customer $r+1$ is seated at a new table, then the event ρ_{r+1} occurs with probability

$$Pr(\rho_{r+1}|\rho_r) = \frac{\pi(\rho_{r+1})}{\pi(\rho_r)} = \frac{\alpha}{\alpha + r} \tag{5}$$

After m steps this process results in a partition $\rho = \rho_m$ from the eppf in (2).

As made evident from the Chinese restaurant process, the *a priori* cluster assignment is independent of the values of the non-observable labels $\underline{\theta}$'s; it is worth noting that, however, this is not true *a posteriori*. Two SNPs will be clustered together when mixing two contributions: Chinese restaurant process and sharing of similar multinomial distributions.

2.2. Computation and Inference of Partition and Clusters

To perform posterior inference, we set up a "standard" Polya urn Gibbs sampler for DPM mixture model (see [37] and [38] for details). We consider the formulation of our model given in (3) and we build a Gibbs sampler on the space of parameters, namely $(\rho, \underline{\psi}_1, \ldots, \underline{\psi}_K)$. To do that, we first observe that

$$\mathcal{L}(\rho, \underline{\psi}_1, \ldots, \underline{\psi}_K | \mathcal{S}_1, \ldots, \mathcal{S}_m) \propto \prod_{k=1}^{K} \prod_{i \in C_k} Mult(\mathcal{S}_i, \underline{\psi}_k) \times \prod_{k=1}^{K} P_0(d\underline{\psi}_k)\pi(\rho). \tag{6}$$

We can integrate out $\underline{\psi}_1, \ldots, \underline{\psi}_K$ from the last expression obtaining

$$\mathcal{L}(\rho | \mathcal{S}_1, \ldots, \mathcal{S}_m) \propto \prod_{k=1}^{K} m(\mathcal{S}_{C_k})\pi(\rho),$$

where $\mathcal{S}_{C_k} = \{\mathcal{S}_i; i \in C_k\}$ is the set of observations in cluster C_k, and $m(\mathcal{S}_{C_k})$ is the joint marginal of this set under the Bayesian multinomial-Dirichlet model, that is:

$$m(\mathcal{S}_{C_k}) = \int \prod_{i \in C_k} Mult(\mathcal{S}_i, \underline{\psi}) P_0(d\underline{\psi}) = \frac{\Gamma(a)}{\Gamma(a + nm_k)} \prod_{j=0}^{2} \frac{\Gamma(a_j + m_{j,C_k})}{\Gamma(a_j)}$$

where $m_{j,C_k} = \sum_{i \in C_k} S_{ij}$.

For each $i = 1, \ldots, m$, let ρ^{-i} be the partition with i removed. We can update ρ using Gibbs sampling, so that the cluster assignment of one SNP frequency $\mathcal{S}_i$ is updated at each step. To formalize the Gibbs sampler, we denote the cluster assignment of $\mathcal{S}_i$ with a variable z_i such that $z_i = C$ denotes the event that $\mathcal{S}_i$ is assigned to cluster $C \in \rho^{-i}$, and $z_i = \emptyset$ denotes the event that it is assigned a new cluster. By using (4) and (5) we have that for each $i = 1, \ldots, n$,

$$Pr\left(z_i = C | \rho^{-i}, \mathcal{S}_1, \ldots, \mathcal{S}_m\right) \propto \begin{cases} \frac{\#C}{\alpha+m-1} \frac{m(\{\mathcal{S}_i\} \cup \mathcal{S}_C)}{m(\mathcal{S}_C)} & \text{for } C \in \rho^{-i} \\ \frac{\alpha}{\alpha+m-1} m(\{\mathcal{S}_i\}) & \text{for } C = \emptyset \end{cases} \tag{7}$$

Therefore, the clustering parameter ρ is updated at each step of the Gibbs sampler using (7). To complete the design of the algorithm, we have to specify how to update the parameters $\underline{\psi}_1, \ldots, \underline{\psi}_K$. It is easy to realize from (6) that, conditionally on ρ, those parameters are independent and

$$\mathcal{L}(\underline{\psi}_k | \rho, \mathcal{S}_1, \ldots, \mathcal{S}_m) \propto \prod_{i \in C_k} Multi(\mathcal{S}_i, \underline{\psi_k}) P_0(d\underline{\psi}_k), \quad k = 1, \ldots, K.$$

Since the Dirichlet distribution P_0 is conjugate with respect to the multinomial model, $\mathcal{L}(\underline{\psi}_k | \rho, \mathcal{S}_1, \ldots, \mathcal{S}_m)$, for each k, is the law of Dirichlet distributions with parameter $(a_{0,C_k}, a_{1,C_k}, a_{2,C_k})$, where $a_{j,C_k} = a_j + m_{j,C_k}$, $j = 1, \ldots, 3$, for each $k = 1, \ldots, K$.

Model (3) highlights that both the estimation of the number of clusters and the allocation of SNPs to clusters are possible once the partition ρ is determined. Here we consider posterior means, i.e., the Bayesian estimators of parameters, since they minimize the posterior expected loss under a squared loss function. To estimate the partition ρ, we adopt the Binder loss minimization method [39].

In practice, we look for a loss function $L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}})$ giving the cost of estimating the "true" $\boldsymbol{\rho}$ by $\hat{\boldsymbol{\rho}}$. Then the proposed estimate is given by any $\hat{\boldsymbol{\rho}}$ which minimizes the posterior expectation of the loss function, i.e.

$$\hat{\boldsymbol{\rho}} \in \arg\min_y \mathbb{E}[L(\boldsymbol{\rho}, y) | data].$$

We consider the so called Binder loss function [39], which assigns the cost b when two elements are erroneously clustered together and the cost a when two elements

are wrongly assigned to different clusters, i.e.

$$L(\boldsymbol{\rho}, \hat{\boldsymbol{\rho}}) = \sum_{i<j\leq n} \left(a\, \mathbf{1}\{\mathcal{S}_i \overset{\boldsymbol{\rho}}{\leftrightarrow} \mathcal{S}_j, \mathcal{S}_i \overset{\hat{\boldsymbol{\rho}}}{\nleftrightarrow} \mathcal{S}_j\} + b\, \mathbf{1}\{\mathcal{S}_i \overset{\boldsymbol{\rho}}{\nleftrightarrow} \mathcal{S}_j, \mathcal{S}_i \overset{\hat{\boldsymbol{\rho}}}{\leftrightarrow} \mathcal{S}_j\} \right), \qquad (8)$$

where $\overset{\boldsymbol{\rho}}{\leftrightarrow}$ and $\overset{\hat{\boldsymbol{\rho}}}{\leftrightarrow}$ stand for the equivalence relations induced by the partitions $\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}$, respectively.

To minimize $L(\rho, \hat{\rho})$, Lau and Green [26] proposed a sophisticated optimization method considering a binary integer programming problem; their method is computationally infeasible unless the sample size is really small. Here we consider a simpler method: we run the MCMC chain once in order to estimate the posterior incidence probabilities $\mathbb{P}(\mathcal{S}_i \overset{\boldsymbol{\rho}}{\leftrightarrow} \mathcal{S}_j | data)$, then we plug the estimates in the posterior mean of (8) and run the MCMC algorithm a second time, obtaining posterior sample configurations. Finally, we choose $\hat{\boldsymbol{\rho}}$ as the configuration that minimizes the expected loss among the previously sampled ones. Of course, the result is affected by the value selected for $b/(a+b)$, which can be seen as the proportion of the cost to pay by putting together two elements, when they should be actually separated. In this work, 0.5 is fixed so that the two costs are equally weighted.

The resulting estimate of ρ completes, at the same time, the inference on the number of clusters, the assignment of SNPs to each cluster, as well as the genotype probability $\underline{\psi}_k$ for each cluster. These marker-sets can now be used to perform the multiple-marker association studies.

3. Association Studies for Clusters and SNPs

Once we have fitted and implemented a DPM model for each chromosome region in our data, we have a point estimate of the K clusters in each region, $\hat{\rho} = (C_1, \ldots, C_K)$, and the parameter identifying the multinomial kernel density of the data within each cluster $\boldsymbol{\psi} = (\underline{\psi}_1, \ldots, \underline{\psi}_K)$. For each SNP-set cluster C_k, we construct a genetic score $G_{p,k}$ for the $p-$th subject,

$$G_{p,k} = \frac{\ln Pr(\mathbf{X}_{p,C_k} | \underline{\psi}_k)}{\sharp C_k} \qquad (9)$$

where $Pr(\mathbf{X}_{p,C_k} | \underline{\psi}_k) = \prod_{X_{ip} \in C_k} \psi_{k0}^{I(X_{ip}=0)} \psi_{k1}^{I(X_{ip}=1)} \psi_{k2}^{I(X_{ip}=2)}$ is the product of all genotype probabilities of the SNPs belonging to the cluster C_k for the $p-$th subject

and $\sharp C_k$ is the cluster size, i.e the number of SNPs in the cluster. The denominator $\sharp C_k$ is designed so that the various cluster sizes can be standardized.

The key point of our analysis is the following: we want to use the result of the clustering procedure as an input for an association study where the regressors are obtained by the genetic score (9) [33]. A scheme of our modelling idea is given in Figure 1. After the SNPs covariates are clustered as shown at the bottom of the figure, the genetic scores are calculated for each cluster and then used in the GLMM. Conditionally on the partition, for each patient, we can compute

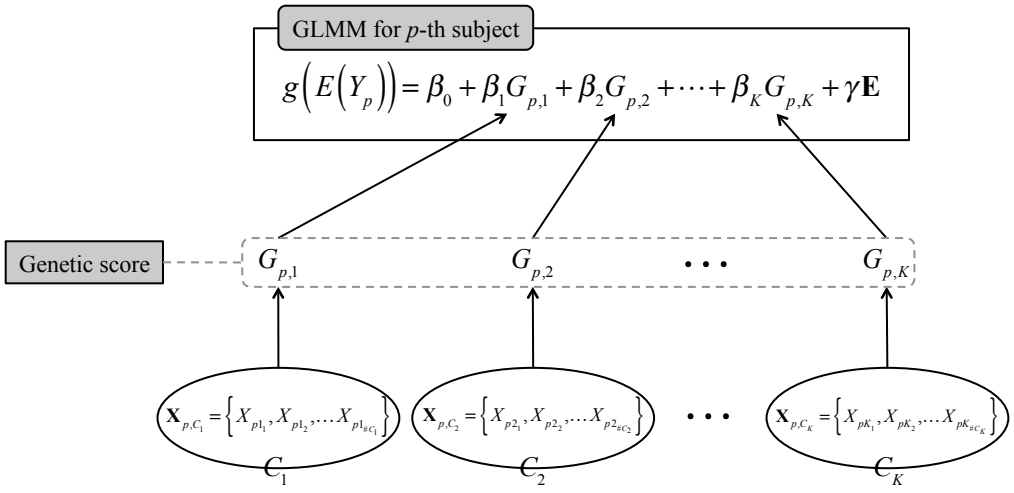

Figure 1: Scheme of our model. The bottom part indicates the clustering of the SNP covariates derived under the DPM model. In the upper panel, the genetic scores are obtained via formula (9).

the (standardized) log-probability of observing the SNP configuration within each cluster, that is our genetic score. In this way we can build the covariate matrix (the G's values on the right hand side of Figure 1) and set up a GLMM to study the association. To be more formal, let Y_p be the disease status indicator variable for patient $p = 1, \ldots, n$. We consider the following GLMM

$$g(E(Y_p)) = \beta_0 + \beta_1 G_{p,1} + \ldots + \beta_K G_{p,K} + \boldsymbol{\gamma}\mathbf{E}, \tag{10}$$

with link function g, genetic scores $G_{p,k}$, and possibly other environmental covariates $\mathbf{E}$. For instance, in presence of binary disease status ($Y_p \in \{0, 1\}$), as in the application here, a logit link can be assumed for the link function g while an

identity link function can be considered for quantitative traits.

A vague prior is assumed on the regression parameters $(\beta_1, \ldots, \beta_k)$ (i.e, independent Gaussian with mean 0 and variance 100), and the posterior distribution is computed via a MCMC algorithm. We resort to a variable selection procedure [hard shrinkage, 40] to study the association between the disease and the clusters of SNPs: a group C_k is not significative if the posterior 90% credible interval of the corresponding parameter β contains zero; if the posterior credible interval is entirely contained in $\mathbb{R}^+$ ($\mathbb{R}^-$), then it denotes positive (negative) association between the clusters and the disease phenotype Y_i.

The likelihood in (10) is conditioned not only on the parameters β_p's, but also on the partition ρ and the corresponding ψ's. Therefore, under this model, we may alternatively consider two strategies to perform the association study. On one hand we can plug in the Bayesian point estimates $\hat{\rho}$ and $\hat{\underline{\psi}}$ to compute first the genetic score (9) and later infer through the regression model. In this way we take into account the collective association effect of multiple markers (SNPs) to the disease. On the other hand, we can incorporate the uncertainty on the clustering in our association study, i.e. we use the information contained in the whole posterior distribution of ρ. In this case observe that, given ρ and ψ, once we have chosen a rule to classify a cluster (whether associated or not to the disease), for each $j = 1, \ldots, m$, we can define single marker effect indexes as

$$A_j := A_j(\rho, \underline{\psi}) = \begin{cases} 1 & \text{if SNP}j \text{ belongs to a positively associated cluster} \\ 0 & \text{if SNP}j \text{ belongs to a non associated cluster} \\ -1 & \text{if SNP}j \text{ belongs to a negatively associated cluster.} \end{cases}$$

If $(\rho^{(1)}, \underline{\psi}^{(1)}), \ldots, (\rho^{(G)}, \underline{\psi}^{(G)})$ is a sample from $\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\psi}|X)$, we can evaluate the posterior frequencies that $A_j = a$ with $a = -1, 0, 1$. If the mode of these posterior frequencies, for each SNP j, occurs at -1, 1, and 0, then the SNP is classified as carrying negative, positive, or no association with the disease, respectively. Notice that here we evaluate the strength of association for each cluster by estimating the β's regressors, but only infer the effect direction for each single SNP marker.

4. Application Study of Crohn's Disease

We utilized the Crohn's disease (CD) dataset from the Wellcome Trust Case Control Consortium study (WTCCC) [34] as an example to demonstrate the performance of our proposed method. This dataset included 1748 patients with CD and 2938 shared controls. We only analyzed the 3737 SNPs with minor allele

frequency (MAF) ≥ 0.01 in the control group and in Hardy-Weinberg equilibrium located on chromosome 1p31.3, 2q37.1, 5p13.1, 10q24.2 and 16q12.1. These regions were selected as candidate regions because previous studies have shown a strong association between these regions and CD [34, 20]. The genotype coding is assigned based on the minor allele copies; that is, as defined earlier, S_{ij} represents the number of subjects carrying j ($j = \{0, 1, 2\}$) minor alleles on the $i-$th SNP.

4.1. Clustering Structure

We ran twice the Gibbs sampler of Section 2.2 to fit the DPM model for each of the five chromosome regions. The first chain is obtained after a burn-in of 1000 iterations, a thinning of 5 iterations and a final sample size of 4000. We use this chain to estimate the incidence probability $Pr(\mathcal{S}_i \overset{\rho}{\leftrightarrow} \mathcal{S}_j|data)$, then we plug the estimates in the posterior mean of (8) and run the Gibbs sampler a second time for an extra 1000 iterations (thinned every 5) obtaining a posterior sample configurations (see Section 2.2). Convergence was checked by standard statistical test available by the `coda` R package.

Our population consists of 3737 SNPs from the five chromosomes. The numbers of clusters K for five chromosomes are estimated as 28, 31, 32, 28, and 37, respectively, through the Binder loss information.

The posterior histogram of K for each chromosome is displayed in Figure 2 (A)-(E). Figure 3 (A)-(E) displays the observed genotypes of each SNP, colored according to its corresponding cluster assignment. The axes of the plots reported there are encodings for the frequencies of 1 or 2 minor alleles (i.e., Aa and aa), so that each point is a representation of a $\mathcal{S}$ data belonging to the 2-dimensional simplex. In fact, our clustering model is based on a "closure" (similarity) relationship between SNPs in the 2-dimensional simplex, so that by visualising these scatter plots we gain a spacial intuition of our clustering. As such, neighboring points share similar observed genotype frequencies $\underline{\theta}_i$ and have been clustered together. Some clusters are large, containing as many as 45 SNPs, while some are small with only two SNPs. Table 1 lists the descriptive statistics for the number of SNPs assigned to clusters and the number of clusters per chromosome region.

4.2. Association of Clusters

We used a generalized linear mixed model with a logit link (eq. (10)) to analyze the association between SNP clusters and the disease of interest. To fit the Bayesian logistic regression we used the `MCMCpackage` R-package. Our results are based on chains obtained after 1000 iterations of burn-in, a thinning of 5 and

Table 1: Descriptive statistics for SNPs and clusters per chromosome region.

Region	No. of SNPs	$\hat{k}$	Cluster size			Association	
			min	median	max	positive	negative
1p31.3	1357	37	12	19	36	6	5
2q37.1	662	31	7	22	45	4	8
5p13.1	554	32	2	15	44	3	5
10q24.2	390	28	2	12.5	36	0	4
16q12.1	742	37	1	17	45	4	7

a final simple size of 5000. As earlier, convergence of the posterior chain was checked via the `coda` library.

The strength of association for each cluster is assessed with the posterior distributions and 90% credible intervals of β_k, as seen in Figure 4 (A)-(E). Intervals lying above the zero horizontal line indicate a deleterious effect from the corresponding cluster; while intervals below represent protective effects. For instance, the first cluster in 10q24.2 suggests an overall protective effect for SNPs in this cluster, with an estimated odds ratio $\exp(-0.21) = 0.81 = 1/1.23$, indicating an approximate 20% reduction in risk with one unit increase in the genetic score $G_{p,1}$ for the p-th subject. The numbers of deleterious and protective clusters are listed in the last two columns of Table 1. In summary, among the five chromosomes 17 clusters show a strong deleterious effect in association with CD and 29 clusters present protective association with CD. These 46 important clusters exert effects to different degrees. Taking 1p31.3 for example, the strength of such association can be as large as $\exp(+0.96) = 2.61$ for the second cluster in 1p31.3 region and $\exp(-0.63) = 0.53$ for the 16-th cluster in 1p31.3. For each of the five chromosome regions, there are 11, 12, 8, 4, and 11 important clusters, respectively, as seen in Figure 4 (A)-(E). We refer to these clusters as significant clusters.

To assess the estimates of $\underline{\psi}_k$, the average genotype frequency vector in cluster k, we display the fitting performance on the left hand side of all the panels in Figure 5. Axes Y of the plots indicate the deviation of the estimate $\hat{\psi}_{k2}$ from the observed percentage of S_{i2} if the $i-$th SNP belongs to cluster C_k. Axes X report the identifier of the Binder clustering (i.e., cluster 1, cluster 2, and so on). All deviations are very small, indicating a good fit. The colors in the left panels represent the effect direction of the clusters. Thus all SNPs in the same cluster share the same color. For instance, the gray color for the first cluster in 1p31.3 indicates no association with the disease. However, the red color for the second cluster implies that the SNPs in this cluster collectively exert a positive association

with the disease. That is, the larger the genetic score for this cluster the higher the risk in developing the Crohn's disease.

All these significant clusters were later incorporated together in a logistic regression model to evaluate the change in effects if all clusters are fitted together. Figure 6 contrasts the 90% credible intervals when the clusters were fitted separately versus the fit altogether. These probability intervals are similar regardless of the approach. Thirty-eight intervals remain significant when all genetic regions were fitted collectively. The 46 significant clusters include 106 genes of which 16 genes have been reported previously in literature as associated with Crohn's disease, including *ATG16L1, IL23R, NOD2* and *PTGER4* [41, 42, 43, 44, 45]. The clusters with larger effect size on CD contain many SNPs in such genes. For example, Cluster 2 and 5 in 1p31.3 are part of the gene *IL23R*. In addition, there are six SNPs in significant clusters, which have been identified as associated with CD. Located in gene *IL23R* in 1p31.1, rs11805303, rs2201841 and rs10889675 are in Cluster 16, 16, and 26, respectively. Rs6871834 is in Cluster 11 (5p13.1), and rs2066843 and rs18617559 are in Cluster 11 and 27 (16q12.1), respectively.

4.3. Effect Direction of SNPs

Once the influential clusters are identified, the next question would be to select the important SNPs, e.g., those carrying protective effect or considered risk variants. This could be useful if the target therapy is of interest. To determine the effect direction of each single SNP after accounting for the uncertainty in the cluster assignment, we performed the Monte Carlo procedure described at the end of section (3). The right hand side plots in all the panels of Figure 5 demonstrate the effect direction of each SNP, with red for deleterious effect, green for protective and gray for none. It is clear that, for most clusters, their component SNPs show consistent effect direction. When comparing left and right plots of each panel in Figure 5, it is clear that the colors remain the same for most SNPs even after accounting for the uncertainty in the cluster assignment. This is an indication of robustness for this nonparametric clustering procedure. Specifically 30 clusters remain important and thus deserve special attention for further examination. The significant clusters are the ones labeled 15, 16, 17, 26, 35 in 1p31.3; the ones labeled 1, 3, 6, 9, 11, 14, 22, 23, 26 in 2q37.1; the ones number 2, 6, 11, 17, 18, 23, 25, 27 in 5p13.1; the ones labeled 1, 9, 13 in 10q24.2 and finally the ones labeled 1, 3, 6, 7, 9 in 16q12.1.

These important clusters reveal interesting findings. For instance, the important clusters numbered 2, 5, 15, 16 and 26 are all in 1p31.3, and they are all part of the gene *IL23R*. These clusters, however, present effects of different directions:

cluster 16 is protective while clusters 2, 5, 15 and 26 are deleterious. Such discordant effect directions may be the reason why inconsistent findings about the association between *IL23R* and Crohn's disease have been reported in literature [42, 46, 47]. As an illustration, we list in Table 2 for 1p31.3 the estimates of each cluster effect and the gene symbols containing the SNPs. Table S1 displays the information for the other four regions.

4.4. Comparison analysis

To evaluate the performance of the proposed DPM method with other analyses of genetic association studies, we consider the traditional single-marker analysis with χ^2 test, Bayesian mixture model with hybrid procedures (BMIX) [20], sequence kernel association test (SKAT) [48] with linkage disequilibrium blocks (termed as LD.SKAT hereafter), and the normalized generalized gamma mixture model (NGG) [33].

Table 3 lists the number of clusters and influential SNPs identified by these methods. The single-marker test is performed on every SNP, not on clusters; while the BMIX simply classifies all SNPs into two groups, one associated with the disease and the other not. Therefore, these two methods do not identify clusters. The SKAT is used on sets of markers and thus the SNPs are grouped first to form blocks based on linkage disequilibrium statistics and Haploview [49]. Each haplotype block is then tested with SKAT. If the block is statistically significant, then all SNPs in this block are considered influential. Note that the single-marker test and LD.SKAT are corrected with Bonferroni's corrections to prevent the inflation of type I errors. The NGG mixture model is similar to DPM; however, it considers a wider class of mixing distributions.

The last column in Table 3 represents the number of common SNPs identified by at least three out of the five methods. These common 65 SNPs are located in 8 genes known to be linked to CD, including *IL23R*, *IL12RB2*, *CYLD*, *NRU2*, *ATG16L1*, *RPL37*, *NKX2-3* and *NOD2*, while some SNPs are in intergenic regions. This finding confirms the ability of DPM to locate the susceptible genes. Furthermore, the DPM identifies genes that hold potential to be considered in treatments. For instance, the *PTGER4* in 5p13.1 was recently reported to be associated with the Crohn's disease [50] but only DPM and NGG pick up this signal. Therefore, these influential SNPs may deserve further investigation.

These five methods can indeed be categorized into two groups. The first group contains the single-marker χ^2 test and BMIX. Both tests do not consider clustering structure. The single-marker χ^2 test considers all SNPs independent; while

in BMIX the SNPs are dependent. The prior distribution assumed in BMIX simply takes into account that all SNPs can be grouped into two groups, one with associated SNPs and the other without. These two methods identify 50 SNPs in common. The high concordance between these two results from the slight difference in their assumptions. The second group contains the other three methods; all assume clustering structure. The clusters in LD.SKAT have to contain nearby SNPs, limiting the inclusion of related SNPs located in non-neighboring areas. In addition, the important SNPs are defined by LD.SKAT if they are located in significant clusters. Therefore, LD.SKAT has reported more important SNPs than the previous χ^2 and BMIX test. It shares 54 SNPs in common with χ^2 and 45 in common with BMIX. The other two methods, NGG and DPM, put no restriction on the distance between SNPs in the same cluster. These two methods are more similar to each other except that they assume different prior distributions for cluster allocation. This similarity in models leads to the high concordance in their findings; a total of 616 SNPs are in common. However, the large number of important SNPs should be taken carefully. In both NGG and DPM, the SNPs are defined important based on the modes (positive, negative, and none) instead of the posterior probabilities. That is, more SNPs are estimated to be associated with the disease under our Bayesian nonparametric method, but they might not be if probability is used. The major purpose of NGG and DPM is to find significant clusters. NGG has 25, 22, 39 SNPs in common with χ^2, BMIX, and LD.SKAT, respectively; while DPM has 23, 20, 38 SNPs in common with χ^2, BMIX, and LD.SKAT, respectively.

5. Discussion

In summary, we proposed the Bayesian Dirichlet process mixture model to cluster SNPs via their genotype frequencies, followed by association tests with the Bayesian logistic regression model with genetic scores for SNP clusters. The advantages of this approach include no pre-specification of the number of clusters, statistical inference of this number, incorporation of the uncertainty in SNP allocations to clusters in analysis of associations, and posterior inference of the susceptibility for each cluster and SNP. The strength of the joint effects of clustered SNPs can be evaluated; while any single SNP effect direction can be assessed concurrently as well. These two purposes can both be pursued in the analysis, unlike most existing approaches that simply focus on one of them.

Our proposed procedure will cluster SNPs of similar genotype frequencies, regardless of their effect directions. Take two SNPs for example. If all patients

carry AA and all healthy controls carry aa genotype, while all patients carry bb and all controls carry BB genotypes, then both SNPs correspond to the same genotype frequency vector $\underline{\theta} = (0.5, 0, 0.5)$, and will be clustered together, even though these two SNPs are of completely opposite effect directions. In other words, a cluster may contain both positive and deleterious SNPs. This phenomenon has been observed in Figure 5, especially when MAF is around 50%. However, if the component probabilities are less balanced, such as (0.6, 0, 0.4) and (0.4, 0, 0.6) for SNP A and SNP B, then the clustering procedure will pick up the difference and will not cluster them. Therefore, SNPs of the same effect direction would be clustered, as seen in other clusters containing the same color of SNPs. In general, there should be no constraint in clustering SNPs of the same or different effect directions, when no biological assumption is made. Users of the proposed model should bear in mind this limitation. An alternative illustration would be to plot the single SNP effect direction versus the currently proposed approach.

Although the approach is nonparametric, the resulting clustering structure clearly carries some biological information. For instance, in the Crohn's disease study, the SNPs effects are of the same direction in most clusters. They are mostly all protective, or deleterious, as shown in the lower panel of Figure 5. In addition, if different SNPs in the same gene exert effects of different directions, then they will be clustered into different clusters. A good example is the interleukin 23 receptor (*IL23R*) in 1p31.3, where cluster 16 is protective but four other clusters (numbered 2, 5, 15, and 26) are not. This gene produces the protein that is involved in cell membrane of several immune system cells, and thus is related to inflammation and immune response. Some coding and non-coding variants in *IL23R* have been shown to associate with CD [42, 46]. One meta-analysis confirmed the association between *IL23R* and inflammatory bowel disease (IBD) [47]. Such association may not be observed in Crohn's disease-specific association [47], but Crohn's disease and ulcerative colitis are two common forms of IBD. This observation may provide an explanation for inconsistent results in previous literature, and this gene certainly deserves more investigation in its molecular functions in smaller regions.

The demonstration of this strategy on Crohn's disease identifies 46 important clusters from 3737 SNPs. This clustering information can be used in future analysis of interaction to save computational cost in an exhaustive search for gene-gene interactions and in future laboratory studies for molecular interactions. Such clustering structure can also be considered as complementary results to pathway analysis, if one is more interested in smaller groups of SNPs rather than large sets of genes.

In [33] we presented a similar model to the one discussed in this paper, by considering a wide class of process, namely the normalized generalized gamma process (NGG), as mixing distribution in the hierarchical model (1). The class of NGG processes has been recently introduced in the statistical literature [51]. This class encompasses the Dirichlet process, and has been proved to be very flexible in its clustering ability. Here we would like to highlight pros and cons of the two models.

From the application point of view we point out that the DPM model is a thoroughly explored method, widely used in applications and generally accepted by the statistical community and users of statistical methods in science, unlike the NGG whose efficacy in real applications is still under investigation, although it is quite promising. For instance, the clustering induced by the DPM can be more easily interpreted in term of the Chinese restaurant process, than the one obtained with the NGG: while the allocation of a new observation in the former is determined by weights corresponding to a Polya urn scheme (see formula (4) and (5)), the weights in the latter are obtained using special functions (i.e. integrals unavailable in closed form) and are quite hard to interpret from an application point of view [52]. The different clustering allocation rules, that are dictated by the predictive characterization of the two processes, have implication also on the computational aspects, making the DPM model more attractive. For instance, the Polya urn algorithm to obtain posterior inference in the current work has been implemented with the statistical software **R** (the code is available upon request to the corresponding author), and the computational time is fairly reasonable (around 20 minutes for the nonparametric mixture models for each region); on the other hand the code for the NGG mixture model in [33] is much more elaborated and the authors had to resort to C to obtain comparable computational times.

With this consideration in mind, we point out that the focus here is the application to the Crohn's disease association study, while in the paper [33] we were more interested in proving the feasibility of NGG mixture in real problems, in modelling and, furthermore, in providing a review of the model and its current applications. A thorough comparison between the DPM and the NGG, especially when applied to genetic problems, is planned for the future when we hope to be able to develop a formal test to select the best model, as a complement to the previous statements and the obvious observation that the NGG could improve (but at the costs mentioned above) upon the DPM just because the latter is a particular case of the former.

As far as the considered genetic application is concerned, some extensions and issues of the proposed model are worth of being noted. First, the coding system

demonstrated here was for genotypes. It can be changed to 0/1 coding for carrying specific alleles/genotypes, if the dominance or recessive inheritance model is assumed or if a certain allele is of major interest. Second, we have applied the logistic regression model for the case-control study design in this research. For quantitative phenotypes, this approach can accommodate continuous variables as response vectors. All Bayesian computations are then carried out in a similar manner. Third, in this analysis we considered only common variants with minor allele frequencies larger than 0.01. It is not clear if the proposed approach is robust to rare variants. The major challenge would be the increase in computational load. We are currently investigating this issue in small experiments. Finally, as pointed out in earlier sections, the genetic score is individual and cluster specific. In other words, the configuration of SNPs in each cluster for every individual can be denoted with this genetic score, and the combined effects for SNPs in the same cluster as well as the overall risk across all clusters can be calculated. Such risk can serve as an indicator for health assessment. The utilization of this index requires further investigations.

Acknowledgements

The research was performed with the support of the joint CNR (National Research Council of Italy) - MOST (Ministry of Science and Technology of Taiwan) programme on scientific and technological cooperation. This research was supported in part by NSC 101-2923-B-002-003 (CW, CKH).

[1] P.-C. Chen, S.-Y. Huang, W. J. Chen, C. K. Hsiao, A new regularized least squares support vector regression for gene selection, BMC Bioinformatics 10 (1) (2009) 44.

[2] N. Malo, O. Libiger, N. J. Schork, Accommodating linkage disequilibrium in genetic-association analyses via ridge regression, Am. J. Hum. Genet. 82 (2) (2008) 375–385.

[3] H. Zhou, M. E. Sehl, J. S. Sinsheimer, K. Lange, Association screening of common and rare genetic variants by penalized regression, Bioinformatics 26 (19) (2010) 2375–2382.

[4] L. S. Chen, C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, L. Hsu, Insights into colon cancer etiology via a regularized approach to gene set analysis of gwas data, Am. J. Hum. Genet. 86 (6) (2010) 860–871.

[5] J. Li, K. Das, G. Fu, R. Li, R. Wu, The Bayesian lasso for genome-wide association studies, Bioinformatics 27 (4) (2011) 516–523.

[6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (43) (2005) 15545–15550.

[7] B. Efron, R. Tibshirani, On testing the significance of sets of genes, Ann. Appl. Stat. 1 (1) (2007) 107–129.

[8] J. Hu, J.-Y. Tzeng, Integrative gene set analysis of multi-platform data with sample heterogeneity, Bioinformatics 30 (11) (2014) 1501–1507.

[9] V. K. Ramanan, L. Shen, J. H. Moore, A. J. Saykin, Pathway analysis of genomic data: concepts, methods, and prospects for future development, Trends Genet. 28 (7) (2012) 323–332.

[10] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res. 21 (7) (2011) 1109–1121.

[11] M.-H. Lee, J.-Y. Tzeng, S.-Y. Huang, C. K. Hsiao, Combining an evolution-guided clustering algorithm and haplotype-based lrt in family association studies, BMC Genet. 12 (1) (2011) 48.

[12] Y.-H. Huang, M.-H. Lee, W. J. Chen, C. K. Hsiao, Using an uncertainty-coding matrix in Bayesian regression models for haplotype-specific risk detection in family association studies, PloS One 6 (7) (2011) e21890.

[13] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, P. Drineas, PCA-correlated snps for structure identification in worldwide human populations, PLoS Genet. 3 (9) (2007) e160.

[14] C. Wang, W.-H. Kao, C. K. Hsiao, Using hamming distance as information for snp-sets clustering and testing in disease association studies, PLoS ONE 10 (8) (2015) e0135918.

[15] P. Zhang, X. Wang, P. X.-K. Song, Clustering categorical data based on distance vectors, J. Am. Stat. Assoc. 101 (473) (2006) 355–367.

[16] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. Ser. B-Stat. Methodol. 63 (2) (2001) 411–423.

[17] J. Wang, Consistent selection of the number of clusters via crossvalidation, Biometrika 97 (4) (2010) 893–904.

[18] T. Broderick, M. I. Jordan, J. Pitman, et al., Cluster and feature modeling from combinatorial stochastic processes, Stat. Sci. 28 (3) (2013) 289–312.

[19] A. Y. Lo, On a class of Bayesian nonparametric estimates: I. density estimates, Ann. Stat. 12 (1) (1984) 351–357.

[20] Y.-C. Wei, S.-H. Wen, P.-C. Chen, C.-H. Wang, C. K. Hsiao, A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies, Eur J. Hum. Genet. 18 (8) (2010) 942–947.

[21] J. Wakefield, Bayes factors for genome-wide association studies: comparison with p-values, Genet. Epidemiol. 33 (1) (2009) 79–86.

[22] J. Wakefield, A Bayesian measure of the probability of false discovery in genetic epidemiology studies, Am. J. Hum. Genet. 81 (2) (2007) 208–227.

[23] D. Aldous, Exchangeability and related topics, École d'Été de Probabilités de Saint-Flour XIII1983 (1985) 1–198.

[24] J. Pitman, Some developments of the Blackwell-Macqueen urn scheme, in: T. S. Ferguson, L. S. Shapley, M. J. B. (Eds.), Statistics, Probability and Game Theory: Papers in Honor of David Blackwell, Vol. 30 of IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward (USA), 1996, pp. 245–267.

[25] R. M. Neal, Markov chain sampling methods for dirichlet process mixture models, J. Comput. Graph. Stat. 9 (2) (2000) 249–265.

[26] J. W. Lau, P. J. Green, Bayesian model-based clustering procedures, J. Comput. Graph. Stat. 16 (3) (2007) 526–558.

[27] R. Argiento, A. Cremaschi, A. Guglielmi, A "density-based" algorithm for cluster analysis using species sampling gaussian mixture models, J. Comput. Graph. Stat. 23 (4) (2014) 1126–1142.

[28] M. Medvedovic, K. Y. Yeung, R. E. Bumgarner, Bayesian mixture model based clustering of replicated microarray data, Bioinformatics 20 (8) (2004) 1222–1232.

[29] D. B. Dahl, Modal clustering in a class of product partition models, Bayesian Anal. 4 (2) (2009) 243–264.

[30] A. Fritsch, K. Ickstadt, Improved criteria for clustering based on the posterior similarity matrix, Bayesian Anal. 4 (2) (2009) 367–391.

[31] M. Papathomas, J. Molitor, C. Hoggart, D. Hastie, S. Richardson, Exploring data from genetic association studies using Bayesian variable selection and the dirichlet process: application to searching for gene$\times$ gene patterns, Genet. Epidemiol. 36 (6) (2012) 663–674.

[32] A. Onogi, M. Nurimoto, M. Morita, Characterization of a bayesian genetic clustering algorithm based on a dirichlet process prior and comparison among bayesian clustering methods, BMC bioinformatics 12 (1) (2011) 263.

[33] R. Argiento, A. Guglielmi, C. Hsiao, F. Ruggeri, C. Wang, Modeling the association between clusters of snps and disease responses, in: R. Mitra, P. Mller (Eds.), Nonparametric Bayesian Inference in Biostatistics, Frontiers in Probability and the Statistical Sciences, Springer International Publishing, 2015, pp. 115–134.

[34] The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 447 (7145) (2007) 661–678.

[35] J. Sethuraman, A constructive definition of dirichlet priors, Stat. Sin. 4 (1994) 639–650.

[36] M. Abramowitz, I. A. Stegun, Handbook of mathematical functions: with formulas, graphs, and mathematical tables, no. 55, Courier Corporation, 1964.

[37] S. N. MacEachern, P. Müller, Estimating mixture of dirichlet process models, Journal of Computational and Graphical Statistics 7 (2) (1998) 223–238.

[38] R. M. Neal, Markov chain sampling methods for dirichlet process mixture models, Journal of computational and graphical statistics 9 (2) (2000) 249–265.

[39] D. Binder, Bayesian cluster analysis, Biometrika 65 (1) (1978) 31–38.

[40] I. M. Johnstone, B. W. Silverman, Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences, Ann. Statist. 32 (2004) 1594–1649.

[41] J. H. Cho, The genetics and immunopathogenesis of inflammatory bowel disease, Nat. Rev. Immunol. 8 (6) (2008) 458–466.

[42] R. H. Duerr, K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, et al., A genome-wide association study identifies il23r as an inflammatory bowel disease gene, Science 314 (5804) (2006) 1461–1463.

[43] J.-P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J.-P. Cézard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, et al., Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease, Nature 411 (6837) (2001) 599–603.

[44] Y. Ogura, D. K. Bonen, N. Inohara, D. L. Nicolae, F. F. Chen, R. Ramos, H. Britton, T. Moran, R. Karaliuskas, R. H. Duerr, et al., A frameshift mutation in nod2 associated with susceptibility to crohn's disease, Nature 411 (6837) (2001) 603–606.

[45] C. Libioulle, E. Louis, S. Hansoul, C. Sandor, F. Farnir, D. Franchimont, S. Vermeire, O. Dewit, M. De Vos, A. Dixon, et al., Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13. 1 and modulates expression of ptger4, PLoS Genet. 3 (4) (2007) e58.

[46] J. Glas, J. Seiderer, M. Wetzke, A. Konrad, H.-P. Török, S. Schmechel, L. Tonenchi, C. Grassl, J. Dambacher, S. Pfennig, et al., rs1004819 is the main disease-associated il23r variant in german crohn's disease patients: combined analysis of il23r, card15, and octn1/2 variants, PloS One 2 (9) (2007) e819.

[47] L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, et al., Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease, Nature 491 (7422) (2012) 119–124.

[48] I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, X. Lin, Sequence kernel association tests for the combined effect of rare and common variants, The American Journal of Human Genetics 92 (6) (2013) 841–853.

[49] J. C. Barrett, B. Fry, J. Maller, M. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2) (2005) 263–265.

[50] M. Prager, J. Büttner, C. Büning, Ptger4 modulating variants in crohns disease, International journal of colorectal disease 29 (8) (2014) 909–915.

[51] A. Lijoi, R. H. Mena, I. Prünster, Controlling the reinforcement in Bayesian nonparametric mixture models, J. R. Stat. Soc., B 69 (2007) 715–740.

[52] R. Argiento, A. Guglielmi, A. Pievatolo, Bayesian density estimation and model selection using nonparametric hierarchical mixtures, Comput. Stat. Data Anal. 54 (4) (2010) 816 – 832.

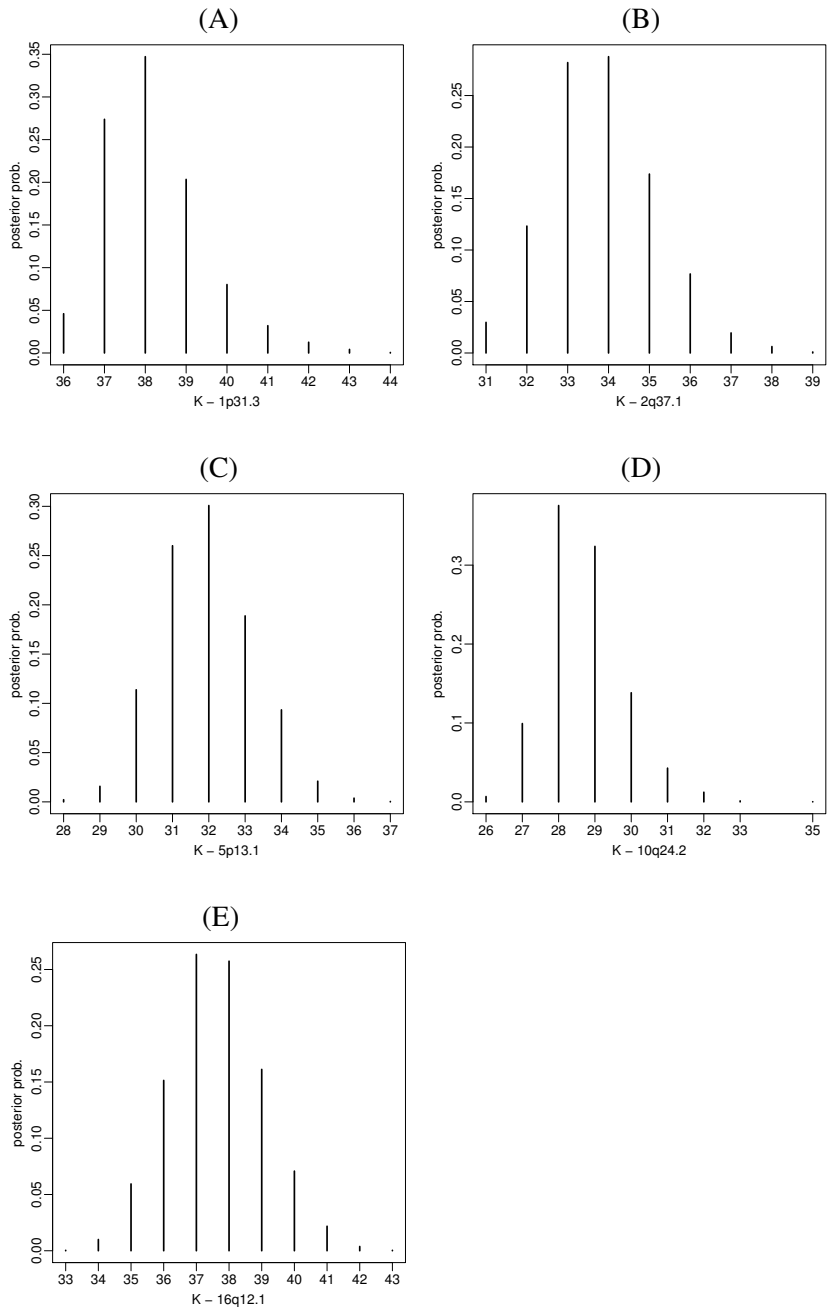

Figure 2: posterior histogram of K for each chromosome

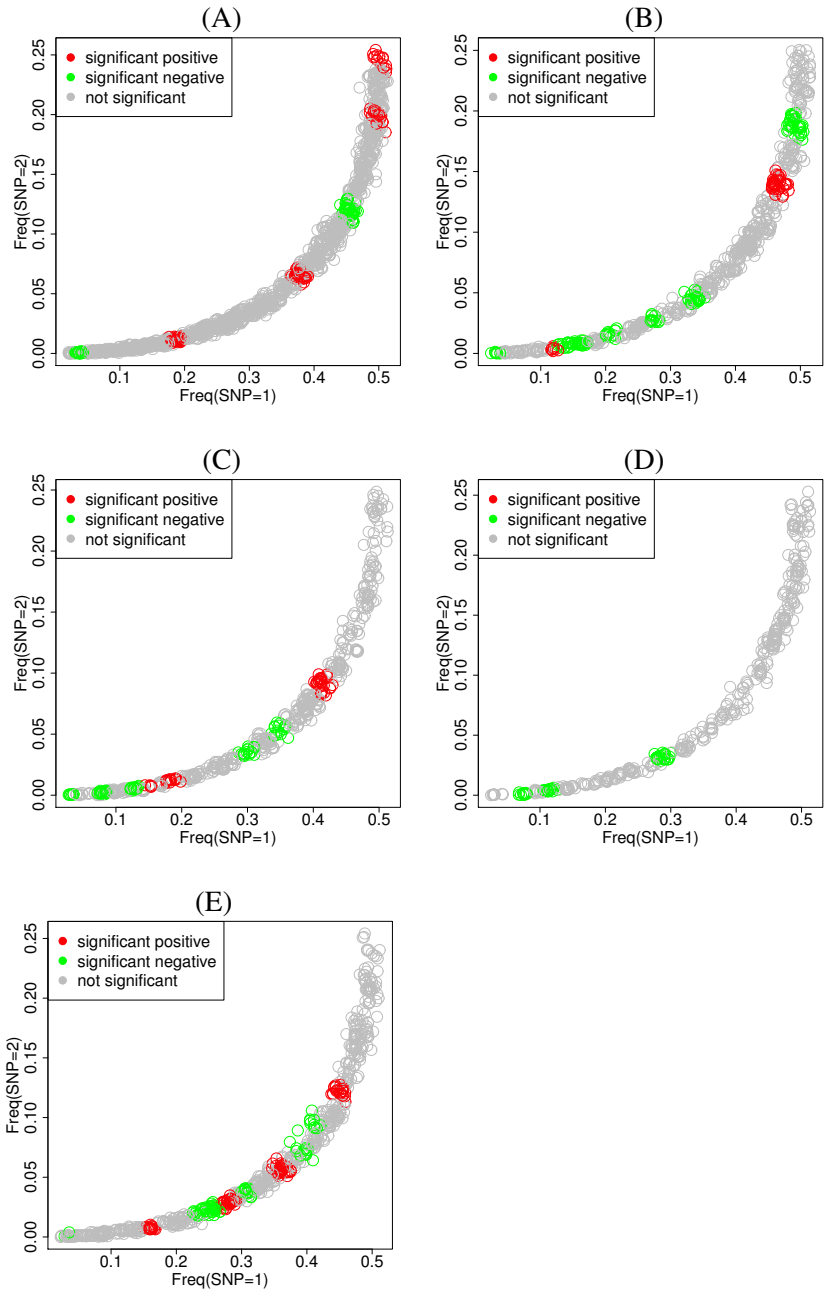

Figure 3: Observed frequencies S_1 and S_2 for each SNP in the five regions. The points are colored in red (green) if they belong to a cluster that is positively (negatively) associated to Crohn's disease, while gray points represent SNPs in cluster which were not associated to the disease.

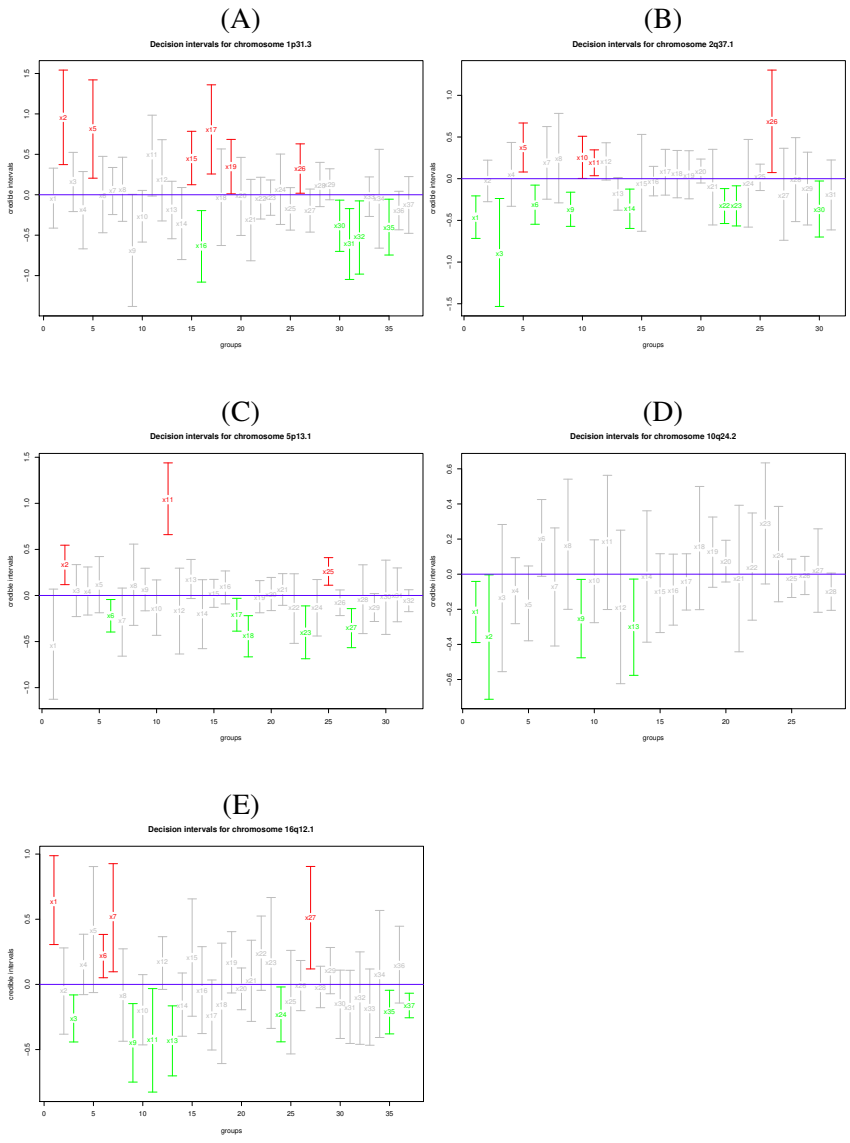

Figure 4: 90% credible intervals of β_k, for each of the five chromosomes colored according the significance: red for deleterious effect, green for protective and gray for none

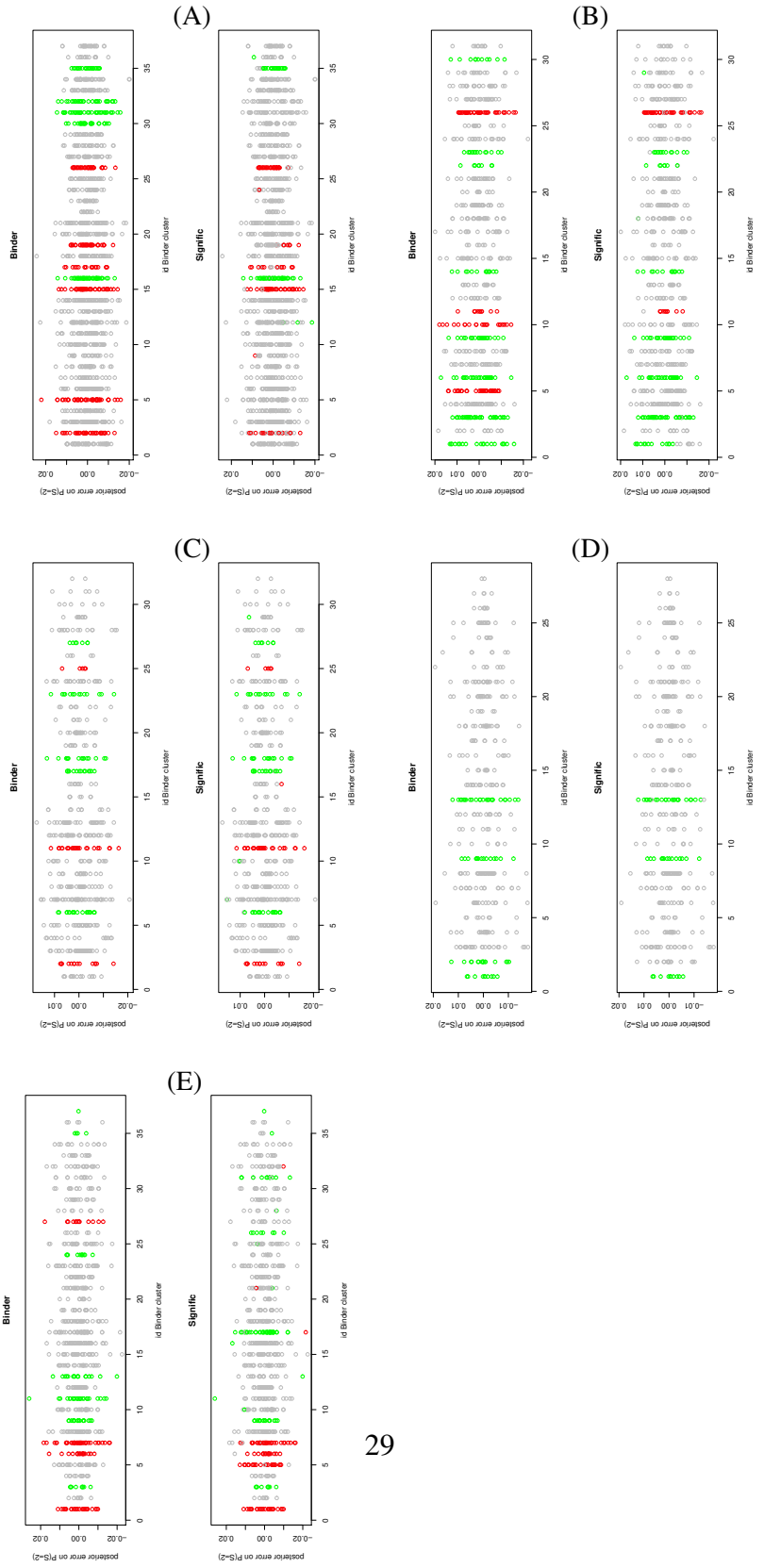

Figure 5: Comparison between the SNP effects estimated by the Binder clustering, single SNP effect estimated as in Section 4.3.

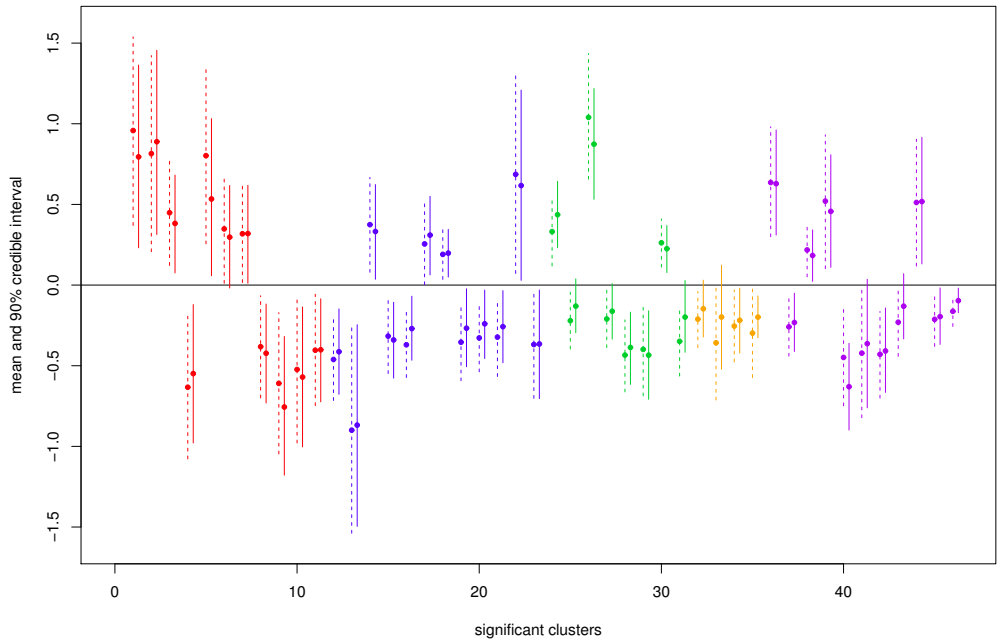

Figure 6: The comparison of 90% credible intervals for the clusters fitted separately and clusters fitted collectively. The dashed lines stand for the effects of the clusters when the clusters were fitted by the regions, and the solid lines stand for the effects of the clusters when all clusters were fitted together. Red lines stand for the significant cluster on 1p31.3; blue for the significant clusters on 2q37.1; green for the significant clusters on 5p13.1; orange for the significant clusters on 10q24.2; purple for the significant clusters on 16q12.1. X-axis represents the significant cluster ID's in the order.

Table 2: Estimates of association for SNP clusters in 1p31.3.

Cluster ID	$\hat{\beta}$	90% CI for β	$\hat{\text{OR}}$	No. of SNPs	Gene Symbol
1p31.3					
2	0.958	(0.370, 1.539)	2.606	42	*CACHD1, GADD45A, GNG12, IL12RB2, IL23R, INADL, KANK4, LEPR, NFIA, PDE4B, ROR1, SGIP1, TM2D1, USP1*
5	0.815	(0.207, 1.423)	2.259	54	*ALG6, C1orf141, EFCAB7, GADD45A, IL23R, INADL, INSL5, ITGB3BP, KANK4, LEPR, NFIA, PDE4B, PGM1, ROR1, RPE65, SGIP1, USP1, WDR78*
15	0.448	(0.122, 0.783)	1.566	48	*CACHD1, IL23R, INADL, MIER1, NFIA, RAVER2, ROR1, SGIP1, SLC35D1, TM2D1, WDR78*
16	-0.633	(-1.078, -0.193)	0.531	53	*ATG4C, DOCK7, IL12RB2, IL23R, ITGB3BP, KANK4, LEPR, NFIA, PDE4B, PGM1, RAVER2, ROR1, SGIP1, SLC35D1, WDR78*
17	0.802	(0.255, 1.358)	2.231	18	*ALG6, INADL, ITGB3BP, PDE4B, ROR1, SGIP1, UBE2U*
26	0.318	(0.016, 0.626)	1.374	39	*CACHD1, IL23R, NFIA, RAVER2, ROR1, SGIP1, UBE2U, USP1, WLS*
30	-0.382	(-0.700, -0.066)	0.682	28	*CACHD1, GADD45A, GNG12, L1TD1, NFIA, PDE4B, ROR1, UBE2U, USP1, WLS*
31	-0.609	(-1.047, -0.171)	0.544	39	*GNG12, INADL, LEPR, NFIA, PDE4B, ROR1, SGIP1, USP1*
32	-0.523	(-0.979, -0.074)	0.593	31	*C1orf141, INADL, NFIA, PDE4B, PGM1, RAVER2, ROR1, SGIP1, TM2D1, UBE2U, USP1*
35	-0.404	(-0.747, -0.057)	0.668	28	*INADL, KANK4, NFIA, PDE4B, ROR1, RPE65, USP1*

Table 3: Number of significant clusters (C) and SNPs (S) under each method

	Single-marker		BMIX		LD.SKAT		NGG		DPM		Common
	C	S	C	S	C	S	C	S	C	S	SNPs
1p31.3	-	28	-	15	6	35	9	180	11	192	17
2q37.1	-	4	-	5	3	11	10	184	12	197	6
5p13.1	-	23	-	22	6	48	10	121	8	118	26
10q24.2	-	8	-	7	3	18	4	50	4	48	6
16q12.1	-	21	-	12	3	17	13	162	11	111	10
Total	-	84	-	61	21	129	46	697	46	666	65