A Mutual Information-based method to select informative pairs of variables in case-control genetic association studies to improve the power of detecting interaction between genetic variants

Mathieu Emily, Chloé Friguet

Journal de la Société Française de Statistique
Vol. 159 No. 2 (2018)

A Mutual Information-based method to select informative pairs of variables in case-control genetic association studies to improve the power of detecting interaction between genetic variants.

Mathieu Emily[1] and Chloé Friguet[2]

Abstract: We propose a novel procedure for tagging Single Nucleotide Polymorphisms (SNPs), called *EpiTag*, to deal with interaction detection in Genome-Wide Association Studies. The aim of our method is to select a set of tag-SNPs that optimally represents the whole set of pairs of SNPs whereas usual approaches are univariate. The linkage between two pairs of SNPs is measured by the Normalized Mutual Information. The proposed algorithm is assessed considering the power of interaction detection compared to a no-tagging strategy and a usual one-dimensional tagging procedure, both on simulated and real genotype structures. *EpiTag* demonstrates good power performances along with various signal strengths or data sizes *w.r.t* the competing methods.

Résumé : Nous proposons une nouvelle méthode de sélection de marqueurs biologiques, appelée *EpiTag*, permettant la détection d'interaction de gènes dans les études d'association à l'échelle du génome. Notre méthode extrait un sous-ensemble de marqueurs qui caractérise de façon optimale la variabilité de la totalité des couples de marqueurs, là où les approches usuelles considèrent les marqueurs de façon univariée. Nous proposons de quantifier le lien entre couples de marqueurs par l'Information Mutuelle Normalisée. La faisabilité de notre méthode est validée à partir d'une étude de la puissance de détection d'interaction sur un ensemble de jeu de données avec une structure de dépendance simulée ou bien provenant de donnéées réelles. *EpiTag* réalise de bonnes performances en terme de puissance, et ce quelque soit la force du signal ou la dimension des données testées, par rapport aux autres méthodes.

Keywords: Genome-wide association studies, Gene-gene interaction, Mutual information, Selection of pairs of variables
Mots-clés : Etudes d'association à l'échelle du génome, Inéteraction entre gènes, Information mutuelle, Selection de paires de variables
AMS 2000 subject classifications: 62F03, 62F07, 62P10

1. Introduction

The sequencing of the human genome combined with the finalization of the HapMap project (International HapMap Consortium, 2003) have allowed the development of association studies at the genome scale. The aim of these Genome-Wide Association Studies (GWAS) is to detect differences in genetic variants associated to a specific trait (a disease for example). Genetic variant usually refers to Single Nucleotide Polymorphism (SNP), defined as one base pair on the genome that is polymorphic in the studied population. Single-locus approaches, whereby a large

[1] Agrocampus OUEST, IRMAR, Rennes
E-mail: mathieu.emily@agrocampus-ouest.fr
[2] Univ. Bretagne-Sud, IRISA, Vannes
E-mail: chloe.friguet@univ-ubs.fr

 ISSN: 2102-6238

number of SNPs are tested independently for association, have first been developed to analyse GWAS (Lewis, 2002). Although such single-locus approaches have successfully identified regions of disease susceptibility (Hindorff et al., 2009), findings were of modest effect and a large proportion of the genetic heritability is still not covered for common complex diseases (Maher, 2008; Manolio et al., 2009).

Single-marker strategy is usually considered as one of the main limiting factors for GWAS to detect causal variants. Since human complex diseases are generally caused by the combined effect of multiple genes, the detection of genetic interactions (also called epistasis) is thus essential to improve our knowledge of the etiology of complex diseases (Cordell, 2009; Hindorff et al., 2009). However, the detection of gene-gene interaction in GWAS remains very challenging. First, when considering SNP arrays with 1,000,000 SNPs, an exhaustive testing requires extensive computing resources to perform, store and post-process the 5×10^{11} possible interaction tests (Ritchie, 2015). From a statistical point-of-view, the detection of gene-gene interactions raises issues related to the statistical power of the proposed methods, such as the data structure and the complexity of the models of interaction. GWAS data are first characterized by their high-dimension and by the correlation between variables inherited from the complex architecture of the genome. Furthermore, the lack in power is enhanced by the number of factors known to influence the power of statistical methods in GWAS (Emily and Friguet, 2017; Emily, 2016b) and by the vast amount of epistatic models (Li and Reich, 2000; Hallgrimsdottir and Yuster, 2008). Although the past few years have therefore seen the development of methods dedicated to the detection of association between a case-control phenotype and the interaction between pairs of SNPs (Wan et al., 2010; Ueki and Cordell, 2012; Emily, 2012), and pairs of SNP-sets (Larson and Schaid, 2013; Emily, 2016a), efforts are still needed to improve the power of detection of pairwise interaction.

Another major limitation of GWAS is the indirect association testing due to the tagging step. Because of the block structure of the genome and to reduce the technological cost of genotyping, SNP-arrays are designed to genotype a (relatively) small part of the SNPs, called tag-SNPs. Tag-SNPs are selected to capture a high proportion of the genetic variation all along the genome. Therefore, the actual causal SNPs, where the mutation responsible for the disease has truly occurred, may not be genotyped in the observed dataset. In that case, the causal SNPs may be detected thanks to its tag-SNPs, thus testing for indirect association. The loss of power, induced by indirect association, has been well studied in the case of single-marker association (Weir, 2008; Emily and Friguet, 2017; Emily, 2016b) and a strong link between power loss and the amount of correlation between causal and tag-SNPs has been demonstrated in several studies (Pritchard and Przeworski, 2001; Carlson et al., 2004; Nielsen et al., 2004).

Since the selection of maximum informative tag-SNPs is an NP-complete problem, a substantial literature has been dedicated to provide computational solutions to the tagging issue (Carlson et al., 2004; de Bakker et al., 2005; Ao et al., 2005; Frommlet, 2010; Sicotte et al., 2011; Bush and Moore, 2012). Tagging algorithms currently focus on one-dimensional quality measures of tagging and therefore aim at maximizing the information retrieved for each SNP individually. For example, in the widely used method *Tagger*, each SNPs is tagged by at least one tag-SNP with a correlation coefficient r^2 higher than a cutoff, usually $r^2 \geq 0.8$ (de Bakker et al., 2005). However, when searching for SNPxSNP interaction, there is no guarantee that maximizing the information for each SNP of the causal pair using a one-dimensional strategy is optimal to re-

 ISSN: 2102-6238

cover the maximum of information of the pair. In other words, a pair of markers is not necessarily well represented by the pair of the one-dimensional associated tag-SNPs. Thus, it is very likely that only accounting for a one-dimensional pattern of correlation between SNPs may generate an uncontrolled loss in power. Although variable and feature selection is of main interest in various domains, the selection of *pairs* of variables is still not much studied in the statistical literature. To our knowledge, the only study dealing with pair selection can be found in Ng (2004) where the author proposes a bivariate variable selection method but for supervised classification problems.

In this paper, we propose a new approach for tagging SNPs, called *EpiTag*, to deal with interaction detection. The aim of our method is to select a set of tag-SNPs that optimally represents the total set of all pairs of SNPs. The linkage between two pairs of SNPs is measured by the Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002). NMI aims at quantifying the amount of information (from an entropy point-of-view) obtained about one random variable, through the other random variable. Compared to the correlation coefficient, NMI is not limited to real-valued random variables and NMI can therefore be applied to pairs of variables which is of major interest in our context of interaction detection. Based on the NMI, *EpiTag* can be seen as a greedy algorithm that allows each pair of the genome to be tagged by a pair of SNPs, at fixed level of NMI.

After introducing the notations, Section 2 describes the most commonly used one-dimensional tagging strategy called *Tagger*. Our proposed method *EpiTag*, based on two-dimensional information, is detailed in Section 3. In Section 4, we evaluate the performance of our method compared to a non-tagging strategy, where no selection is performed before testing for association and a usual one-dimensional strategy. Perspectives are discussed in Section 5.

2. Notations and usual tagging strategy

2.1. Notations for genotype data

Let us consider that the genotype of an individual is measured through a collection of p SNPs. In more details, for $i = 1, \ldots, p$, let X_i be a random variable modeling the genotype of the i^{th} SNP. Although various modeling of the X_i's can be considered, we focus here on the raw representation of a SNP where X_i is a categorical variable with three levels denoted by $X_i \in \{AA, Aa, aa\} = \{0, 1, 2\}$. States AA and aa correspond to the two homozygote genotypes while Aa is the heterozygote state, where A (resp. a) is the major (resp. minor) allele of SNP i. In the following, we focus on sets of SNPs, $\mathbb{X}^1 = [X_1^1, \ldots, X_{p_1}^1]$ and $\mathbb{X}^2 = [X_1^2, \ldots, X_{p_2}^2]$, respectively located within two genomic regions $\mathscr{R}_1$ and $\mathscr{R}_2$ composed of p_1 and p_2 SNPs.

Let us further consider a sample of n individuals. The observed genotypes from $\mathscr{R}_1$ can be represented by a $n \times p_1$ matrix $\mathbf{X^1} = \left[x_{\ell,i}^1\right]_{\ell \in 1 \ldots n;\ i \in 1 \ldots p_1}$, where $x_{\ell,i}^1$ is the observed genotype for SNP i carried by individual ℓ. Therefore, $x_{\ell,i}^1$ is the realization of a random variable characterized a multinomial distribution as introduced in the previous paragraph. Using similar notations for genotypes observed from $\mathscr{R}_2$, a typical genotype dataset can be summarized as follows:

 ISSN: 2102-6238

$$\mathbf{X^1} = \begin{bmatrix} x^1_{1,1} & \dots & x^1_{1,p_1} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ x^1_{n,1} & \dots & x^1_{n,p_1} \end{bmatrix} \quad \mathbf{X^2} = \begin{bmatrix} x^2_{1,1} & \dots & x^2_{1,p_2} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ x^2_{n,1} & \dots & x^2_{n,p_2} \end{bmatrix}$$

2.2. *One-dimensional tagging: Tagger*

The most widely used method for tag-SNP selection is called *Tagger* (de Bakker et al., 2005). *Tagger* algorithm aims at selecting a maximally informative set of SNPs based on Linkage Disequilibirum (LD) (Carlson et al., 2004). LD refers to the non-random association of alleles at different SNPs and therefore introduces dependency between SNPs. The amount of LD in a set of SNPs is commonly measured by the pairwise correlation r^2 between SNPs (Hill and Robertson, 1968). When considering two biallelic SNPs, X_i and X_j, with respective marginal allele frequencies π_A and π_B, and π_{AB} as joint allele frequency, the r^2 is defined by:

$$r^2(X_i, X_j) \quad = \quad \frac{\left(\pi_{AB} - \pi_A \times \pi_B\right)^2}{\pi_A \times \pi_a \times \pi_B \times \pi_b}$$

Given the set of SNPs $\mathbf{X^1}$ from the genomic region $\mathscr{R}_1$, the *Tagger* procedure starts by computing the pairwise correlation matrix of size $p_1 \times p_1$. As displayed in the flowchart of Figure 4 (see Appendix B), SNPs are then iteratively partitioned into different blocks (also called bins) according to the relevance of SNPs ($r^2 \geq \gamma$, with $\gamma = 0.8$ usually in practice). Each iteration is decomposed into two steps: (1) the identification of the current tag-SNP and (2) the building of the current bin. The identification of the current tag-SNP is performed by selecting the SNP exceeding the r^2 threshold with the maximum number of other SNPs in the current set. Then, the current bin is built by aggregating the current tag-SNP and its associated SNPs if any. In the first iteration, the current set of SNPs is the whole set of SNPs and then, at each iteration, the current set of SNPs is updated by removing SNPs from the current bin until all SNP belongs to a single bin.

After the final iteration performed on $\mathscr{R}^1$, the *Tagger* procedure provides two distinct outputs: a set of tag-SNPs denoted I^1_{Tagger} and a set of bins. I^1_{Tagger} refers to the set of tag-SNPs indices so that $I^1_{Tagger} \subseteq [1, \dots, p_1]$. Let us also introduce t_1 a mapping function defined by:

$$\begin{array}{rccc} t_1 : & [1, \dots, p_1] & \longrightarrow & I^1_{Tagger} \\ & i & \longmapsto & t_1(i) \end{array} \tag{1}$$

where $t_1(i)$ gives the index of the tag-SNP for SNP i. Besides, it is noteworthy that :

$$\forall i \in I^1_{Tagger} : t_1(i) = i.$$

Tagger relies on the r^2 measure that can be seen as a quality measure that specifies how well a set of tag-SNPs reports the information carried by the whole set of SNPs. Common quality measure introduced in the literature are one-dimensional in the sense that they focus on the

 ISSN: 2102-6238

recovered information for single SNP only, while information from pairs, triplet or any n-tuple with $n > 1$ is not considered. Therefore, in the design setting introduced in Section 2.1, the search for tag-SNPs in regions $\mathscr{R}_1$ and $\mathscr{R}_2$ is performed independently. As consequence, when applying *Tagger* to region $\mathscr{R}_2$, similar notations can be used to define the set of of tag-SNPs, namely I^2_{Tagger}, and the map function $t_2 : [1,\dots,p_2] \to I^2_{Tagger}$ that provides the index of the tag-SNP for any SNP $i \in [1,\dots,p_2]$.

3. EpiTag: our two-dimensional tagging procedure

This section is dedicated to the detailed presentation of our tagging method, called *EpiTag*. To overcome the main limitation of one-dimensional methods, such as *Tagger*, that consider $\mathscr{R}_1$ and $\mathscr{R}_2$ independently, we proposed a quality measure based on the information carried by pairs of SNPs.

3.1. The Normalized Mutual Information: a quality measure for pairs of SNPs

The Normalized Mutual Information (NMI) is a measure of the mutual dependence between two variables (Strehl and Ghosh, 2002). NMI aims at quantifying the amount of information (from an entropy point-of-view) obtained about one random variable, through the other random variable. Compared to the correlation coefficient, NMI is not limited to real-valued random variables. NMI can therefore be applied to pairs of variables which is of major interest in our context of interaction detection as defined hereafter.

Let us first introduce $Z_{ij} = (X^1_i, X^2_j)$ as the random variable characterizing the couple of the two random variables X^1_i and X^2_j, where X^1_i (resp. X^2_j) is the i^{th} SNP of $\mathscr{R}_1$ (resp. j^{th} SNP of $\mathscr{R}_2$). If we consider two pairs of SNPs Z_{ij} and Z_{rs}, the NMI between Z_{ij} and Z_{rs} is defined by:

$$NMI\Big(Z_{ij}, Z_{rs}\Big) \quad = \quad \frac{I\big[Z_{ij}, Z_{rs}\big]}{\sqrt{H(Z_{ij})H(Z_{rs})}}$$

with

$$\begin{aligned} I\Big(Z_{ij}, Z_{rs}\Big) &= \sum_{(x_i,x_j,x_r,x_s)\in\{0,1,2\}^4} p_{(i,j,r,s)} \log\left(\frac{p_{(i,j,r,s)}}{\mathbb{P}\big(Z_{ij}=z_{ij}\big)\mathbb{P}\big(Z_{rs}=z_{rs}\big)}\right) \\ H\left(Z_{ij}\right) &= I\Big(Z_{ij}, Z_{ij}\Big) \\ p_{(i,j,r,s)} &= \mathbb{P}\Big((X^1_i, X^2_j, X^1_r, X^2_s) = (x_i, x_j, x_r, x_s)\Big) \end{aligned}$$

I and H denote respectively the mutual information and the entropy (Kullback, 1959).

In Appendix A, the benefit of using a two-dimensional measure of information such as the NMI compared to combining two one-dimensional measures of correlation is investigated. If we consider two pairs, Z_{ij} and Z_{rs} such that $r^2(X^1_i, X^1_r) = 0.8$, $r^2(X^2_j, X^2_s) = 0.8$ and all other pairwise $r^2 = 0$, it can be remarked that $NMI(Z_{ij}, Z_{rs})$ falls in the range $[0.7, 0.85]$ (see Figure 3). Thus, only accounting for the one-dimensional correlation when tagging pairs is likely to miss

 ISSN: 2102-6238

a certain amount of variability since the whole degrees-of-freedom are not considered. Such a result strengthens our working hypothesis and motivates the proposal of two-dimensional tagging strategies.

3.2. *EpiTag procedure*

Given the set of pairs of SNPs from the genomic regions $\mathscr{R}_1 \times \mathscr{R}_2$, the *EpiTag* procedure starts by computing the pairwise NMI matrix of size $p_1p_2 \times p_1p_2$. The main purpose of *EpiTag* is to select a subset of pairs (*i.e.* a subset of SNPs) that tags all pairs with a predefined amount of NMI denoted τ. Therefore, the goal of the *EpiTag* procedure is to provide a subset $I_{EpiTag} \subseteq [1,\dots,p_1] \times [1,\dots,p_2]$ such as:

$$\forall (r,s) \in [1,\dots p_1] \times [1,\dots,p_2], \quad \exists (i,j) \in I_{EpiTag} \ : \ NMI\Big(Z_{ij}, Z_{rs}\Big) \geq \tau \tag{2}$$

To solve the optimization problem raised in Equation (2), we propose the following iterative algorithm. Let us introduce U_k and T_k the sets of indices of respectively Untagged and Tagged pairs of SNPs after iteration k. The *EpiTag* algorithm is straightforwardly initialized by $U_0 = [1,\dots,p_1] \times [1,\dots,p_2]$, $T_0 = \emptyset$ and $I_{EpiTag} = \emptyset$. Let us now consider the k^{th} iteration of the *EpiTag* algorithm and define Q as, $\forall (i,j) \in U_{k-1}$:

$$Q_k(i,j) \quad = \quad \sum_{(r,s)\in U_{k-1}} \mathbb{1}\Big\{NMI\Big(Z_{ij}, Z_{rs}\Big) \geq \tau\Big\} \tag{3}$$

The indices of the tag-pair of SNPs $\Gamma_k = (i_k, j_k)$ is defined as:

$$\Gamma_k = \underset{(i,j)\in U_{k-1}}{\operatorname{argmax}} \ Q_k(i,j)$$

We denote by $T_k^* = \Big\{(r,s) \in U_{k-1} : NMI\Big(Z_{\Gamma_k}, Z_{rs}\Big) \geq \tau\Big\}$ the set of pairs that are tagged at iteration k. Then, at the end of iteration k, the following updates are performed:

$$U_k = U_{k-1} \setminus T_k^* \ ; \ T_k = T_{k-1} \cup T_k^* \ ; \ I_{EpiTag} = I_{EpiTag} \cup \Gamma_k$$

Tag-SNPs selection stops when $U_k = \emptyset$. Algorithm 1 describes the main steps of the *EpiTag* procedure.

Besides, as shown in Figure 5 in Appendix B, it can be remarked that the *EpiTag* algorithm has similarities with the *Tagger* algorithm and can therefore be seen as an extension of this algorithm considering (1) pairs of SNPs as input variables instead of single SNPs and (2) the NMI between pairs of SNPs as a measure of similarity between input variables instead of the correlation coefficient. We also introduce a mapping function denoted t_{EpiTag} that gives the index of the tag-SNPs pair for each pair of SNPs.

$$\begin{array}{rcccl} t_{EpiTag} & : & [1,\dots,p_1] \times [1,\dots,p_2] & \longrightarrow & I_{EpiTag} \\ & & (i,j) & \longrightarrow & t_{EpiTag}(i,j) \end{array} \tag{4}$$

where $t_{EpiTag}(i,j)$ gives the index of the tag-pair for the pair of SNPs Z_{ij}.

It is noteworthy that :

$$\forall i \in I_{EpiTag} : t_{EpiTag}(i,j) = (i,j).$$

 ISSN: 2102-6238

Algorithm 1 *EpiTag* algorithm

Require: $\mathbf{M_{NMI}}$ {the matrix of *NMI* values between all pairs of SNPs from regions $\mathscr{R}_1$ and $\mathscr{R}_2$}
Require: τ {the threshold for blocks partitionning}
$p \leftarrow dim(\mathbf{M_{NMI}})[1]$ {total number of pairs of SNPs between region $\mathscr{R}_1$ and region $\mathscr{R}_2$ = nb of rows in M_{NMI}. Note that $p = p_1 \times p_2$}
$nb.pairs.tagged \leftarrow 0$ {counter for the number of tagged pairs of SNPs}
$I_{EpiTag} \leftarrow vect(0,p)$ {0/1 p-vector indicating wether a pair of SNPs is a tag or not}
$U \leftarrow vect(1,p)$ {0/1 p-vector indicating wether a pair of SNPs has been tagged or not}
$B \leftarrow vect(0,p)$ {p-vector indicating for each pair of SNPs the index of its tag-pair}
$k = 0$ {Iteration counter}
while $nb.pair.tagged < p$ **do**
  for $t \in U == 1$ **do**
    Compute $Q(t)$ as in (3) {For each pair of SNPs, count the nb of pairs with which the mutual information is greater than τ. Note that $t = 1 \Leftrightarrow (i,j) = (1,1), t = 2 \Leftrightarrow (i,j) = (1,2), \ldots, t = p \Leftrightarrow (i,j) = (p_1, p_2)$.}
  end for
  $tag.pair \leftarrow which.max(Q)$ {The pair of SNPs exceeding the MNI threshold τ with the maximum number of other pairs of SNPs is identified as a tagging pair. If several pairs of SNPs are candidate, choose the one with the highest MNI average.}
  $tag.pair.bin \leftarrow which(\mathbf{M_{NMI}}[tag.pair, U == 1] > \tau)$ {The pair of SNPs indexed by $tag.pair$ tags for all pairs of SNPs with $MNI > \tau$ among those that have not been tagged yet}
  $I_{EpiTag}[tag.pair] \leftarrow 1$ {Update: the pair of SNPs indexed by $tag.pair$ is a tag}
  $B[tag.pair.bin] \leftarrow tag.pair$ {Update: the pair of SNPs indexed by $tag.pair$ is the tag-pair for those indexed by $tag.pair.bin$}
  $B[tag.pair] \leftarrow tag.pair$ {Update: the pair of SNPs indexed by $tag.pair$ is its own tag-pair}
  $U[tag.pair.bin] \leftarrow 0$ {Update: pairs of SNPs indexed by $tag.pair.bin$ are no longer candidates for tagging}
  $nb.pairs.tagged \leftarrow p - sum(U)$ {Update: total nb of pairs of SNPs that have been tagged}
end while
return I_{EpiTag}, B

4. Numerical evaluation of EpiTag

To evaluate the performances of the *EpiTag* procedure, power studies are performed, based on simulated disease models. The aim is to compare the power of detection of *EpiTag* with two commonly used strategies: a one-dimensional tagging strategy, namely *Tagger*, and a "no-tagging" strategy that we call *NoTag* where no selection is performed. Although there exists a direct relationship between coverage and power in single association testing (Pritchard and Przeworski, 2001), such a relationship is not valid anymore when dealing with SNPxSNP interaction. Therefore, compared to previous studies where emphasis is put on the coverage of the genome (Carlson et al., 2004), we focus here on the statistical power of detection as a measure of comparison.

In the remainder of this section, we first provide details regarding the pipeline of simulation proposed to compare *EpiTag* with the two other tagging procedures, *Tagger* and *NoTag*. Indeed, existing simulators for GWAS does not account for correlation between two regions and the setting is not straightforward. The functions used to run *EpiTag* are available in the following github repository: `https://github.com/MathieuEmily/EpiTag`. Then, results obtained with either a simulated genotype structure or an observed genotype structure are presented.

 ISSN: 2102-6238

4.1. Simulation pipeline

Our simulation pipeline can be decomposed into four main successive steps: (1) the design of genotype structures, (2) the simulation of the phenotype, (3) the testing for SNPxSNP associations and (4) the estimation of power. These steps are detailed hereafter.

Step ♯1: Design of genotype structures The design of genotype data is performed by simulating the two regions $\mathscr{R}_1$ and $\mathscr{R}_2$ simultaneously. Our purpose is to simulate genotype structures with various patterns of r^2 and pairwise *NMI* defined according to the setting of a probabilistic setup. This probabilistic setup is decomposed into three main steps: (1) an initialization step where two triplets of variables (one triplet in region $\mathscr{R}_1$ and one triplet in region $\mathscr{R}_2$) are simulated simultaneously, (2) the addition of a single variable in either $\mathscr{R}_1$ or $\mathscr{R}_2$, (3) the addition of a pair of variables with one variable in $\mathscr{R}_1$ and one variable in $\mathscr{R}_2$. The following paragraphs provide details regarding the three simulation steps.

Initialization. To initialize the simulation process, we start by defining a joint probability distribution denoted $\mathbb{P}$ between the two triplets (X_1^1, X_2^1, X_3^1) and (X_1^2, X_2^2, X_3^2) as defined in Equation (5). (X_1^1, X_2^1, X_3^1) and (X_1^2, X_2^2, X_3^2) correspond to the three first SNPs of $\mathscr{R}_1$ and $\mathscr{R}_2$ respectively.

$$\forall (i,j,k,r,s,t) \in [0,1,2]^6 : \mathbb{P}[X_1^1 = i, X_2^1 = j, X_3^1 = k, X_1^2 = r, X_2^2 = s, X_3^2 = t] = p_{ijkrst} \quad (5)$$

The definition of $\mathbb{P}$ allows the control both of the r^2 between all pairs of variables and of $NMI\left(Z_{ij}, Z_{rs}\right)$ for $(i,j,r,s) \in [0,1,2]^4$.

Single-SNP adding. A single variable can be added to either region $\mathscr{R}_1$ or $\mathscr{R}_2$ by specifying the r^2 between the adding variable and a variable already included the simulation setup. For example, variable X_1^4 is added by setting $r^2(X_1^1, X_1^4) = 0.8$.

SNP-pair adding. A single pair of variable can be added to the simulation setup by specifying the *NMI* between the added pair and an existing pair. For example, SNP pair $Z_{4,4}$ is added by defining $NMI\left(Z_{44}, Z_{11}\right) = 0.7$.

Details regarding the two above mentioned operations (*Single-SNP adding* and *SNP-pair adding*) are provided in Appendix D.

To simulate two observed genotype structures, $\mathbf{X}^1$ and $\mathbf{X}^2$, we set the number n of individuals to be simulated. Then given a simulation setup defined by an initialization step and a series of single-SNP and SNP-pair adding: we first use Equation (5) to simulate n observations of the 6 variables $(X_1^1, X_2^1, X_3^1, X_1^2, X_2^2, X_3^2)$. Then single-SNPs and/or SNP-pairs are added conditionally to existing variables. The output of the simulation procedure is two matrices, $\mathbf{X^1}$ and $\mathbf{X^2}$, with n observations (or rows) and respectively p_1 and p_2 variables (or columns), as described in Section 2.1.

 ISSN: 2102-6238

Step ♯2: Phenotype simulation The main purpose of the phenotype simulation is to draw a response variable with respect to a given disease model. The response variable, denoted by Y and corresponding to the disease status, is modeled as a random binary variable ($Y \in \{0,1\}$) where $Y = 0$ stands for an healthy individual (control status) and $Y = 1$ a diseased individual (case status).

Given two SNP-sets, $\mathbf{X}^1$ and $\mathbf{X}^2$, obtained as outputs of **Step ♯1**, the simulation of Y is performed as follows. At first, a pair of causal SNPs, denoted by $Z_c = (X^1_{c_1}, X^2_{c_2})$ where $c_1 \in [1, p_1]$ and $c_2 \in [1, p_2]$, is chosen at random. Then, Y is defined according the following logistic regression model:

$$\text{logit}\Big(\mathbb{P}[Y = 1 | Z_c = (x_{c_1}, x_{c_2})]\Big) \quad = \quad \beta_0 + \beta_1 \mathbb{1}\{(x_{c_1} \geq 1) \cap (x_{c_2} \geq 1)\} \tag{6}$$

Equation (6) refers to a dominant-dominant disease model that has been studied in a large number of studies (Marchini et al., 2005; Emily, 2012). Such a model allows the simulation of the disease status for each observed individual summarized in the following $n-uplet$:

$$\mathbf{Y} = [y_i, \ldots, y_n]'$$

Step ♯3: Testing for SNPxSNP associations The association between Y and the two SNP-sets $\mathbf{X^1}$ and $\mathbf{X^2}$ is tested by performing all pairwise interaction tests between SNP-pairs $Z_{ij} = (X^1_i, X^2_j)$ ($1 \leq i \leq p_1$ and $1 \leq j \leq p_2$) and Y. More precisely, association testing between Y and a single pair Z_{ij} is performed by a Likelihood Ratio Test that aims at comparing the two logistic models $\mathscr{M}_{\text{Inter}}$ and $\mathscr{M}_{\text{NoInter}}$, defined as follows (the reference level is 0):

$$\mathscr{M}_{\text{NoInter}} \ : \ \text{logit}\Big(\mathbb{P}[Y = 1 | Z_{ij} = (x_1, x_2)]\Big) = \beta^N + \sum_{i=1}^{2} \gamma_i^N \mathbb{1}\{x_1 = i\} + \sum_{j=1}^{2} \delta_j^N \mathbb{1}\{x_2 = j\}$$

$$\mathscr{M}_{\text{Inter}} \ : \ \text{logit}\Big(\mathbb{P}[Y = 1 | Z_{ij} = (x_1, x_2)]\Big) = \beta^I + \sum_{i=1}^{2} \gamma_i^I \mathbb{1}\{x_1 = i\} + \sum_{j=1}^{2} \delta_j^I \mathbb{1}\{x_2 = j\} + \sum_{\substack{1 \leq i \leq 2 \\ 1 \leq j \leq 2}} \gamma\delta_{ij}^I \mathbb{1}\{x_1 = i\}\mathbb{1}\{x_2 = j\}$$

$$LRT = \mathscr{D}(\mathscr{M}_{\text{NoInter}}) - \mathscr{D}(\mathscr{M}_{\text{Inter}}) \overset{H_0}{\sim} \chi^2(4)$$

where $\mathscr{D}$ is the deviance. Such a Likelihood Ratio Test allows for testing the significance of the interaction between X^1_i and X^2_j in association with Y. As usual, the significance of the test is summarized by a p-value denoted by $pval_{(i,j)}(Y)$. Therefore, the output of **Step ♯3** is a set of $p_1 \times p_2$ p-values, stored in the **Pval** uplet as follows:

$$\mathbf{Pval}(Y) = \Big[pval_{(1,1)}(Y), \ldots, pval_{(i,j)}(Y), \ldots, pval_{(p_1,p_2)}(Y)\Big]$$

Step ♯4: Power estimation Based on the observation of two SNP-sets, $\mathbf{X}^1$ and $\mathbf{X}^2$, obtained as outputs of Step ♯1, the association test for tag-pairs can be performed for each of the three compared strategies: *NoTag*, *Tagger* and *EpiTag*.

 ISSN: 2102-6238

NoTag. Since there is no SNP selection, all possible pairwise tests are performed. We introduce I_{NoTag} as the set of indices for tested pairs:

$$I_{NoTag} = [1\ldots,p_1]\times[1\ldots,p_2].$$

Given a response variable Y, obtained as an output of **Step ♯2**, the set of computed p-values is then given by:

$$\mathbf{Pval}_{NoTag}(Y) = \Big\{ pval_{(i,j)}(Y),\ (i,j)\in I_{NoTag} \Big\}.$$

To account for multiple comparisons, we aim at controlling the False Discovery Rate at a given level α by applying the Benjamini-Hochberg (BH) correction to $\mathbf{Pval}_{NoTag}(Y)$ (Benjamini and Hochberg, 1995). Let us denote by $\widetilde{\mathbf{Pval}}_{NoTag}(Y)$ the obtained vector of BH-corrected p-values where:

$$\widetilde{\mathbf{Pval}}_{NoTag}(Y) = \Big\{ \widetilde{pval}^{NoTag}_{(i,j)}(Y),\ (i,j)\in I_{NoTag} \Big\}.$$

The ability for the *NoTag* strategy to correctly detect interaction, with respect to Y, is therefore measured by:

$$pval^{*}_{NoTag}(Y) = \widetilde{pval}^{NoTag}_{(c_1,c_2)}(Y),$$

where (c_1,c_2) is the causal pair used to simulate Y. It is noteworthy that $(c_1,c_2)\in I_{NoTag}$, meaning that the causal pair is always tested in the *NoTag* strategy. To estimate the power of detection, we simulate *nb.sim* response variables by repeating **Step ♯2** *nb.sim* times. We therefore obtain a collection of *nb.sim* vectors of phenotypes $(Y_1,\ldots,Y_{nb.sim})$, and the power is estimated with:

$$\widehat{Power}_{NoTag}(\alpha) = \frac{1}{nb.sim}\sum_{k=1}^{nb.sim} \mathbb{1}\Big\{ pval^{*}_{NoTag}(Y_k)\leq\alpha \Big\}.$$

Tagger. As described in Section 2.2, the one-dimensional tagging strategy *Tagger* results in the independent selection of two sets of tag-SNPs for region $\mathscr{R}_1$ and $\mathscr{R}_2$. The set of tested pairs indices can be defined as:

$$I_{Tagger} = I^1_{Tagger}\times I^2_{Tagger},$$

where I^1_{Tagger} (resp. I^2_{Tagger}) denotes the set of tag-SNPs indices within $\mathscr{R}_1$ (resp. $\mathscr{R}_2$).

Similarly to the *NoTag* strategy, the set of p-values for a given Y, $\mathbf{Pval}_{Tagger}(Y)$, and corresponding BH-corrected p-values, $\widetilde{\mathbf{Pval}}_{Tagger}(Y)$ obtained with the *Tagger* strategy are:

$$\mathbf{Pval}_{Tagger}(Y) = \Big\{ pval_{(i,j)}(Y),\ (i,j)\in I_{Tagger} \Big\},\ \text{and}\ \widetilde{\mathbf{Pval}}_{Tagger}(Y) = \Big\{ \widetilde{pval}^{Tagger}_{(i,j)}(Y),\ (i,j)\in I_{Tagger} \Big\}.$$

It is noteworthy that, since $I_{Tagger}\subseteq I_{NoTag}$, the Benjamini-Hochberg correction is likely to be different between both strategies as the correction depends on the number of tests. Furthermore, it is likely that $(c_1,c_2)\notin I_{Tagger}$, so that the actual causal pair (c_1,c_2) may not be directly tested with the *Tagger* strategy. In that case, the causal pair (c_1,c_2) is tested indirectly by the pair $\Big(t_1(c_1),t_2(c_2)\Big)$, where $t_i(c_i)$ is the tag for c_i in I^i_{Tagger}(see Equation (1)).

Therefore, the ability for *Tagger* to detect the causal pair is controlled by $p^{*}_{Tagger}(Y)$ defined by:

 ISSN: 2102-6238

$$pval^{*}_{Tagger}(Y) = \widetilde{pval}^{Tagger}_{(t_1(c_1),t_2(c_2))}(Y).$$

Similarly to the *NoTag* stategy, the estimation of power is performed by averaging over *nb.sim* simulated response variables obtained with the procedure described in **Step ♯2**:

$$\widehat{Power}_{Tagger}(\alpha) = \frac{1}{nb.sim}\sum_{k=1}^{nb.sim} \mathbb{1}\Big\{pval^{*}_{Tagger}(Y_k) \leq \alpha\Big\}.$$

EpiTag. The main output of the *EpiTag* algorithm is the set of tag-pairs indices denoted by I_{EpiTag} (See Section 3). Therefore, for a given Y, the set of p-values $\mathbf{Pval}_{EpiTag}(Y)$ and BH-corrected p-values $\widetilde{\mathbf{Pval}}_{EpiTag}(Y)$ obtained with the *EpiTag* strategy can be defined as follows:

$$\mathbf{Pval}_{EpiTag}(Y) = \Big\{pval_{(i,j)}(Y),\ (i,j) \in I_{EpiTag}\Big\},\ \text{and}\ \widetilde{\mathbf{Pval}}_{EpiTag}(Y) = \Big\{\widetilde{pval}^{EpiTag}_{(i,j)}(Y),\ (i,j) \in I_{EpiTag}\Big\}.$$

As for the *Tagger* strategy, it is likely that $(c_1,c_2) \notin I_{EpiTag}$ and the true signal may be tested indirectly by $t_{EpiTag}(c_1,c_2)$, the tag-pair for (c_1,c_2) obtained with *EpiTag* algorithm (see Equation (4)). Therefore, the ability for *EpiTag* to detect the causal pair is controlled by:

$$pval^{*}_{EpiTag}(Y) = \widetilde{pval}^{EpiTag}_{(t_{EpiTag}(c_1,c_2))}(Y).$$

Here again, power estomation is performed by averaging over *nb.sim* simulated response variables obtained with the procedure described in **Step ♯2**:

$$\widehat{Power}_{EpiTag}(\alpha) = \frac{1}{nb.sim}\sum_{k=1}^{nb.sim} \mathbb{1}\Big\{pval^{*}_{EpiTag}(Y_k) \leq \alpha\Big\}.$$

4.2. Results

To evaluate the ability for the *EpiTag* procedure to identify an association between a binary phenotype Y and a causal pair of SNPs (X_{c_1},X_{c_2}), we conduct power studies based on either simulated structures of genotypes and an observed structure of genotype. The power is strongly related to several parameters of the data design such as the sample size, the odds-ratio of the phenotype (odds of a case with respect to a control) and the frequency of the desease in the population. In the present simulation study, the sample sizes among cases and controls are fixed and equal and we consider various strength of the signal by choosing several β_1 values in model (6). Significance is determined at the $\alpha = 5\%$ experiment-wide level. Power evaluation for *Tagger* is performed by setting the r^2 threshold to $\gamma = 0.8$ as usual. We further fix the *NMI* threshold for *EpiTag* to $\tau = 0.75$ since it corresponds to the average NMI value obtained when simulating pairs with $r^2 = 0.8$, as shown in Appendix A.

 ISSN: 2102-6238

Simulated genotype structures

Settings. The first series of power studies is based on simulated genotype structures using the procedure described in the simulation pipeline in Section 4.1. Three simulated genotype structures are tested, considering different number of variables per region:

- Scenario ♯1: $p_1 = p_2 = 3$
- Scenario ♯2: $p_1 = p_2 = 6$
- Scenario ♯3: $p_1 = p_2 = 10$

In each scenario, the major allele frequency for each variable is set to 0.6 ($\pi_i^1 = 0.6 \ \forall i = 1,\ldots,p_1$ and $\pi_j^2 = 0.6 \ \forall j = 1,\ldots,p_2$). Furthermore, the between correlation structure is defined so that variables from $\mathscr{R}_2$ are not correlated with variables from $\mathscr{R}_1$:

$$\forall (i,j) \in [1,p_1] \times [1,p_2] \quad : \quad r^2(X_i^1, X_j^2) = 0$$

According to **Step ♯1**, the six first variables defined in the initializing step are set according to the following parameters:

- the within correlation structure in region $\mathscr{R}_1$ and $\mathscr{R}_2$ is defined so that two variables within the same region are correlated with $r^2 = 0.8$:

$$\forall i \in [1,2,3], j \in [1,2,3] \text{ and } i \neq j \quad : \quad r^2(X_i^1, X_j^1) = 0.8 \quad \text{and} \quad r^2(X_i^2, X_j^2) = 0.8,$$

- the information between pairs of variables is chosen so that the NMI between Z_{11} and Z_{22} is minimal and the NMI between Z_{11} and Z_{33} is maximal:

$$NMI(Z_{11}, Z_{22}) = 0.7 \quad \text{and} \quad NMI(Z_{11}, Z_{33}) = 0.85.$$

Then, pairs of variables are added with respect to the following parameters:

$$\forall i \in [4,\ldots,p_1], \ r^2(X_i^1, X_1^1) = 0.7 \quad \text{and} \quad \forall j \in [4,\ldots,p_2], \ r^2(X_j^2, X_1^2) = 0.7$$

In order to mimic a Genome-Wide Association Study, a large population of $N = 100,000$ individuals is simulated. Then, the set of tested pairs I_{Tagger} and I_{EpiTag} are estimated based on the total population. Next, a response variable Y is simulated according to the logistic model (6) introduced in **Step ♯2** with several values for $\beta_1 \in [0;0.6]$. For each $nb.sim = 1,000$ simulations, 1,000 cases ($Y = 1$) and 1,000 controls ($Y = 0$) are picked at random in the population and SNPxSNP associations are tested according to **Step♯3**. We then use the procedure described in **Step♯4** to estimate the power of each compared method.

Results. Figure 1 displays the estimated power (mean over 1,000 replicates) for interaction detection along with the different values for β_1, for the 3 tagging procedures (*EpiTag*, *Tagger* and *noTag*) and the 3 scenarios (3, 6 and 10 SNPs per region). It can be firstly remarked that *EpiTag* outperforms *Tagger* and *NoTag* strategies. The gain in using *EpiTag* is even more substantial when the relative risk is high (highest values of β_1). The difference between *EpiTag* and *NoTag* seems to be equivalent from one scenario to another. As the number of SNPs per region increases, it is noteworthy that *EpiTag* is more efficient than other methods but, more surprisingly, the use

 ISSN: 2102-6238

of one-dimensional tagging (*Tagger*) appears to be similar or slightly less efficient that the no-tagging strategy in this case. The efficiency of the *EpiTag* method is related to its ability to select a small but sufficiently representative subset of SNP pairs. As reported in Table 1, the number of SNP pairs tested with *EpiTag* is low, compared to the number of SNP pairs tested with the *Tagger* method. From Table 1 it can also be remarked that the difference in the number of tested SNP pairs sharply increases along with the number of SNPs per region.

TABLE 1. *Mean number of SNP pairs and relative Standard Deviation (SD) computed for each simulated scenario and each tagging strategy over the nb.sim=1,000 simulations.*

	$p_1 = p_2 = 3$		$p_1 = p_2 = 6$		$p_1 = p_2 = 10$	
	Mean	SD	Mean	SD	Mean	SD
NoTag	9	0	36	0	100	0
Tagger	3.98	2.64	27.04	6.4	81.1	11.89
EpiTag	2.48	0.54	3.4	0.7	11.24	1.4

Observed genotype structure

Settings. An observed pattern of genotype data is used in this second power study. For that purpose, we focus our investigation on two genes, Adenomatous Polyposis Coli (*APC*) and the IQ-domain GTPase-activating protein 1 (*IQGAP1*), that have previously been reported to interact in susceptibility with Crohn's disease (Emily et al., 2009). Genes *APC* and *IQGAP1* are located on chromosome 5 ($p_1 = 13$ SNPs) and 15 ($p_2 = 14$ SNPs) respectively. The *Tagger* procedure selects 2 tag-SNPs for each region, and *EpiTag* tags 3 pairs of SNPs. To study the power of the tagging procedures on a real pattern of genotype, we consider $5,000$ genotype samples for *APC* and *IQGAP1*. Then, a total of $nb.sim = 1,000$ responses variables are simulated with respect to the logistic model introduced in **Step♯2** with several values for $\beta_1 \in [0;0.6]$. Power is then computed using the procedure described in **Step♯4**.

Results. Figure 2 displays the estimated power (mean over $1,000$ replicates) for interaction detection along with the different values for β_1, for the 3 tagging procedures (*EpiTag*, *Tagger* and *noTag*). Similarly to the simulation-based power studies, *EpiTag* outperforms *Tagger* and *NoTag*. In more details, it can be remarked that, although the gain in power for *EpiTag* is obvious for moderate values of β_1, the difference between *EpiTag* and *NoTag* becomes smaller for $\beta_1 = 0.6$. Finally, similarly to the results obtained with simulated genotype structures, using one-dimensional tagging seems to be much less efficient for detecting interaction.

5. Conclusion and discussion

In this paper, we introduce *EpiTag*, a novel method for tagging SNPs in genetic association studies. Contrary to existing methods, *EpiTag* is based on the Normalized Mutual Information, a two-dimensional measure of similarity, that allows a better tagging for pairs of SNPs. Our results prove that our method, by maximizing the information of SNP pairs instead of single SNP, improves the statistical power to detect interaction between SNPs. More precisely, when testing for interaction, accounting for the information at the level of the pair is more efficient

 ISSN: 2102-6238

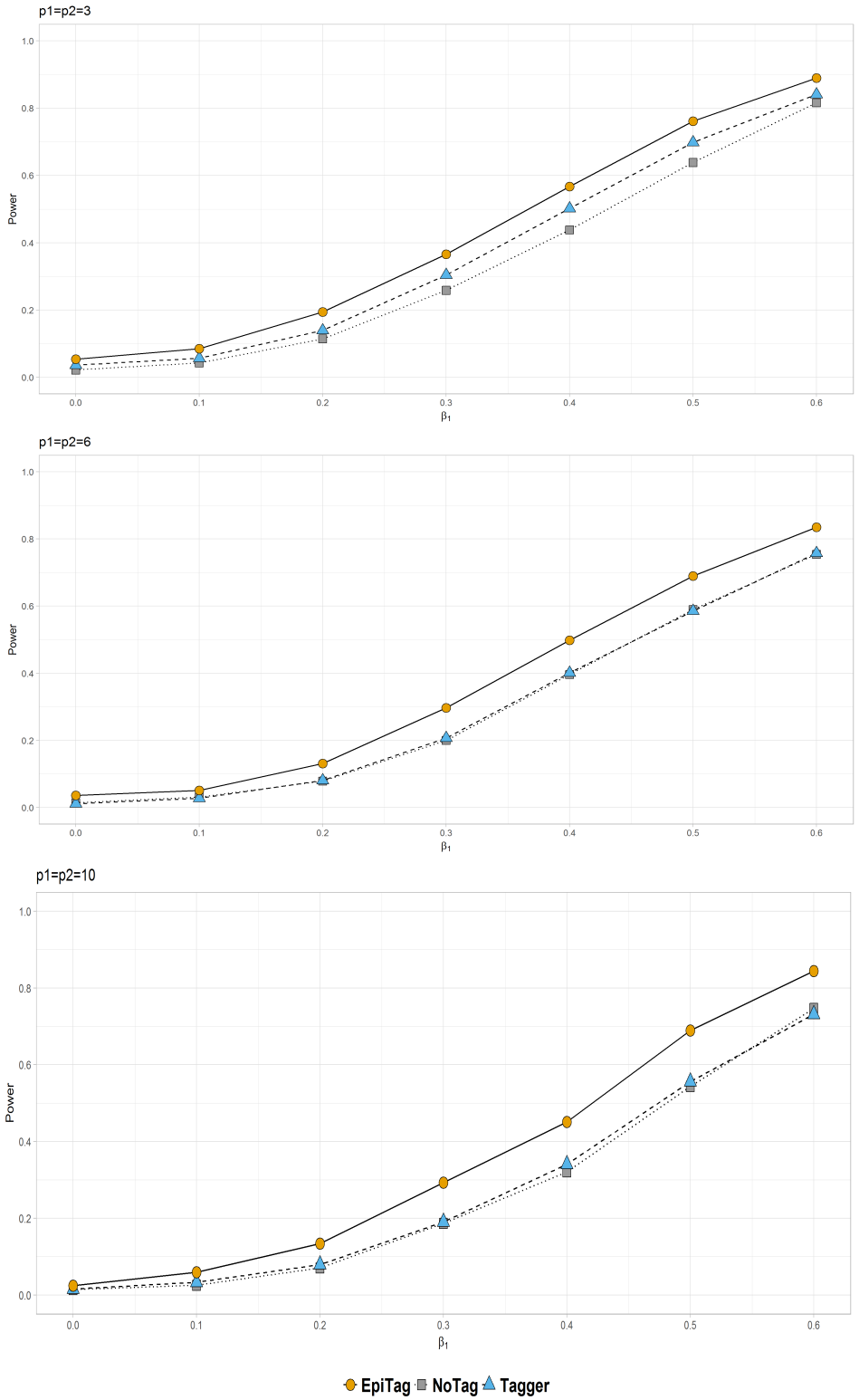

FIGURE 1. *Simulated genotype structures: Power for interaction detection along with different values for* β_1*, considering 3 scenarios for the region sizes.*

 ISSN: 2102-6238

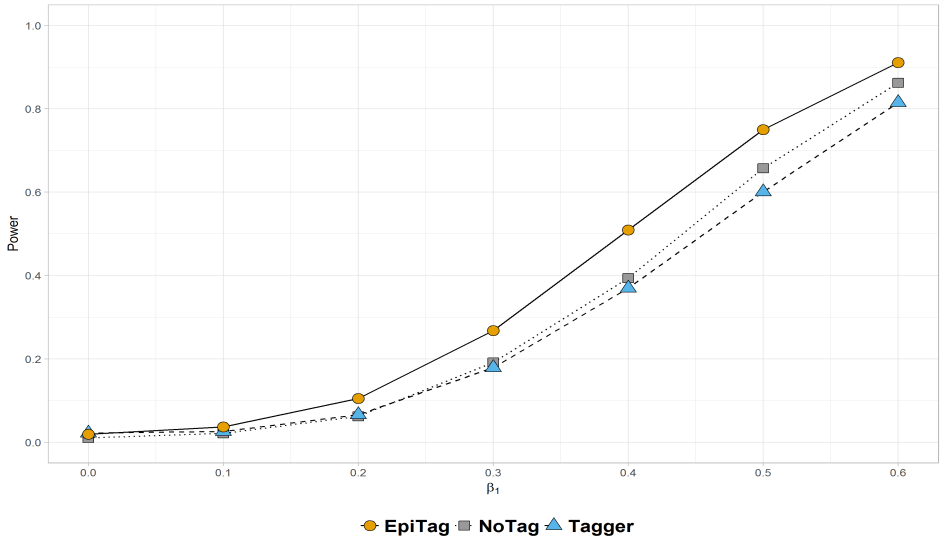

FIGURE 2. *Observed genotype structure: Power for interaction detection along with different values for* β_1.

than no selection and than using one-dimensional tagging. More surprisingly, our results show that using a one-dimensional tagging strategy is less efficient than avoiding tag selection. Such a result strengthens our hypothesis that one-dimensional tagging is not appropriate to detect interactions.

Despite the good performances of *EpiTag*, its practical use suffers from few limitations. First, similarly to the one-dimensional strategy *Tagger*, *EpiTag* is based on the choice of a threshold τ (Equation (2)) on the NMI values to build bins of SNP-pairs. In practice, the threshold is tuned by the user, depending on a balance between the level of coverage of the whole set of SNPs and a parcimonious set of tag-SNPs. In our proposal, we set $\tau = 0.75$ to mimic the mean of computed MNI values between 2 pairs with $r^2 = 0.8$ and the use of other thresholds shows a relative sensitivity to τ.

Although the ultimate goal of statistical methods for detecting interaction in GWAS is to analyze the whole genome at a time, whole-genome investigation of interaction still raises the issue of the scalability of classical methods. Since the number of SNPs is very large (estimated to 10 millions in humans), the number of SNP-pairs is even much larger ($\approx 5 \times 10^{13}$) and *EpiTag* is therefore hardly scalable. Computing and storing the whole NMI matrix, of size $(5 \times 10^{13}) \times (5 \times 10^{13})$, raises computational issues. Furthermore, if feasible, the computation of such a matrix is likely to be highly time-consuming. However, in the context of whole-genome gene-gene interaction testing, many different statistical strategies have been proposed in the recent literature such as logic regression (Kooperberg and Ruczinski, 2005), penalized regression (Park and Hastie, 2008), Random Forest (McKinney et al., 2009), Multifactor Dimensionality Reduction (Gola et al., 2016) or Support Vector Machines (Chen et al., 2008). Although we can hope that improvements in data management, data storage, computer performances and distributed computing could help reducing the computational cost, such a strategy would still remain too demanding to be performed in routine. Nowadays, the computational limitation of

 ISSN: 2102-6238

gene-gene interaction can only be overcome by the incorporation of prior biological knowledge in the analyse (Emily, 2018). For example, in the context of pathway disease-associated detection or gene-gene interaction network investigation, *EpiTag* can be performed on blocks of SNPs under consideration.

In the same way as *Tagger*, *EpiTag* is based on a greedy algorithm to select the most informative set of SNP-pairs. However, tag-SNP selection based on multivariate statistical techniques also exists in the literature. For example, in a one-dimensional strategy, cluster analysis has been proposed by considering $1 - r^2$ as a distance measure between SNPs (Ao et al., 2005; Frommlet, 2010). A similar strategy can be performed in two-dimensional tagging by considering $1 - NMI$ as a distance measure between SNP-pairs. Note that such a procedure is also sensitive to the tuning of a threshold parameter associated to the distance between clusters. Preliminary results regarding the use of such approach in a two-dimensional tagging show a lack in power in all situations.

Finally, we focus on the benefits of using a two-dimensional tagging strategy to improve the power to detect interaction (or epistasis). However, such a two-dimensional strategy is not appropriate to test for single-marker association since the output of *EpiTag* is a set of tag-SNP pairs. Therefore, one-dimensional strategy and two-dimensional strategy can be seen as complementary and combining these two approaches may be very efficient to detect simultaneously marginal effects, pure epistasis and epistasis with marginal effects.

References

Ao, S., Yip, K., Ng, M. K., Cheung, D., Fong, P.-Y., Melhado, I., and Sham, P. C. (2005). Clustag: Hierarchical clustering and graph methods for selecting tag-SNPs. *Bioinformatics*, 21(8):1735–1736.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, 8(12):1–11.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics*, 74(1):106 – 120.

Chen, S.-H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B.-L., Zheng, S. L., Gronberg, H., Xu, J., and Hsu, F.-C. (2008). A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology*, 32(2):152–167.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics*, 10(2):392–404.

de Bakker, P. I. W., Yelensky, R., Peer, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, 37:1217–1223.

Emily, M. (2012). Indor: a new statistical procedure to test for SNPxSNP epistasis in genome-wide association studies. *Statistics in Medicine*, 31(21):2359–2373.

Emily, M. (2016a). Aggregator: A gene-based gene-gene interacttion test for case-control association studies. *Statistical Application in Genetics and Molecular Biology*, 15(2):151–171.

Emily, M. (2016b). Power comparison of cochran-armitage test of trend against allelic and genotypic tests in case-control genetic association studies. *Statistical Methods in Medical Research*, page In press.

Emily, M. (2018). A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies. *Journal de la Société Française de Statistique*, 159(1):27–67.

Emily, M. and Friguet, C. (2017). Power evaluation of asymptotic tests for comparing two binomial proportions to detect direct and indirect association in large-scale studies. *Statistical Methods in Medical Research*, 26(6):2780–2799.

 ISSN: 2102-6238

Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17:1231–1240.

Frommlet, F. (2010). Tag-SNP selection based on clustering according to dominant sets found using replicator dynamics. *Advances in Data Analysis and Classification*, 4(1):65–83.

Gola, D., Mahachie John, J. M., van Steen, K., and Knig, I. R. (2016). A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 17(2):293.

Hallgrimsdottir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, 9(17).

Hill, W. G. and Robertson, A. (1968). Linkage diseqilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and A., T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceeding of the National Academy of Sciences*, 106(23):9362–9367.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426:789–796.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting snps using monte carlo logic regression. *Genetic Epidemiology*, 28:157–170.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Larson, N. B. and Schaid, D. J. (2013). A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, 37(7):695–703.

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2):146–153.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50(6):334–349.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456:18–21.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417.

McKinney, B. A., Crowe, Jr, J. E., Guo, J., and Tian, D. (2009). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLOS Genetics*, 5(3):1–12.

Ng, V. W. Y. (2004). *Univariate and Bivariate Variable Selection in High Dimensional Data*. PhD thesis, University of California at Berkeley, Berkeley, CA, USA. AAI3167213.

Nielsen, D. M., Ehm, M. G., Zaykin, D. V., and Weir, B. S. (2004). Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics*, 168(2):1029–1040.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9:30–50.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics*, 69:1 – 14.

Ritchie, M. D. (2015). *Finding the Epistasis Needles in the Genome-Wide Haystack*, pages 19–33. Springer New York, New York, NY.

Sicotte, H., Rider, D. N., Poland, G. A., Dhiman, N., and Kocher, J.-P. A. (2011). Snppicker: High quality tag snp selection across multiple populations. *BMC Bioinformatics*, 12(1):129.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Ueki, M. and Cordell, H. J. (2012). Improved statistics for genome-wide interaction analysis. *PLoS Genet*, 8(4):e1002625.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., and Yu, W. (2010). Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87:325–340.

Weir, B. S. (2008). Linkage disequilibrium and association mapping. *Annual Review of Genomics and Human Genetics*, 9:129–142.

 ISSN: 2102-6238

Appendix A: NMI behaviour w.r.t. the data design

We consider the simulation of two pairs of SNPs on two genomic regions ($p_1 = p_2 = 2$). The aim is to illustrate the variability of NMI values between the pairs, according to the r^2 correlation between SNPs within a region. The major allele frequency for each SNP is set to $\pi_1^1 = \pi_2^1 = \pi_1^2 = \pi_2^2 = 0.6$. Furthermore, the between correlation structure is defined so that variables from $\mathscr{R}_2$ are not correlated with variables from $\mathscr{R}_1$. The r^2 coefficient within each region are equal: $r^2 = r^2(X_1^1, X_2^1) = r^2(X_1^2, X_2^2)$ with $r^2 \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Each setup is repeated $1,000$ times.

For $r^2 = 0.8$ (as in the simulation study presented in Section 4), the mean value of NMI is about 0.75. This is the value we have chosen for the threshold τ in the simulation study.

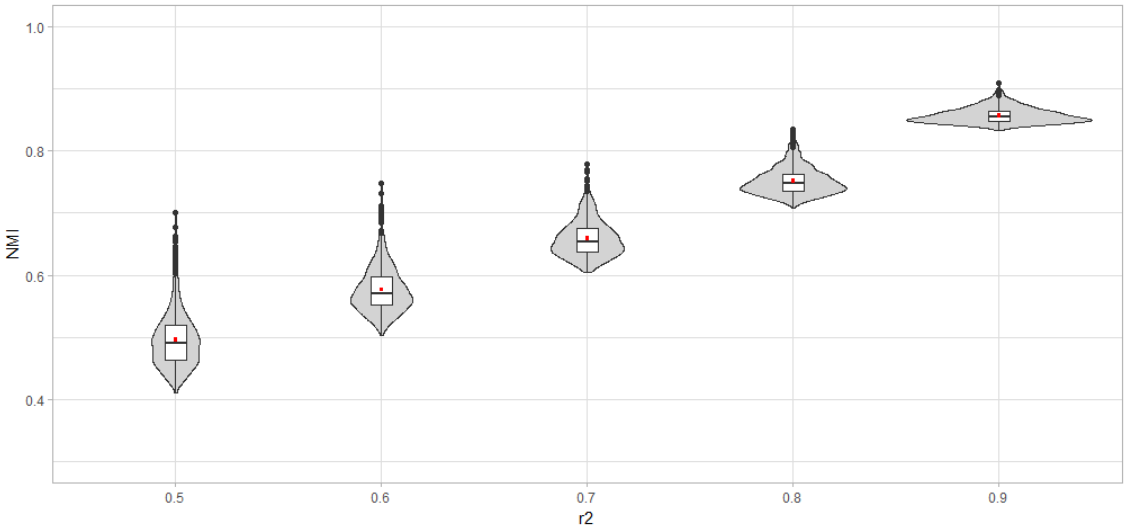

FIGURE 3. *NMI between two pairs over* $1,000$ *simulations with different settings for the* r^2 *coefficient within each region.*

Appendix B: Graphical scheme for *Tagger* and *EpiTag*

Figures 4 and 5 display the flowcharts of two tagging methods: the one-dimensional *Tagger* and our pairwise tagging method *EpiTag*. From Figures 4 and 5, it can be remarked that the core of the two algorithms is similar. Both algorithms are indeed based (1) on an initializing step that aims at computing a matrix of similarity, and (2) on iterative steps used to build the set of tags and the set of bins. However, since *EpiTag* is applied to an initial set of SNP pairs and uses the pairwise NMI as quality measure, it allows the tagging of pairs of SNPs.

 ISSN: 2102-6238

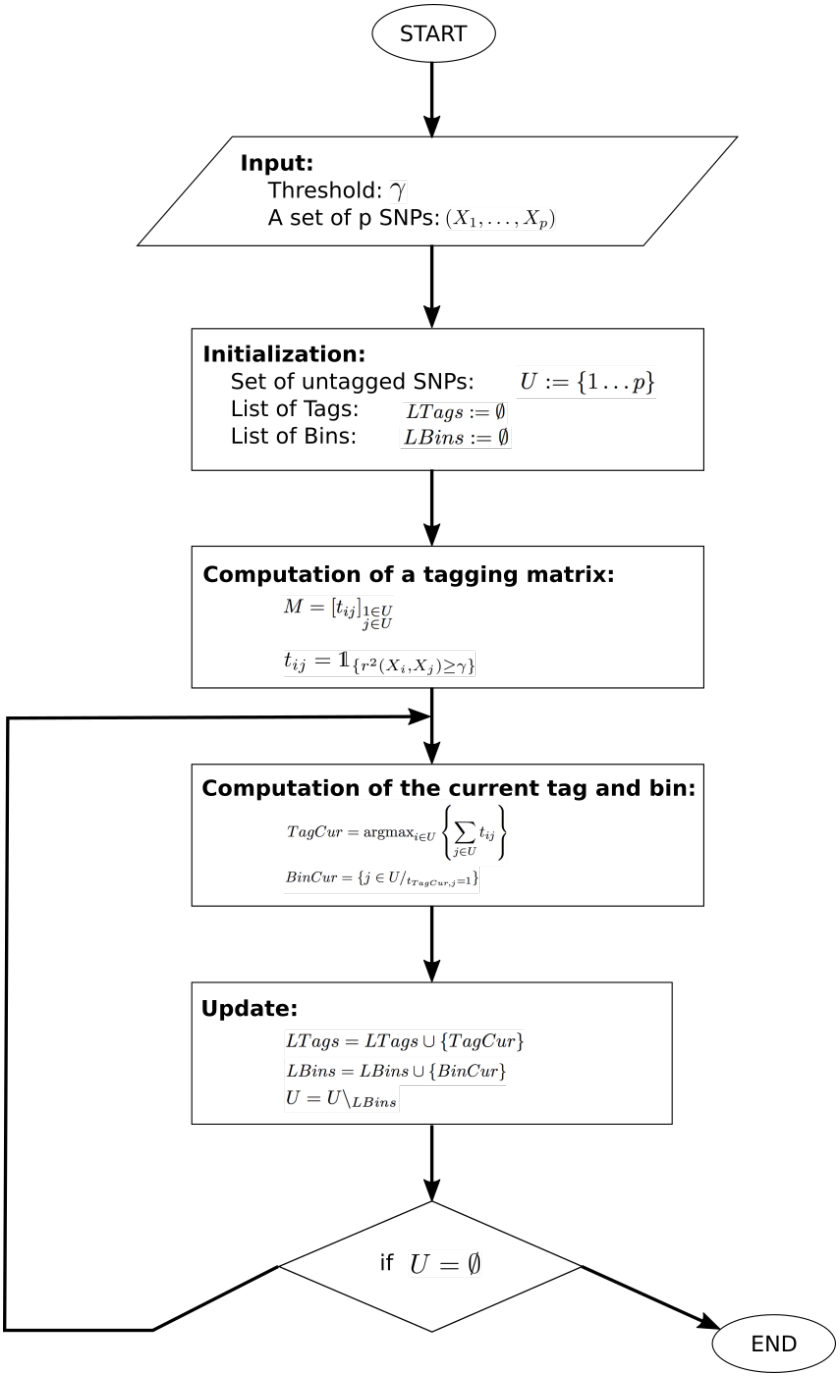

FIGURE 4. *Flowchart summarizing the main steps of the Tagger algorithm.*

 ISSN: 2102-6238

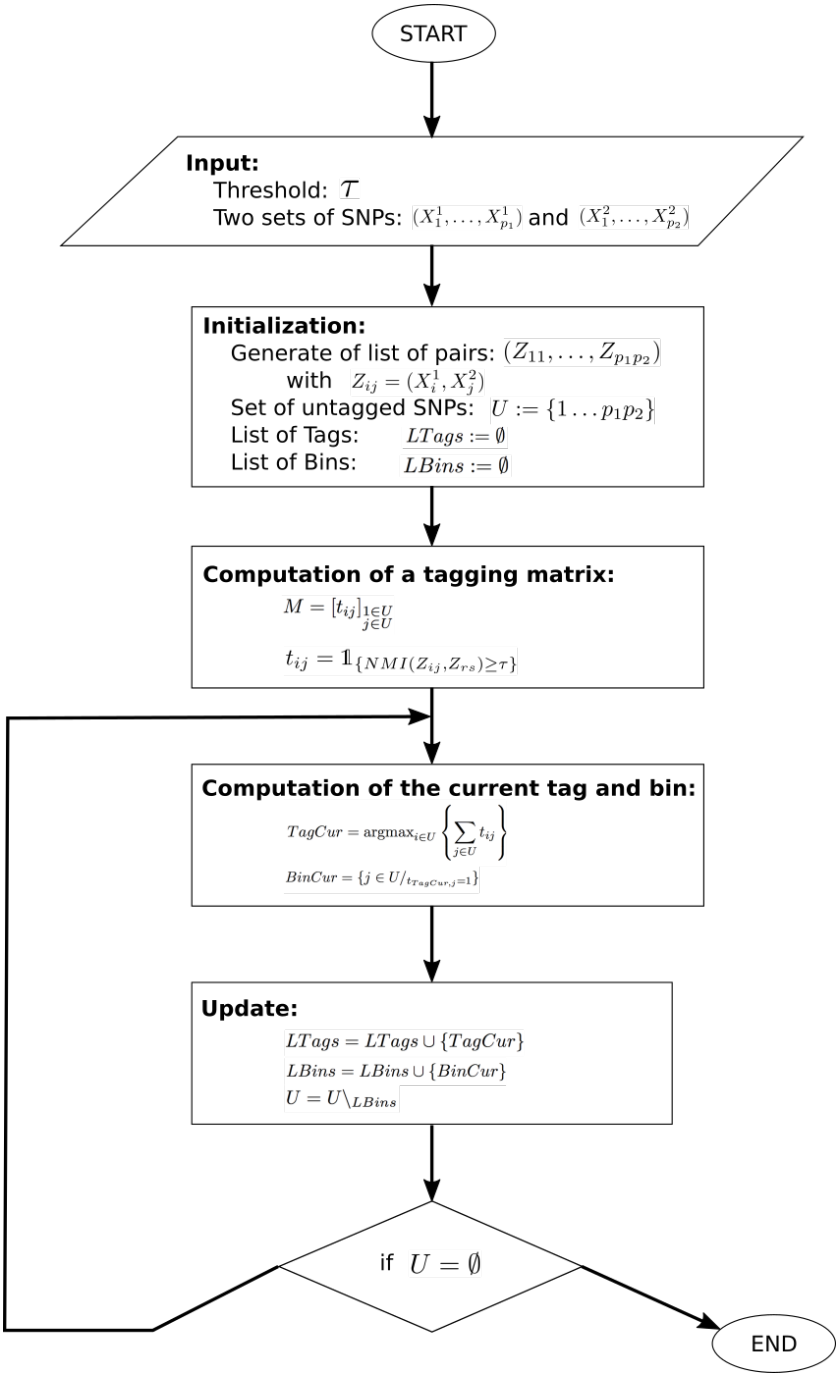

FIGURE 5. *Flowchart summarizing the main steps of the EpiTag algorithm.*

 ISSN: 2102-6238

Appendix C: Study of the scalability of the method

Due to its lack of scalability, the practical use of *EpiTag* at the genome scale is not yet feasible. However, in this Appendix, details regarding the memory needs and the computational time are provided.

As shown in Figure 5, one of the main step in the *EpiTag* algorithm is the computation of the initial tagging matrix. This step is the most consuming step in terms of memory needs. If no prior information is used to restrict the computation to predefined subset of SNPs, such a matrix has to be initialize with all combinations of pairs of SNPs. If we consider a set of p SNPs, the total number of elements in the tagging matrix that have to be computed is given by $\binom{nPairs}{2}$ where $nPairs = \binom{p}{2}$. Such a number is growing along with p (and more precisely p^4) which is not tractable for a whole-genome tagging. For example, if we consider a set of $p = 1,000,000$ SNPs, the total number of elements is approximatively 1.25×10^{23}. The storage of such object in memory is hardly possible. However, to save memory, efficient algorithmic paradigms may be used, such as Branch and Bound solution for example.

In terms of computational time, the two steps mostly demanding are: (1) the initialization of the tagging matrix and (2) the iterative building of the list of tags and the list of bins (see Figure 5). First, as for the memory, the computation time required to compute the tagging matrix is exponential in the number of SNPs. Next, the iterative building of the lists tags and bins depends on the mutual information structure of the SNPs. However, in the worth-case scenario where each pair is only tagged by itself, the computational time is of the order of the number of SNP pairs, *i.e.* also exponential in the number of SNPs. Furthermore, the sample size also plays a role in the computation time.

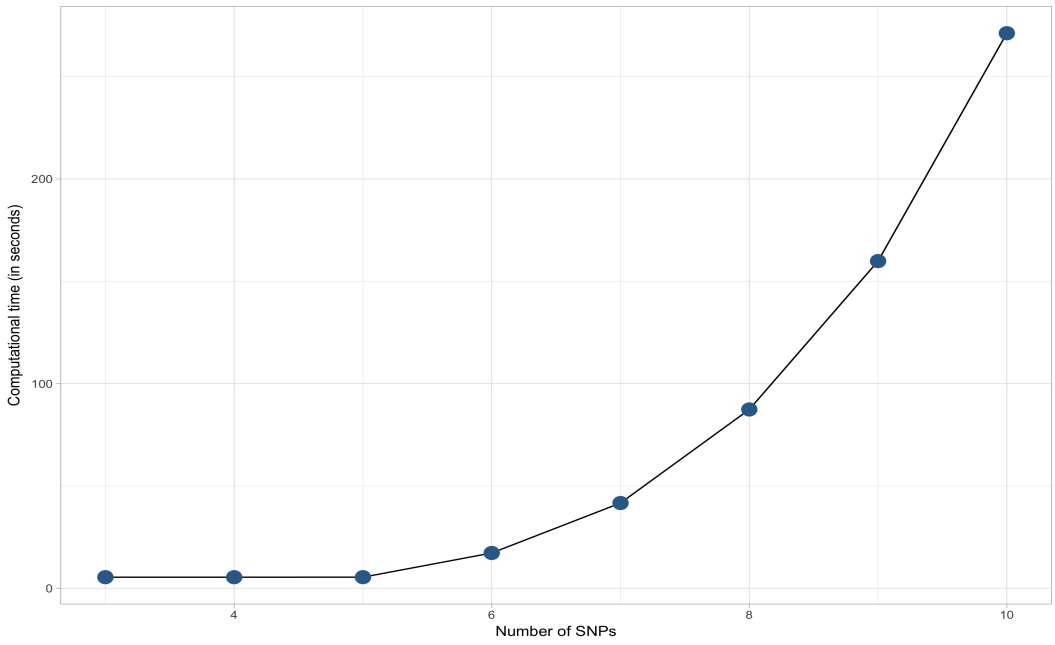

FIGURE 6. *Evolution of the computation time with respect to the number of SNPs.*

 ISSN: 2102-6238

Figure 6 displays the evolution of the computation time with respect to the number of SNPs where it can be remarked that the shape of the curve is indeed exponential. Computation time is obtained for a sample size of 10,000 individuals on a single processor with the following characteristics: 2,7 GHz Intel Core i7. Functions distributed in the following github repository https://github.com/MathieuEmily/EpiTag are used to run *EpiTag*. It can be remarked that *EpiTag* takes about 270 seconds to run for 10 SNPs. However in both steps, execution time can be drastically reduced using both algorithmic optimization and parallelization.

Appendix D: Details on genotype simulation

In this appendix, we provide details regarding the simulation of genotype data as introduced in Step #1, Section 4.1. We focus on two specific operations, namely the adding of a single SNP and the adding of a SNP-pair. For both operations, we consider that SNPs are in Hardy-Weinberg equilibrium so that each chromosome is simulated independently. More precisely, for a given SNP $X_i \in \{0,1,2\}$, the allele on each chromosome is simulated independently through two random variables $XA_i \in \{0,1\}$ and $XB_i \in \{0,1\}$, X_i being obtained with: $X_i = XA_i + XB_i$.

D.1. Single-SNP adding

The addition of a single SNP to a dataset is performed conditionally to one other SNP. Let consider that the SNP X_i has already been simulated where each allele XA_i and XB_i have an observed allelic frequency p_i. Let p_j be the desired allelic frequency for both alleles XA_j and XB_j of SNP X_j and r_{ij} be the correlation between SNP i and j. The correlation r_{ij} can be formulated in each chromosome as for example:

$$r_{ij}^2 = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$$

with $p_{ij} = \mathbb{P}[XA_i = 1 \cap XA_j = 1]$. The joint probability distribution is given by:

$$\begin{aligned}
\mathbb{P}[XA_i = 0 \cap XA_j = 0] &= (1-p_i)(1-p_j) + r_{ij}\sqrt{p_i(1-p_i)p_j(1-p_j)} \\
\mathbb{P}[XA_i = 1 \cap XA_j = 0] &= p_i(1-p_j) - r_{ij}\sqrt{p_i(1-p_i)p_j(1-p_j)} \\
\mathbb{P}[XA_i = 0 \cap XA_j = 1] &= (1-p_i)p_j - r_{ij}\sqrt{p_i(1-p_i)p_j(1-p_j)} \\
p_{ij} = \mathbb{P}[XA_i = 1 \cap XA_j = 1] &= p_i p_j + r_{ij}\sqrt{p_i(1-p_i)p_j(1-p_j)}
\end{aligned}$$

The conditionnal distribution of SNP allele XA_j can easily be obtained since

$$\mathbb{P}[XA_j = k | XA_i = \ell] = \frac{\mathbb{P}[XA_j = k \cap XA_i = \ell]}{\mathbb{P}[XA_i = \ell]}.$$

It is noteworthy that each combination of p_i, p_j and r_{ij}^2 are not possible since joint probabilities may be negative. Finally, to simulate SNP allele XA_j, the genotype for each individual is straightforwardly computed according to the conditional probability $\mathbb{P}[XA_j = 1 | XA_i = x_i]$. Similarly and independently, SNP allele XB_j can be simulated conditionally to XB_i. Using Hardy-Weinberg equilibrium SNP X_j is obtained by $X_j = XA_j + XB_j$.

 ISSN: 2102-6238

D.2. SNP-pair adding

To add a pair of SNP $Z_{uv} = (X_u^1, X_v^2) = (XA_u^1 + XB_u^1, XA_v^2 + XB_v^2)$ to an existing set of SNPs, we specify the relationship between Z_{uv} and an already simulated pair $Z_{rs} = (X_r^1, X_s^2) = (XA_r^1 + XB_r^1, XA_s^2 + XB_s^2)$. For that purpose, the joint probability distribution of the quadruplet $(X_r^1, X_s^2, X_u^1, X_v^2)$, given by $p_{ijk\ell} = \mathbb{P}[X_r^1 = i \cap X_s^2 = j \cap X_u^1 = k \cap X_v^2 = \ell]$ with $(i, j, k, \ell) \in [0,1,2]^4$, is built under a set of constraints. More precisely, alleles in both chromosomes are first simulated according to the joint probability distribution of the quadruplets $(XA_r^1, XA_s^2, XA_u^1, XA_v^2)$ and $(XB_r^1, XB_s^2, XB_u^1, XB_v^2)$. As displayed in Table 2, the joint distribution of both quadruplets $(XA_r^1, XA_s^2, XA_u^1, XA_v^2)$ and $(XB_r^1, XB_s^2, XB_u^1, XB_v^2)$ is given by 16 probabilities, namely $q_{ijk\ell}$ with $(i, j, k, \ell) \in [0,1]^4$. $q_{ijk\ell}$'s are then derived with respect to four types of constraints that are details hereafter: (A) an overall constraint on the probability distribution, (B) constraints regarding the marginal distribution of each SNP, (C) constraints regarding the pairwise relationship between each pair and (D) a constraint related to the mutual information between Z_{rs} and Z_{uv}.

TABLE 2. *Joint distribution of the quadruplet* $(XA_r^1, XA_s^2, XA_u^1, XA_v^2)$ *where* $q_{ijk\ell} = \mathbb{P}\left[XA_r^1 = i \quad \cap \quad XA_s^2 = j \quad \cap \quad XA_u^1 = k \quad \cap \quad XA_v^2 = \ell\right]$.

		$XA_u^1 = 0$		$XA_u^1 = 1$	
		$XA_v^2 = 0$	$XA_v^2 = 1$	$XA_v^2 = 0$	$XA_v^2 = 1$
$XA_r^1 = 0$	$X_s^2 = 0$	q_{0000}	q_{0001}	q_{0010}	q_{0011}
	$XA_s^2 = 1$	q_{0100}	q_{0101}	q_{0110}	q_{0111}
$XA_r^1 = 1$	$X_s^2 = 0$	q_{1000}	q_{1001}	q_{1010}	q_{1011}
	$XA_s^2 = 1$	q_{1100}	q_{1101}	q_{1110}	q_{1111}

Similarly and independently, the quadruplet $(XB_r^1, XB_s^2, XB_u^1, XB_v^2)$ is simulated using the same joint distribution as in Table 2.

Finally, $\forall (i, j, k, \ell) \in [0,1,2]^4$,

$$\begin{aligned} p_{ijk\ell} &= \mathbb{P}\left[X_r^1 = i \quad \cap \quad X_s^2 = j \quad \cap \quad X_u^1 = k \quad \cap \quad X_v^2 = \ell\right] \\ &= \sum_{(iA,jA,kA,\ell A)\in[0,1]^4} \left(\mathbb{P}\left[XB_r^1 = i - iA \quad \cap \quad XB_s^2 = j - jA \quad \cap \quad XB_u^1 = k - kA \quad \cap \quad XB_v^2 = \ell - \ell A\right] \right. \\ &\quad \left. \times \quad \mathbb{P}\left[XA_r^1 = iA \quad \cap \quad XA_s^2 = jA \quad \cap \quad XA_u^1 = kA \quad \cap \quad XA_v^2 = \ell A\right] \right) \\ &= \sum_{(iA,jA,kA,\ell A)\in[0,1]^4} q_{i-iA,j-jA,k-kA,\ell-\ell A} \quad \times \quad q_{iA,jA,kA,\ell A} \end{aligned} \tag{7}$$

(A). Constraint on the overall probability distribution

The first constraint **C0** is straightforward since we deal with a probability distribution.

$$\textbf{C0:} \quad \textstyle\sum_{i,j,k,\ell} q_{ijk\ell} = 1$$

 ISSN: 2102-6238

(B). Constraints on marginal probabilities for each SNP

Marginal distribution of each SNP are considered as fixed and the next constraints present the corresponding relationship between the $q_{ijk\ell}$.

Constraints on X_r^1

$$\textbf{C1:} \quad \sum_{j,k,\ell} q_{1jk\ell} = p_{XA_r^1} = \mathbb{P}[XA_r^1 = 1]$$
$$\textbf{C1b:} \quad \sum_{j,k,\ell} q_{0jk\ell} = 1 - p_{XA_r^1} = \mathbb{P}[XA_r^1 = 0]$$

Constraints on X_s^2

$$\textbf{C2:} \quad \sum_{i,k,\ell} q_{i1k\ell} = p_{XA_s^2} = \mathbb{P}[XA_s^2 = 1]$$
$$\textbf{C2b:} \quad \sum_{i,k,\ell} q_{i0k\ell} = 1 - p_{XA_s^2} = \mathbb{P}[XA_s^2 = 0]$$

Constraints on X_u^1

$$\textbf{C3:} \quad \sum_{i,j,\ell} q_{ij1\ell} = p_{XA_u^1} = \mathbb{P}[XA_u^1 = 1]$$
$$\textbf{C3b:} \quad \sum_{i,j,\ell} q_{ij0\ell} = 1 - p_{XA_u^1} = \mathbb{P}[XA_u^1 = 0]$$

Constraints on X_v^2

$$\textbf{C4:} \quad \sum_{i,j,k} q_{ijk1} = p_{XA_v^2} = \mathbb{P}[XA_v^2 = 1]$$
$$\textbf{C4b:} \quad \sum_{i,j,k} q_{ijk0} = 1 - p_{XA_v^2} = \mathbb{P}[XA_v^2 = 0]$$

It is straighforward to show that constraints **C1b**, **C2b**, **C3b** and **C4b** are respectively equivalent to **C1**, **C2**, **C3** and**C4**. Therefore, for each SNP, there is only one link between the $q_{ijk\ell}$.

(C). Constraints on the joint probability for each pair of SNPs

In our pipeline, we consider that relationships between pairs of SNPs are fixed and parameterized by the statistical correlation. Setting the pairwise correlation induces constraints on the $q_{ijk\ell}$. In the following, we focus on the pair (X_r^1, X_s^2) and then extend the expression to all the pairs.

Constraints on (X_r^1, X_s^2). The correlation is denoted by $r_{X_r^1,X_s^2}$, with:

$$r^2_{X_r^1,X_s^2} = \frac{\left(\mathbb{P}[XA_r^1 = 1 \cap XA_s^2 = 1] - \mathbb{P}[XA_r^1 = 1]\mathbb{P}[XA_s^2 = 1]\right)^2}{\mathbb{P}[XA_r^1 = 0]\mathbb{P}[XA_r^1 = 1]\mathbb{P}[XA_s^2 = 0]\mathbb{P}[XA_s^2 = 1]}$$

Therefore, we have:

$$\begin{aligned}\mathbb{P}[XA_r^1 = 1 \cap XA_s^2 = 1] &= \mathbb{P}[XA_r^1 = 1]\mathbb{P}[XA_s^2 = 1] \\ &+ r_{XA_r^1,XA_s^2}\sqrt{\mathbb{P}[XA_r^1 = 0]\mathbb{P}[XA_r^1 = 1]\mathbb{P}[XA_s^2 = 0]\mathbb{P}[XA_s^2 = 1]}\end{aligned}$$

$$\Updownarrow$$

$$\sum_{k\ell} q_{11k\ell} = \sum_{j,k,\ell} q_{1jk\ell} \sum_{i,k,\ell} q_{i1k\ell} + r_{XA_r^1,XA_s^2}\sqrt{\sum_{j,k,\ell} q_{0jk\ell} \sum_{j,k,\ell} q_{1jk\ell} \sum_{i,k,\ell} q_{i0k\ell} \sum_{i,k,\ell} q_{i1k\ell}}$$

 ISSN: 2102-6238

Therefore, constraints regarding the whole joint probability distribution for the pair (X_r^1, X_s^2) are given by:

$$\textbf{C5:}\quad \sum_{k\ell} q_{11k\ell} = \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell} + r_{XA_r^1,XA_s^2} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell}}$$

$$\textbf{C5b:}\quad \sum_{k\ell} q_{01k\ell} = \sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell} - r_{XA_r^1,XA_s^2} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell}}$$

$$\textbf{C5c:}\quad \sum_{k\ell} q_{10k\ell} = \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} - r_{XA_r^1,XA_s^2} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell}}$$

$$\textbf{C5d:}\quad \sum_{k\ell} q_{00k\ell} = \sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} + r_{XA_r^1,XA_s^2} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell}}$$

It is crucial to remark that given the marginal probabilities ($\mathbb{P}[XA_r^1 = 0]$ and $\mathbb{P}[XA_s^2 = 0]$), some values of correlation $r^2_{X_r^1,X_s^2}$ are not possible. According to constraints **C5, C5b, C5c, C5d**, joint probabilities can be either negative or higher than one with unacceptable combination of $\mathbb{P}[XA_r^1 = 0]$, $\mathbb{P}[XA_s^2 = 0]$ and $r^2_{XA_r^1,XA_s^2}$.

Moreover, considering that $\mathbb{P}[XA_r^1 = 0] = 1 - \mathbb{P}[XA_r^1 = 1]$, $\mathbb{P}[XA_s^2 = 0] = 1 - \mathbb{P}[XA_s^2 = 1]$ and $\mathbb{P}[XA_r^1 = 0 \cap XA_s^2 = 0] + \mathbb{P}[XA_r^1 = 0 \cap XA_s^2 = 1] + \mathbb{P}[XA_r^1 = 1 \cap X_s^2 = 0] + \mathbb{P}[XA_r^1 = 1 \cap XA_s^2 = 1]$, it is straightforward to show that constraints **C5b, C5c** and **C5d** are equivalent to **C5**.

Extension to the other pairs. Therefore, for each pair, there is only one constraint linking the $q_{ijk\ell}$. More precisely, these constraints are:

- (XA_r^1, XA_u^1)

$$\textbf{C6:}\sum_{j\ell} q_{1j1\ell} = \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,j,\ell} q_{i,j,1,\ell} + r_{XA_r^1,XA_u^1} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,j,\ell} q_{i,j,0,\ell} \sum_{i,j,\ell} q_{i,j,1,\ell}}$$

- (XA_r^1, XA_v^2)

$$\textbf{C7:}\sum_{jk} q_{1jk1} = \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,j,k} q_{i,j,k,1} + r_{XA_r^1,XA_v^2} \sqrt{\sum_{j,k,\ell} q_{0,j,k,\ell} \sum_{j,k,\ell} q_{1,j,k,\ell} \sum_{i,j,k} q_{i,j,k,0} \sum_{i,j,k} q_{i,j,k,1}}$$

- (XA_s^2, XA_u^1)

$$\textbf{C8:}\sum_{i\ell} q_{i11\ell} = \sum_{i,k,\ell} q_{i,1,k,\ell} \sum_{i,j,\ell} q_{i,j,1,\ell} + r_{XA_s^2,XA_u^1} \sqrt{\sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell} \sum_{i,j,\ell} q_{i,j,0,\ell} \sum_{i,j,\ell} q_{i,j,1,\ell}}$$

- (XA_s^2, XA_v^2)

$$\textbf{C9:}\sum_{ik} q_{i1k1} = \sum_{i,k,\ell} q_{i,1,k,\ell} \sum_{i,j,k} q_{i,j,k,1} + r_{XA_s^2,XA_v^2} \sqrt{\sum_{i,k,\ell} q_{i,0,k,\ell} \sum_{i,k,\ell} q_{i,1,k,\ell} \sum_{i,j,k} q_{i,j,k,0} \sum_{i,j,k} q_{i,j,k,1}}$$

- (XA_u^1, XA_v^2)

$$\textbf{C10:}\sum_{ij} q_{ij11} = \sum_{i,j,\ell} q_{i,j,1,\ell} \sum_{i,j,k} q_{i,j,k,1} + r_{XA_u^1,XA_v^2} \sqrt{\sum_{i,j,\ell} q_{i,j,0,\ell} \sum_{i,j,\ell} q_{i,j,1,\ell} \sum_{i,j,k} q_{i,j,k,0} \sum_{i,j,k} q_{i,j,k,1}}$$

 ISSN: 2102-6238

(D). Constraints on the link between Z_{rs} and Z_{uv}

To simulate the pair Z_{uv} with respect to the pair Z_{rs}, we set the Normalized Mutual Information (NMI) between Z_{rs} and Z_{uv}. Let's recall that

$$NMI\Big(Z_{rs},Z_{uv}\Big) \quad = \quad \frac{I\big[Z_{rs},Z_{uv}\big]}{\sqrt{H(Z_{rs})H(Z_{uv})}}$$

with

$$\begin{aligned} I\Big(Z_{rs},Z_{uv}\Big) &= \sum_{i,j,k,\ell} p_{i,j,k,\ell} \log\left(\frac{p_{i,j,k,\ell}}{\mathbb{P}\big(Z_{rs}=z_{rs}\big)\mathbb{P}\big(Z_{uv}=z_{uv}\big)}\right) \\ H\left(Z_{ij}\right) &= I\Big(Z_{ij},Z_{ij}\Big) \end{aligned}$$

It can then be deduced that setting $NMI\Big(Z_{rs},Z_{uv}\Big)$ generate non-linear constraints on the $p_{i,j,k,\ell}$'s.

D.3. Simulation of Z_{uv} with respect to Z_{rs}

To add the pair Z_{uv} with respect to Z_{rs}, we first set $p_{X_r^1}$, $p_{X_s^2}$, $p_{X_u^1}$, $p_{X_v^2}$, $r^2_{X_r^1,X_s^2}$, $r^2_{X_r^1,X_u^1}$, $r^2_{X_r^1,X_v^2}$, $r^2_{X_s^2,X_u^1}$, $r^2_{X_s^2,X_v^2}$, $r^2_{X_u^1,X_v^2}$ and $NMI(Z_{rs},Z_{uv})$. The "adding pair" simulation algorithm is divided into two steps: firstly, a collection of joint probability distributions are obtained with respect to $p_{X_r^1}$, $p_{X_s^2}$, $p_{X_u^1}$, $p_{X_v^2}$, $r^2_{X_r^1,X_s^2}$, $r^2_{X_r^1,X_u^1}$, $r^2_{X_r^1,X_v^2}$, $r^2_{X_s^2,X_u^1}$, $r^2_{X_s^2,X_v^2}$, $r^2_{X_u^1,X_v^2}$. Then, the joint distribution with the expected NMI is selected. These two steps are descibed with more details hereafter.

Step#1 To summarize, given the constraints on the marginal distribution of each SNP (*i.e.* $p_{X_r^1}$, $p_{X_s^2}$, $p_{X_u^1}$ and $p_{X_v^2}$) and the constraints on the pairwise joint distribution between each pair (X_r,X_s) (*i.e.* $r^2_{X_r^1,X_s^2}$, $r^2_{X_r^1,X_u^1}$, $r^2_{X_r^1,X_v^2}$, $r^2_{X_s^2,X_u^1}$, $r^2_{X_s^2,X_v^2}$, $r^2_{X_u^1,X_v^2}$), the joint distribution displayed in Table 2 can be computed with respect to the following system of constraints:

$$\mathscr{S}: \begin{cases} \textbf{C0:} & \sum_{i,j,k,\ell} q_{ijk\ell} = 1 \\ \textbf{C1:} & \sum_{j,k,\ell} q_{1jk\ell} = p_{X_r^1} \\ \textbf{C2:} & \sum_{i,k,\ell} q_{i1k\ell} = p_{X_s^2} \\ \textbf{C3:} & \sum_{i,j,\ell} q_{ij1\ell} = p_{X_u^1} \\ \textbf{C4:} & \sum_{i,j,k} q_{ijk1} = p_{X_v^2} \\ \textbf{C5:} & \sum_{k\ell} q_{11k\ell} = p_{X_r^1}p_{X_s^2} + r_{X_r^1,X_s^2}\sqrt{p_{X_r^1}(1-p_{X_r^1})p_{X_s^2}(1-p_{X_s^2})} \\ \textbf{C6:} & \sum_{j\ell} q_{1j1\ell} = p_{X_r^1}p_{X_u^1} + r_{X_r^1,X_u^1}\sqrt{p_{X_r^1}(1-p_{X_r^1})p_{X_u^1}(1-p_{X_u^1})} \\ \textbf{C7:} & \sum_{jk} q_{1jk1} = p_{X_r^1}p_{X_v^2} + r_{X_r^1,X_v^2}\sqrt{p_{X_r^1}(1-p_{X_r^1})p_{X_v^2}(1-p_{X_v^2})} \\ \textbf{C8:} & \sum_{i\ell} q_{i11\ell} = p_{X_s^2}p_{X_u^1} + r_{X_r^1,X_v^2}\sqrt{p_{X_s^2}(1-p_{X_s^2})p_{X_u^1}(1-p_{X_u^1})} \\ \textbf{C9:} & \sum_{ik} q_{i1k1} = p_{X_s^2}p_{X_v^2} + r_{X_s^2,X_v^2}\sqrt{p_{X_s^2}(1-p_{X_s^2})p_{X_v^2}(1-p_{X_v^2})} \\ \textbf{C10:} & \sum_{ij} q_{ij11} = p_{X_u^1}p_{X_v^2} + r_{X_u^1,X_v^2}\sqrt{p_{X_u^1}(1-p_{X_u^1})p_{X_v^2}(1-p_{X_v^2})} \end{cases}$$

 ISSN: 2102-6238

More precisely, we first simulate q_{1111} using a uniform prior with a support ensuring that constraints **C5**, **C5b**, **C5c** and **C5d** does not lead to negative probabilities. Then, the system of constraints is updated and the other probabilities are recursively simulated (using a uniform prior with an appropriate support) until the system is satisfied. It can be remarked that the system $\mathscr{S}$ has 10 linear constraints while Table 2 is composed of 16 probabilities. Therefore, the simulation algorithm has 6 degrees-of-freedom that correspond to the uniform simulation of 6 probabilities.

The joint distribution of the genotypes, *i.e.* the 81 probabilities $p_{ijk\ell}$ with $(i,j,k,\ell) \in [0,1,2]^4$, is then computed with respect to Equation 7.

Step#2. Step#1 is recursively performed until the NMI between Z_{rs} and Z_{uv} reached the targeted value for $NMI(Z_{rs}, Z_{uv})$.

 ISSN: 2102-6238