Gene-based bin analysis of genome-wide association studies

Nicolas Omont[1,2]

Karl Forner[1]

Marc Lamarine[1]

Gwendal Martin[1]

François Képès[2,3]
and

Jérôme Wojcik[1,*]
jerome.wojcik@merckserono.net

ABSTRACT

With the improvement of genotyping technologies and the exponentially growing number of available markers, case-control genome-wide association studies promise to be a key tool for investigation of complex diseases. However new analytical methods have to be developed to face the problems induced by this data scale-up, such as statistical multiple testing, data quality control, biological interpretation and computational tractability. We present a novel method to analyze genome-wide association studies results. The algorithm is based on a Bayesian model that integrates genotyping errors and genomic structure dependencies. Probability values are assigned to genomic regions termed bins, which are defined from a gene-biased partitioning of the genome, and the false-discovery rate is estimated. We have applied this algorithm to data coming from three genome-wide association studies of Multiple Sclerosis. The method practically overcomes the scale-up problems and permits to identify new putative regions statistically associated with the disease.

Subject headings: case-control, association studies, bins, false discovery rate (FDR), inhomogeneous hidden Markov chain, Multiple Sclerosis, Bayesian Model

[1]Merck Serono, 9 chemin des Mines, 1211 Geneva, Switzerland

[2]CNRS UMR 8071 Laboratoire Statistique et Génome, 523 place des Terrasses, 91000 Évry, France

[3]Epigenomics Project, Genopole®, CNRS and Université d'Évry, 523, place des Terrasses, 91034 Évry Cedex, France

[*]corresponding author: Jérôme Wojcik, Merck Serono, 9 chemin des Mines, 1211 Geneva, Switzerland, Phone: +41 22 706 91 61, Fax: +41 22 794 69 65

1. Biological definitions

This part is a glossary to introduce biological concepts used in the article.

DNA is a linear molecule whose duplication is possible. It is a sequence of 4 different "base pairs" (letters). Position on DNA is measured for each chromosome from one of its extremity in number of base pairs (1000 bp = 1 kbp, ex: *chr. 6, 20 003 345 bp*). Each

individual holds 2 versions of the same DNA (except for dissymmetric *XY* male chromosome pair, which is not studied here).

"Genes" are sequences of DNA (∼100 000 bp) that contain *(i)* templates for proteins *(ii)* information on when and where the protein should be produced. Only limits of the templates are well known. They are composed of sequences called "exons". There are about 21 000 known human genes.

"Genotyping" means reading a small sequence of DNA of an individual (∼10 bp). These sequence are called "markers". They are chosen so as *(i)* to be found only one time in all DNA *(ii)* to be variant between individuals ("polymorphic"). They can be thought as composed of a fixed flanking sequence immediately followed by a variant sequence. The different variants of a marker are called "alleles". Each individual holds 2 alleles (one on each DNA version), the combination being called "genotype".

A "Single Nucleotide Polymorphism" (SNP) is a marker with a variant sequence composed of one letter ("base"). In this article, among the 4 possible letters at a given position, only 2 are found in the studied populations (arbitrary noted alleles *a* and *A*, without any reference to the 4 letters). The corresponding genotypes are noted *aa*, *AA* and *Aa*. The first 2 are said to be "homozygotous", and the last one "heterozygotous".

A "complex disease" is a genetic disease (caused by variations of DNA) with complex heredity rules. The complexity comes from genetic and/or environmental interactions.

An "association study" aims at identifying spots on DNA ("loci") which are correlated with a characteristic of patients ("phenotype"). It is "genome-wide" when spots are researched across all DNA. Conversely, it is "candidate gene" based if spots are researched in the region of DNA containing a specific gene. In terms of machine learning, a patient is a sample with a label: its phenotype. The goal is to identify predictors inside the DNA.

"Haplotype blocks" are blocks of strongly correlated neighbor markers. An "haplotype" is a sequence of alleles held by the same version of DNA. For example, let consider 2 neighbor SNP *A* and *B*, and an individual with genotypes *aA* and *bB*. Either *a* and *b* (haplotypes *ab* and *AB*) or *a* and *B* (haplotypes *aB* and *Ab*) are held on the same DNA molecule. There is a correlation if one pair of haplotype is more frequent than the other in a population: Due their proximity: they are most of the time transmitted together from generation to generation. The correlation is therefore stronger between individuals or in populations with near common ancestors. This type of correlation is called "linkage disequilibrium" (LD). Thanks to LD, it is possible to sample the DNA of an individual at predefined markers, and thus to realize association studies without reading all DNA ("full sequencing"): If a variation of DNA causing the phenotype (a "functional mutation") occurs between samples, it is likely to be correlated with these sampled markers. The technology used in the article (Affymetrix GeneChip® human mapping 100K) allows to study around 100 000 SNP.

"Stratification" occurs in an association study if two populations different in terms of genotype frequencies for a set of markers are unknowingly mixed in a collection and if the "prevalence" of the disease (proportion of the population affected by the disease) is different in both populations. As a result, all this set of markers is found associated, whereas only few of them are biologically interesting.

2. Introduction

The last years have shown a tremendous increase in the number of markers available for genome-wide association studies, dealing either with the whole genome at a very low resolution (for instance 5 264 micro-satellites Dib et al. (1996)) or with a carefully chosen region of few mega base-pairs Cardon & Bell (2001); Lewis (2002). Recent technologies allow the genome-wide genotyping of hundred of thousands SNPs) Kennedy et al. (2003). This has arisen the need of new methodological developments to overcome different issues, such as the multiple-testing problem, gene biases, data quality analysis and the computational tractability.

First, the so-called multiple testing problem seems to cause association studies ability to detect associations to decrease as the number of markers increases. The classical genetic analysis strategy, based on an association test for each marker Klein et al. (2005), encounters increasing difficulties as the number of markers available will soon reach a few millions: Increasing the number of tests prevents from the detection of the mild genetic effects expected in complex diseases, as only strong effects emerges from the huge noise generated by the increased number of tests. Methods like False Discovery Rate (FDR) Benjamini & Hochberg

(1995) computation allow to control the error rigorously, but do not increase the statistical power. As an example of this loss of power, let assume a study with only associated marker (one true-positive) whose p-value is 10^{-5}. Testing it with 1 000 other markers allows to select it alone with a FDR threshold of $\approx 1\%$. With 1 000 000 markers, it is selected with 10 other non associated markers, with a subsequent FDR threshold of $\approx 91\%$. Better strategies based on haplotype blocks are being developed, the first step being gathering such block data (see the HapMap project, International_HapMap_Consortium (2005)). They will ultimately allow to test all linked markers together, which should increase detection power. As an illustration, continuing the preceding example, let assume that 2 markers in the same haplotype block are associated, both with the same 10^{-5} p-value and 1 000 000 markers. Separate testing leads to select 10 other markers with a FDR threshold of $\approx 83\%$. On the contrary, assuming that p-values are evaluated from a 1 degree of freedom χ^2 score, markers can simply be tested together using a 2 degree of freedom χ^2 test by adding their scores. The new p-value is $\approx 3.4\,10^{-9}$. The FDR threshold is $\approx 0.3\%$.

Secondly, a genetic association of a given SNP is a statistical feature and does not explain by itself a phenotype. To interprete biologically an associated marker, its haplotype block should first be delimited. Then, the association can be refined by fine-scale genotyping technologies or ideally by full resequencing. This eventually allows to identify functional mutations. Most of the time, these mutations impact relatively close genes. This is a first argument to bias this analysis towards genes. Moreover, even if haplotype blocks are unreachable, DNA might be cut into distinct regions (calls *bins*) on an other basis, so as to limit the multiple-testing problem and make it independent of the number of markers. Combining this two arguments leads to choose one bin for each gene, and to create 'desert' bins in large unannotated regions. It allows to associate a list of genes with a test, which simplifies the analysis of results. The drawbacks are *(i)* that it makes more difficult the study of these 'deserts', however the goal is here to maximize, not the chance of finding an association, but the chance of elucidating a mechanism of a complex disease given the current knowledge *(ii)* that a bin might contain several haplotype blocks, resulting in a dilution of the association signal if only one block is associated. Reciprocally, neighbor bins are not independent because they might share an haplotype block, but neighbor SNP were not independent either. The third argument is also congruent with this process: analysis of the results through genomic approaches based on pathway reconstruction (see Rajagopalan & Agarwal (2005) for example) require the use of biological knowledge, mainly structured by genes and corresponding proteins.

Thirdly, one must keep in mind that such genome-wide genotyping data are obtained by high-throughput experiment which encompass limitations requiring careful statistical methodology. Especially, with Affymetrix GeneChip® human mapping 100K, the trade-off between the call-rate (i.e. errors detected by the genotyping process and resulting in missing genotypes in the data set) and the error-rate (i.e. errors left in the data) is difficult to adjust. Obtaining unbiased statistical results is then conditioned to good pre-processing filters. Indeed spurious markers must be eliminated and missing data correctly managed.

In addition, polymorphism tends to be low for most of SNPs present of Affymetrix GeneChip® human mapping 100K Some genotypes are held by less than few percents of patients, which, given usual collection size of few hundreds, *(i)* is not enough for good asymptotic approximations and *(ii)* should be considered with care given possible high error rate.

Finally, whatever algorithmic solution is developed, because the number of markers available will probably quickly reach a few millions, creating a scalability problem, it has to be linear in the number of markers.

In this paper we present a novel Bayesian algorithm developed to practically analyze genome-wide association studies. This algorithm is based on a gene-based partitioning of DNA into regions, called bins.

A p-value of association is computed for each bin. The model takes into account genotyping errors and missing data and tries to detect simple differences in the haplotype block structure between cases and controls. The study of different collections is allowed. The multiple testing problem is addressed by estimation of FDR.

The method has been applied to analyze the results of three genome-wide case-control association studies of the complex disease Multiple Sclerosis (MS). It identifies putatively associated bins, containing genes previously described to be linked to MS (Dyment et al. (2004) for review) as well as new candidate genes.

3. Materials

Three association studies dealing with MS in three independent collections have been realized in collaboration with Serono Genetics Institute Cohen (2005). Around 600 patients have been recruited for each study, half of them as cases affected by the disease, half of them as controls (table 1). Controls have been matched with cases as best as possible so as to avoid stratification (see Thomas & Witte (2002) for example). For chromosome X, only female patients are studied. No SNP is studied on chromosome Y.

Genotypes of the 116 204 SNPs selected by Affymetrix have been determined for each patient using Affymetrix GeneChip® human mapping 100K technology. SNPs have been mapped on the NCBI build 35 of the Human genome.

4. Method

4.1. Notations

Stochastic variables are noted with a round letter ($\mathcal{V}$), a realization is noted in lower case (v). Indices are noted in lower case (k), ranging from 1 to the corresponding upper case letter (K). Unless needed, this range of indices ($k \in [1,K]$) is omitted. The number of different values is noted $\mathrm{card}\mathcal{V}$.

The n-dimensional table of the number of individuals having the same combination of values for given variables $\mathcal{V}^k, k \in [1,K]$ is noted $n(\mathcal{V}^1,\ldots,\mathcal{V}^K)$. It is the contingency table of these variables. The marginalization of such a contingency table over one variable, for example $\mathcal{V}^1$, is noted $n(\oplus,\mathcal{V}^2,\ldots,\mathcal{V}^K) = \sum_{v\in\mathrm{card}(\mathcal{V}^1)} n(v,\mathcal{V}^2,\ldots,\mathcal{V}^K)$.

Estimation of a probability distribution $P(\mathcal{V})$ is noted with hatted letter, $\widehat{P}(\mathcal{V})$.

Each bin $b \in [1,B]$ contains J_b genetic markers $\mathcal{G}_b^j$ with $j \in [1,J_b]$.

Each patient $i \in [1,I]$ has a phenotype value $s(i)$ (in case-control studies, $\mathrm{card}(\mathcal{S}) = 2$), discrete co-variable values $v_m(i), m \in [1,M]$ (gender: $m = 1$, or collection of origin: $m = 2$), and a genotype value for each marker $g_b^j(i)$ (with SNPs, $\mathrm{card}(\mathcal{G}_b^j) = 3$).

A patient i is therefore represented by the following vector:

$$i = \left[s(i), v_m(i), g_b^j(i)\right] \quad \text{with: } m \in [1,M], b \in [1,B], j \in [1,J_b] \tag{1}$$

The data set is noted $D = \{i\}_{i\in[1,I]}$

A first level of the method aggregates predictors at the bin level. The "restriction" of a patient to a given bin is noted i_b:

$$i_b = \left[s(i), v_m(i), g_b^j(i)\right] \quad \text{with: } m \in [1,M], j \in [1,J_b] \tag{2}$$

The corresponding data set is noted $D_b = \{i_b\}_{i\in[1,I]}$

4.2. Data preprocessing

Due to Affymetrix GeneChip® human mapping 100K technology (the D.M. calling algorithm), all errors are unevenly distributed: It is likelier to make an error on heterozygotous genotypes, because they lie "between" the two homozygotous genotypes in terms of pattern matching, i.e., the heterozygotous genotype pattern has a frontier with each of two homozygotous ones whereas each homozygotous genotype pattern has a frontier only with the heterozygotous one Liu et al. (2003); Rabbee & Speed (2006). It is indeed very unlikely to mistake one homozygotous genotype for the other one. This often results in gaps of heterozygotous genotypes. It can be detected through the evaluation of the probability that the distribution of genotypes observed in controls follows the Hardy-Weinberg equilibrium, which basically states that, under some hypothesis like random mating in the studied population, noting $P(a) = P(aa) + P(Aa)/2$ and $P(A) = P(AA) + P(Aa)/2$:

$$\begin{cases} P(aa) = P(a)^2 \\ P(Aa) = 2P(a)P(A) \\ P(AA) = P(A)^2 \end{cases} \tag{3}$$

Therefore, the following pre-processing filters are applied. SNPs are discarded *(i)* if the number of missing genotypes is higher than 5% because this is an indication that the genotyping process quality was low for this SNP, *(ii)* if the minimum allele frequency in controls MAF $= \min(P(a),P(A))$ is lower than 1%, because the SNP is not polymorphic in the studied population, would lead to bad quality probability estimation and would deter FDR estimations, or *(iii)* if the probability that the SNP follows the Hardy-Weinberg equilibrium in controls is lower than 0.02, because it is the indication of errors that the genotyping process did not self-detect. For chromosome X, only female patients are retained for pre-processing.

TABLE 1
GENOME-WIDE ASSOCIATION MULTIPLE SCLEROSIS COLLECTIONS.

Collection	Origin	#Cases	#Controls	%Females
A	French	314	352	69%
B	Swedish	279	301	71%
C	American	289	289	85%

Cases and controls are matched on age and gender. The risk of being affected is 1.7 times higher for females.

SNPs from chromosome Y have been discarded as well as SNPs with multiple localizations on the assembly *NCBI 35* of the human genome.

4.3. Bin definition

Bins are defined on DNA from protein genes as defined in the version 35.35 of EnsEMBL Birney et al. (2006) of the human DNA sequence. The basic region of a gene lie from the beginning of its first exon to the end of its last exon. Overlapping genes are clustered in the same bin. If two consecutive genes or clusters of overlapping genes are separated by less than 200 kbp, the bin limit is fixed in the middle of the interval. Otherwise, the limit of the upstream bin is set 50 kbp downstream its last exon, the limit of the downstream bin is set 50 kbp upstream its first exon, and a special bin corresponding to a *desert* is created in between the two bins. With these rules, desert bins have a minimum length of 100 kbp (figure 1).

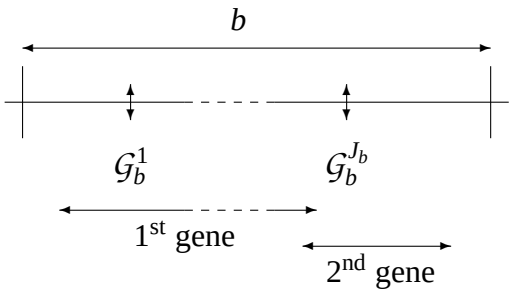

Fig. 1.— Representation of a bin containing two genes and J_b markers. *Bins are defined according to gene boundaries and intergenic distances.*

4.4. Assessing bin association

4.4.1. General model and hypotheses

We assume that each bin constitutes an independent data set. The following ideal probability distribution is defined:

$$\forall b \in [1,B], P(I_b) = P(\mathcal{S}, \mathcal{V}_m, \mathcal{G}_b^j) \tag{4}$$

As experimenters choose cases and controls (phenotypes) of patients in order to have about as many of each kind, as well as they choose for each case a control with the same co-variables, each individual subset of the study is a realization of the conditional distributions $P(\mathcal{G}_b^j|\mathcal{S}, \mathcal{V}_m)$.

Estimations of probability distribution are possible from contingency tables:

$$\widehat{P}(\mathcal{G}_b^j|\mathcal{S}, \mathcal{V}_m) = \frac{n(\mathcal{S}, \mathcal{V}_m, \mathcal{G}_b^j)}{n(\mathcal{S}, \mathcal{V}_m, \oplus)} \tag{5}$$

On the contrary, due to the experimental design, estimations $P(\mathcal{S}, \mathcal{V}_m)$ are impossible.

Finally, the hypothesis that each patient is independently chosen from others is made. In particular, they are supposed to be from different families. If this is false, for example if all cases are from the same family, all SNPs specific to this family will appear to be correlated with the disease.

4.4.2. Statistics

A general way to assess the association of a bin b is to estimate whether $(\mathcal{G}_b^j)_{j\in[1,J_b]}$ is independent from the phenotype $\mathcal{S}$, i.e., whether $P(\mathcal{G}_b^j|\mathcal{V}_m)$ is "far" from $P(\mathcal{G}_b^j|\mathcal{S}, \mathcal{V}_m)$.

$$H_0^b\colon P(\mathcal{G}_b^j|\mathcal{S}, \mathcal{V}_m) = P(\mathcal{G}_b^j|\mathcal{V}_m) \tag{6}$$

However, as only $P(\mathcal{G}_b^j|\mathcal{V}_m,\mathcal{S})$ is estimable, estimation of $P(\mathcal{G}_b^j|\mathcal{V}_m)$ is not possible. Therefore, one estimates $\widehat{P_{H_0^b}}(\mathcal{G}_b^j|\mathcal{V}_m)$ assuming that hypothesis H_0^b is true, as indicated by the subscript.

There are many different ways of computing the "distance" between estimations of $P(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m)$ and $P(\mathcal{G}_b^j|\mathcal{V}_m)$, depending on the parameters used to describe them and on what is understood under "far" from H_0^b. Usually, a score called a statistic which is large when $\widehat{P}(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m)$ is "far" from $\widehat{P}(\mathcal{G}_b^j|\mathcal{V}_m)$ is computed. In this paper, we have chosen likelihood ratio *LR* as a statistic. For each patient, the probability of observing his/her genotypes is computed according to 2 underlying distributions: the *true* underlying distribution $\widehat{P}(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m)$ or the distribution assuming that H_0^b is true $\widehat{P_{H_0^b}}(\mathcal{G}_b^j|\mathcal{V}_m)$. The ratio of these probabilities is the likelihood ratio of the patient:

$$\widehat{P}(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m) = \frac{n(\oplus,\mathcal{V}_m,\mathcal{G}_b^j)}{n(\oplus,\mathcal{V}_m,\oplus)} \tag{7}$$

$$\mathrm{LR}(i_b) = \frac{\widehat{p}(g_b^j(i)|s(i),v_m(i))}{\widehat{p_{H_0^b}}(g_b^j(i)|v_m(i))} \tag{8}$$

As all patients are considered to be independently chosen, the likelihood of the set of patients available is:

$$\mathrm{LR}(D_b) = \prod_{i\in[1,I]} \mathrm{LR}(i_b) \tag{9}$$

The likelihood ratio satisfies $\mathrm{LR}(D_b) \geq 1$. Indeed, $\widehat{P_{H_0^b}}(\mathcal{G}_b^j|\mathcal{V}_m)$ represents a model included in $\widehat{P}(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m)$.

4.4.3. Probability values

Due to randomness and finite sample size, errors are made on estimations of $P(\mathcal{G}_b^j|\mathcal{S},\mathcal{V}_m)$ and $P(\mathcal{G}_b^j|\mathcal{V}_m)$. Therefore, the probability of getting a likelihood ratio as high as or higher than the observed one assuming that H_0^b is true, i.e. the probability that H_0^b is true given the observation, i.e. the *p*-value π_b needs to be computed. This is theoretically achieved by enumerating all possible outcomes $D_b(\sigma)$ of the experiment that lead to the observed data $D_b(\sigma_0)$ (σ is a enumeration parameter to be defined. The following notation simplification is done: $D_b(\sigma_0) = D_b$). Then the probability $p(D_b(\sigma))$ of each outcome assuming that H_0^b is true is computed as well as its likelihood ratio. Finally, the *p*-value is obtained:

$$\begin{aligned}\pi_b &= p(\mathrm{LR}(D_b(\sigma)) \geq \mathrm{LR}(D_b)) \\ &= \sum_{\sigma|\mathrm{LR}(D_b(\sigma))\geq\mathrm{LR}(D_b)} p(D_b(\sigma))\end{aligned} \tag{10}$$

Ways of doing it are either exact (ex.: Fisher's exact *p*-values), asymptotic (ex.: Pearson χ^2, valid when I is large) or permutation based Agresti (2002). Here we have chosen the last one because it handles well complex models with missing values.

Let define the space of possible outcomes. Because of the experimental design which matches cases and controls (phenotypes), all possible outcomes are constrained to have the same contingency table reduced to phenotypes and co-variables $n(\mathcal{S},\mathcal{V}_m,\oplus)$. However, the number of each genotype for each marker is not fixed: It reflects the population genotype frequency, and thus only its mean over outcomes tends to this frequency. Therefore, the space of possible patients in this model is huge, because any combination of genotypes is possible for each patient without any constraint on the number of genotype count $n(\mathcal{G}_b^j)$. To simplify the model, the following hypothesis is made: the only possible patients are the one observed, except that their phenotype can be permuted (permutation of labels). It means that $n(\oplus,\mathcal{V}_m,\mathcal{G}_b^j)$ is unchanged in all possible outcomes. The parameter σ can be defined as a permutation of $[1,I]$ and σ_0 is the identity permutation. The probability of outcomes is uniform: $p(D_b(\sigma)) = 1/I!$.

Sampling the outcome space is possible: random permutations of the phenotypes are drawn and used to compute a likelihood ratio. This is a Monte-Carlo procedure, for which we propose an optimized implementation that guarantees the precision required for FDR estimation (described in appendix). Indeed FDR is very sensitive to bad *p*-value estimations Pounds (2006). Finally, due to this permutation structure, the denominator of equation (8) is constant with respect to the permutations realized, therefore *p*-values can be estimated with the numerator only, this is why the denominator is omitted in the article.

4.4.4. FDR

Finally, to address multiple testing, the method uses an FDR estimation developed originally by Benjamini

& Hochberg (1995) and defined as in Storey & Tibshirani (2003):

$$\mathrm{FDR}(\theta) = \frac{\widehat{\Pi_0}\theta B}{\mathrm{card}(\{b|\pi_b < \theta\})} \tag{11}$$

The numerator is an estimation of the expectation of the number of false-positive with $\pi_b \leq \theta$. In this numerator, $\widehat{\Pi_0}$ is an estimation of the proportion of bins under the null hypothesis. Given that it is expected to be very high in current study, it is (conservatively) fixed at its upper bound: $\widehat{\Pi_0} = 1$. The denominator is the number of tests with *p*-values below θ. The ratio is therefore an estimation of the proportion of false negatives in the set of bins with a *p*-value below θ. Because we want to analyze thoroughly the FDR for around the 10 bins with the lowest *p*-values, the FDR is not controlled at a specified threshold as in Benjamini & Hochberg (1995) but only estimated.

This estimation relies on two main hypothesis: *(i)* tests are independent or positively correlated Benjamini & Yekutieli (2001), *(ii)* *p*-values are continuously and uniformly distributed in $[0,1]$. Assuming that sharing of haplotype block by neighbor bins is the only source of correlation between tests, the positive correlation seems reasonable. Indeed, if the *p*-value of a not associated bin decreases, the *p*-values of bins sharing the same haplotype block are more than likely to decrease too. The uniform distribution is less obvious, because the number of possible contingency tables is finite so that even the null distribution is not uniform. However, the sample size is one to two order of magnitude higher than in other applications of FDR to discrete data in which the problem is acute Pounds & Cheng (2004).

4.4.5. Model of linkage disequilibrium

Due to linkage disequilibrium, neighbor markers may carry information about each other. If the correlation between a marker and its neighbor is very different between cases and controls, it means that haplotypes are different in cases and controls (even if there is no direct correlation between each marker and the phenotype), therefore the region surrounding the 2 markers might be associated with the disease. The model developed tries to implicitly find such differences in haplotypes as well as usual associations. It is an inhomogeneous hidden Markov chain that makes the approximation that the genotypes of two markers separated by a third one are independent conditional to the genotype of this third one. Indeed, as a rough approximation, for each marker, most information will be found on its first neighbor on each direction of DNA. In a directed graphical model, independence assumptions therefore consist in:

$$P(\mathcal{G}_b^j|(\mathcal{G}_b^l)_{l\neq j}) = \begin{cases} P\left(\mathcal{G}_b^j|\mathcal{G}_b^{j-1}\right) & \text{if } j \neq 1 \\ P\left(\mathcal{G}_b^j\right) & \text{if } j = 1 \end{cases} \tag{12}$$

The probability distributions modelled are the same as the one obtained in the reverse order on j. Indeed, as no node has strictly more than one incoming edge, the distributions modelled are the same in the two configurations (Naïm et al. (2004)).

Finally, this assumptions also allow to obtain correct estimations because corresponding contingency tables are sufficiently filled. More precisely, if the minimum cell count $\min(n(\mathcal{V}^k))$ is low, random effects are high because a change of one in the count will lead to a great change in the probability estimation. Therefore contingency tables should not be computed for too many variables at a time. The assumptions implies that contingency tables will be computed for 2 SNPs ($\mathrm{card}(\mathcal{G}_b^j) = 3$), the phenotype ($\mathrm{card}(\mathcal{S}) = 2$) and the co-variables together. The gender co-variable will not be used. It requires the hypothesis that the SNP distribution is independent from it. The only co-variable is the study patients belong to (cf. table 1, $\mathrm{card}(\mathcal{V}_2) = 3$). As collection sizes for a given study are around 600, the average number of patients in each cell of contingency tables is then $\bar{n} = 33$. However the minimum cell count is much less in real data because low minimum allele frequencies and linkage disequilibrium might result in very rare combinations of genotypes.

4.4.6. Model of error

An error model is introduced with observed genotypes $\mathcal{O}_b^j$ (with SNPs, $\mathcal{O}_b^j \in \{aa, Aa, AA, \emptyset\}$, where $\emptyset$ means that the genotype is missing):

$$P(\mathcal{O}_b^j|(\mathcal{G}_b^l)_{l\in[1,J_b]}) \quad = \quad P(\mathcal{O}_b^j|\mathcal{G}_b^j) \tag{13}$$

Indeed, the technology is the same for all determinations of the same marker and there is no correlation between the genomic order of SNPs and the localization of their probes on the genotyping chips, so that

there is no reason why the observed genotype should depend on something else than the real genotype. The model is represented graphically in fig. 2.

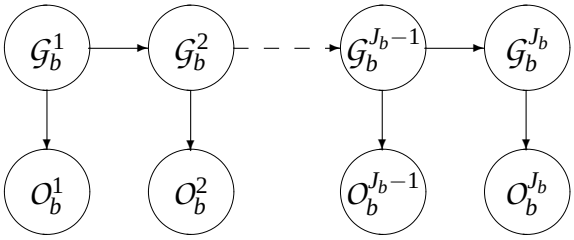

Fig. 2.— Error and linkage disequilibrium model of bin k. *The LD is modelled by dependencies between genotypes $\mathcal{G}_b^j$ and the genotyping errors by the relationships between observed genotypes O_b^j and real genotypes $\mathcal{G}_b^j$. The model is an inhomogeneous hidden Markov chain.*

Since $\mathcal{G}_b^j$ are hidden variables, estimation of a priori probabilities of $P(\mathcal{G}_b^j|\mathcal{G}_b^{j-1})$ and $P(O_b^j|\mathcal{G}_b^j)$ is not straightforward. Usual strategy is to use an Expectation-Maximization (E.-M.) algorithm to infer the state of hidden variables. However, it is not required in order to assess bin associations, and, if used, the likelihood of the model is approximated to the one of the most probable configuration of hidden variables, which is a strong approximation.

Therefore, an alternative strategy is developed. $P(\mathcal{G}_b^j|\mathcal{G}_b^{j-1})$ and $P(\mathcal{G}_b^1)$ are estimated through the removal of patients with missing genotypes:

$$\widehat{P}(\mathcal{G}_b^j|\mathcal{G}_b^{j-1}) = \frac{n(O_b^j, O_b^{j-1}) + C}{n(\oplus, O_b^{j-1}) - n_\emptyset + mC} \tag{14}$$

Where $n_\emptyset$ is the number of patients with either O_b^j or O_b^{j-1} missing and and m is the number of cells. To obtain more regular estimates, a constant is added to all cell counts. It is a Dirichlet prior on parameters (cf. Naïm et al. (2004) for example). This constant is chosen to be $C = \alpha\bar{n}$, where α is the chosen error rate and $\bar{n}$ is the mean number of individuals per cell. This constant means that uncertainty on low cell counts is high, not only because of randomness, but also because of genotyping errors.

On the other hand, given the previously developed structure of errors, the following model of $P(O_b^j|\mathcal{G}_b^j)$ is chosen:

$P(O_b^j|\mathcal{G}_b^j) =$

$O_b^j \backslash \mathcal{G}_b^j$	aa	Aa	AA
aa	$1-\beta$ $\times 1-\alpha$	$1-2\beta$ $\times\alpha$	0
Aa	$1-\beta$ $\times\alpha$	$1-2\beta$ $\times 1-2\alpha$	$1-\beta$ $\times\alpha$
AA	0	$1-2\beta$ $\times\alpha$	$1-\beta$ $\times 1-\alpha$
$\emptyset$	β	2β	β

(15)

The error rate α is estimated during external comparison of Affymetrix GeneChip® human mapping 100K and other technologies. In this study, the error rate is chosen to be $\alpha = 0.05$ (cf. Affymetrix GeneChip® human mapping 100K assay manual p115). The missing rate β is estimated for each marker through the resolution of this non-linear system:

$$\left\{\begin{array}{rl} P(O_b^j = Aa) & = \sum_{g_b^j \in \{aa,Aa,AA\}} P(Aa|g_b^j) P(g_b^j) \\ P(O_b^j = \emptyset) & = \sum_{g_b^j \in \{aa,Aa,AA\}} P(\emptyset|g_b^j) P(g_b^j) \\ 1 & = \sum_{g_b^j \in \{aa,Aa,AA\}} P(g_b^j) \end{array}\right. \tag{16}$$

The system sums up in these equations in which $\beta \in [0, 1/2]$ and $z = P(\mathcal{G}_b^j = Aa) \in [0,1]$ are unknown:

$$\left\{\begin{array}{l} P(O_b^j = Aa) = f(\alpha, \beta, z) \\ \text{with: } f = (1-\beta)\alpha(1-z) + (1-2\beta)(1-2\alpha)z \\ P(O_b^j = \emptyset) = \beta(1+z) \end{array}\right. \tag{17}$$

However the system has no acceptable solution when $\min(f) \geq P(O_b^j = Aa)$. With $\alpha = 0.05$, this happens for SNPs with low diversity. In this case, The a-priori error rate is decreased: Given partial derivatives $\partial f/\partial\beta$ and $\partial f/\partial z$, f decreases when β increases and increases with z under the condition that $\beta < 1 - 3\alpha/2 - 5\alpha$, which is always satisfied because $P(O_b^j = \emptyset) \leq 0.05$ (due to the pre-processing) thus $\beta \leq P(O_b^j = \emptyset) \leq 0.05$. Moreover, the second relation implies that β decreases when z increases. As a result, $\min(f)$ is reached when β is maximum, i.e. for $z = 0$ and $\beta = P(O_b^j = \emptyset)$. When it happens, the error rate is decreased to the maximum possible error-rate $\alpha_{\max}$, obtained for the unrealistic value of $P(O_b^j = Aa) = 0$ (all heterozygotous observed are errors):

$$\alpha_{\max} = \frac{P(O_b^j = Aa)}{1 - P(O_b^j = \emptyset)} \tag{18}$$

If a SNP satisfies the Hardy-Weinberg equilibrium, has no missing values and nonetheless do not satisfy $\alpha_{\max} \geq 5\%$, its minimum allele frequency is lower that approximatively 2.5%, which shows that this problem occurs only for SNP of low diversity.

4.4.7. Likelihood computation

If real genotypes $g_b^j(i)$ were known with certainty (or if they were inferred through an E.-M. algorithm, see Naïm et al. (2004)), the likelihood would be:

$$\begin{aligned} L_0(i_b) &= p\left(g_b^j(i), o_b^j(i)\right) \\ &= \prod_{j>1} p\left(o_b^j(i)|g_b^j(i)\right) p\left(g_b^j(i)|g_b^{j-1}(i)\right) \\ &\quad \times p\left(o_i^1(i)|g_i^1(i)\right) p\left(g_i^1(i)\right) \end{aligned} \tag{19}$$

However, they are not, so one has to compute this likelihood $L_0(i_b)$ for each combination of hidden variables. The likelihood of a patient is therefore:

$$\begin{aligned} &L_1(i_b) = p\left(o_b^j(i)\right) = \\ &\sum_{g_b^j \in [1,\mathrm{card}(\mathcal{G}_b^j)]} \left(\begin{array}{l} \prod_{j>1} p\left(o_b^j(i)|g_b^j\right) p\left(g_b^j(i)|g_b^{j-1}(i)\right) \\ \times p\left(o_b^1(i)|g_b^1\right) p(g_b^1) \end{array} \right) \end{aligned} \tag{20}$$

This a computation in $O\left(\prod \mathrm{card}(\mathcal{G}_b^j)\right) \sim O\left(3^{J_b}\right)$. Some approximations in the model are required to obtain tractable likelihood computations (i.e. linear with the number of markers). The following one is based on two-marker sliding windows and corresponds to the model of fig. 3:

$$\begin{aligned} &L_2(i_b) = \\ &\prod_{j\geq 2} \sum_{g_b^{j-1}, g_b^j} \left(\begin{array}{l} p(o_b^j(i)|g_b^j) \\ \times p(g_b^b j, g_b^{j-1}) \\ \times p(o_b^{j-1}(i)|g_b^{j-1}) \end{array} \right) \end{aligned} \tag{21}$$

This equation considers information coming from two neighbor markers together. Compared to the full model, information flow is limited to pair of markers. The likelihood could be falsely increased in this extreme situation: suppose that a missing genotype is inferred *aa* from its left neighbor and *AA* from its right neighbor, the merging of this two inferences would results in a contradiction and thus a low resulting likelihood. On the contrary, the approximated likelihood will not detect this contradiction and will be falsely increased. This likelihood is named thereafter "two-marker" likelihood.

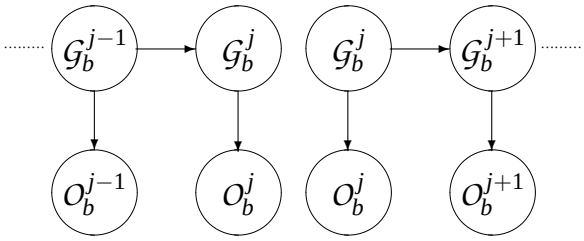

Fig. 3.— Simplified model of two-marker likelihood computation. *It is based on a two-marker sliding window. Compared to the ideal model shown in figure 2, information flow is restricted to couples of neighbor genotypes* $\mathcal{G}_b^j$.

Simplifying further leads to consider markers one by one. There is no model of linkage disequilibrium anymore, but noise is reduced as cells are better filled. This likelihood is named thereafter "naive likelihood" because it corresponds to a naive Bayesian model:

$$L_3(i_b) = \prod_j \sum_{g_b^j} p(o_b^j(i)|g_b^j) p(g_b^j) \tag{22}$$

5. Results

The method has been applied to each of the three collections *A*, *B*, *C* (table 1) as well as to the three collections at once ($A+B+C$), considering the collection of origin as a co-variable. The overall computation time is about 10 days on a single Altix Itanium processor but can be easily parallelized (for instance chromosome by chromosome) to drop down to less than 24h.

The pre-processing filters discard around 20% of SNP: for collection *A*, out of 112 463 SNP, the missing data filter discards 14 407 SNP, the Hardy-Weinberg filter 9 422 SNP, and the MAF filter 12 662. As some SNP are removed by several filters, 84 430 SNP remain. For collection *B* and *C*, respectively 93 548 and 86 652 SNP remain. If all SNP satisfied the Hardy-Weinberg equilibrium, 2 249 SNP are expected to be discarded. Four times as many were. It can be explained *(i)* by artifacts of DM calling algorithm which has a higher error rate on heterozygotous genotypes *(ii)* by deviations from the assumptions underlying this theoretical equilibrium.

The bin partioning algorithm divides the genome into 19 556 gene bins and 1 993 desert bins. Out of these 21 549 bins, only 11 264 (52%) contain one SNP or more after pre-processing in at least one MS collections and are considered for further analyses. To be collection independent, the distribution of the number of SNP per bin is studied before pre-processing. It is heavy tailed: out of 12 512 SNP with one bin or more, 2 781 have only one SNP, and 2 188 bins have 10 SNP or more. The maximum is 210.

Figure 4 shows the FDR plotted against *p*-values computed using the two-marker L_2 or the naive L_3 likelihood for the three collections. With the classical type I error rate threshold of 5%, the FDR is estimated to be 88% for 679 bins and 84% for 713 bins with the naive and two-marker likelihood respectively. Two-marker FDR remains below naive FDR until a *p*-value level of 0.01 and both increase slowly towards 1. This slow increase of FDR may reflect an association noise, like stratification, which results in a large number of lowly associated bins. It seems worse for two-marker FDR because it uses more information: this is quantified by the estimation of the number of true positives, for which a lower bound is $\max_{\theta\in[0,1]}(\mathrm{card}(\{b|\pi_b<\theta\})-\theta B)$. This estimation is 427 for the two-marker likelihood and 275 for the naive one.

FDR against the number of selected SNP plots are detailed by collection in figure 5. As observed in other studies Pounds (2006), the FDR is not monotonous with the *p*-value. The oscillations are less important for the three collection design, maybe because of the three time increase of sample size. With a FDR threshold of 5%, only between 2 and 6 bins are selected depending on the collections and likelihood considered (table 2). Most of them are located on chromosome 6, in the Major Histocompatibility Complex (MHC) region, mainly in the class III subregion. The class II subregion is known to be associated with MS Horton et al. (2004). *(i)* the number of collections studied simultaneously (one or three) and *(ii)* the likelihood type (L_2 or L_3) impact the results. First, the three collection design selects more associated bins than one collection designs, independently on the likelihood. Only one bin (*chr. 6, 32 499 261-32 528 188 bp*, containing the HLA-DRA gene) is found replicated independently on the collection or likelihood considered. Second, let compare likelihoods only for the three collection design. One bin is selected exclusively by the naive likelihood L_3 (*chr. 6, 33 036 631-33 069 204 bp*).

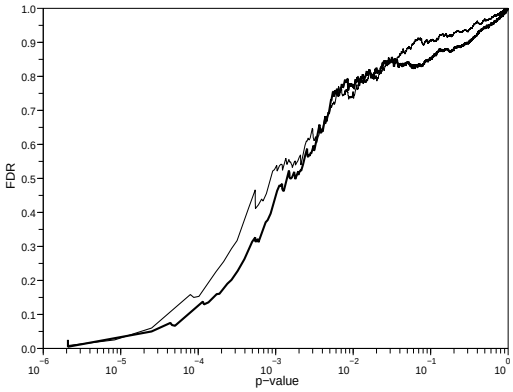

Fig. 4.— FDR versus *p*-values. *The FDR is computed for the three collections $A+B+C$ using the two-marker L_2 (thick line) or the naive L_3 (thin line). FDR is plotted against the p-value of each bin sorted by increasing p-value. The logarithmic scale emphasizes the rapid increase of FDR for low p-values (<0.01) and shows the slow tendency towards 1 for high p-values. FDR is slightly lower for L_2.*

This bin is ranked 9$^{\text{th}}$ using the two-marker likelihood, corresponding to a FDR of 6.6%. On the contrary, two bins are selected exclusively by the two-marker likelihood (*chr. 11, 48 432 965-48 517 820 bp* and *chr. 6, 32 827 002-32 863 903 bp*) and are ranked 5$^{\text{th}}$ and 11 649$^{\text{th}}$ respectively using the naive likelihood. The former could have been selected by the naive likelihood with a slightly less stringent FDR threshold (6%) but the latter can only be captured by the two-marker likelihood. Moreover, as it is not selected in one collection designs, it suggests the ability of the three-collection design to accumulate small evidences. Taken together, these features tend to illustrate greater power *(i)* of the three collection design over the one collection design and *(ii)* of L_2 over L_3 in the three collection design. These conclusions are strengthened by results with a less stringent FDR threshold of 50% (table 3).

But FDR is misleading in this study because the MHC region is known to be associated with MS (Dyment et al. (2004) for review). It leads to an overestimation of the FDR at which bins outside of this region are selected. Indeed, the fact that true positives are known reduces the FDR of the rest of set of selected

TABLE 2

ASSOCIATED BINS AT FDR 5% THRESHOLD.

Chr.	Start	End	$L_3(A)$	$L_3(B)$	$L_3(C)$	$L_3(ABC)$	$L_2(A)$	$L_2(B)$	$L_2(C)$	$L_2(ABC)$	Genes[a]
1	244 919 282	244 952 903		1		1		1		1	OR2T2
2	213 646 850	213 908 230						1			ZNFN1A2
6	32 269 737	32 334 279	1	1	1	1	1	1		1	NOTCH4 and NM_022107.1
6	32 334 280	32 459 062						1	1	1	ENSG00000161877
6	32 499 261	32 528 188	1	1	1	1	1	1	1	1	HLA-DRA
6	32 827 002	32 863 903								1	HLA-DQB2
6	33 036 631	33 069 204				1					HLA-DMA and BRD2
8	67 691 180	67 742 206	1								VCPIP1
11	48 432 965	48 517 820								1	OR4A47
			3	3	2	4	2	6	2	6	

A, *B*, *C* refer to the collections studied independently, *ABC* to the simultaneous study of the 3 collections. L_2 is the two-marker model likelihood model and L_3 the naive one. Bins on *Chr. 6* are assumed to be true positives given previous studies. HLA-DRA is always detected. Selection of OR4A47 with $L_2(ABC)$ illustrates that the three collection designs accumulate small evidence from each collection.

[a]Due to linkage disequilibrium, association of a bin may be caused by functional mutations in adjacent bins, thus it cannot be inferred directly that genes of the bins are linked with the phenotype.

TABLE 3

NUMBER OF ASSOCIATED BINS AT FDR 50% THRESHOLD.

Collection(s)	L_3	L_2	Common
A	6	6	4
B	14	7	5
C	6	28	6
$A+B+C$	20	33	17

The power is greater with the three population design. The L_2 likelihood is more powerful than L_3 for the three collection design.

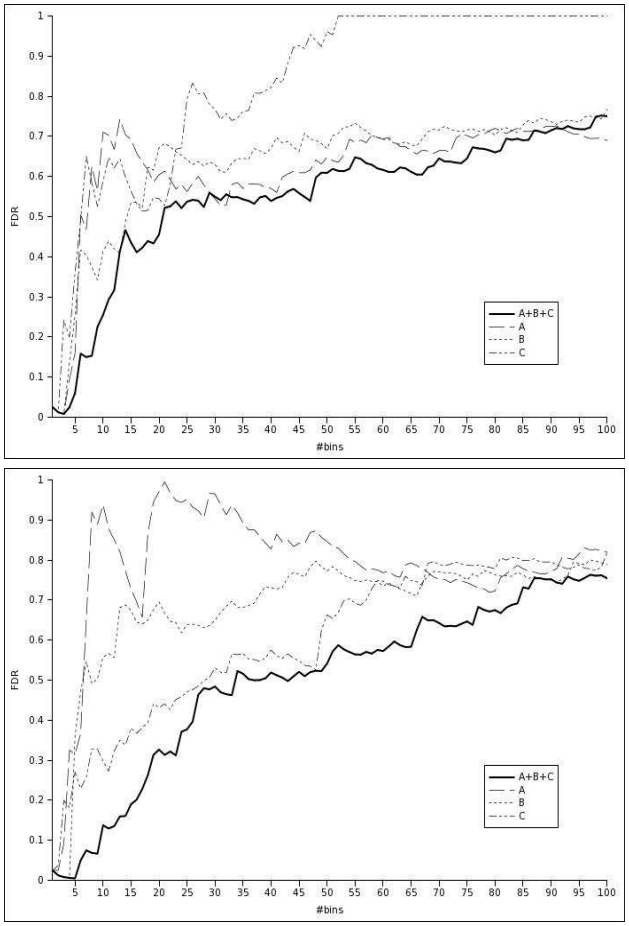

Fig. 5.— FDR using L_3 naive likelihood (top) and L_2 two-marker likelihood (bottom). *The FDR is plotted against the rank of the 100 best bins sorted by increasing p-value, for the three collections independently (A: solid, B: dash, C: dash dot) as well as for the three collections studied together ($A+B+C$: thick). FDR oscillates above* 1 *in the C collection for* L_3 *and has been maxed to* 1*. In both cases, the FDR is lower for the three collection design. It is also more regular.*

bins. If tests corresponding to the 106 bins of this region of 3.6 Mbp are removed from analysis (*chr. 6, 29 720 822-33 276 526 bp*, classical class I, II and III MHC subregions), 12 bins out of the 33 selected by L_2 on the three collection design are also removed (table 3). Assuming that these 12 bins are truly associated, the corrected FDR of the 21 remaining bins is 69%. Results at 50% FDR threshold without the MHC region presented in table 4: only 10 bins are selected. Again, the bins selected with the naive likelihood in $A+B+C$ are either selected or not far from being selected with the two-marker likelihood (the two-marker likelihood ranked 99^{th} or less all the bins selected by the naive likelihood).

Finally, a one test per SNP method was also tested[1]. It is based on exact Fisher *p*-values and FDR estimation following Storey & Tibshirani (2003). Due to exact *p*-value computation limits and genetic heterogeneity (in the same bin, different SNP are expected to be associated in different collections), collections are studied independently. A bin is selected if it contains one bin selected at the chosen threshold. Consequently, FDR is underestimated at the bin level (For example, let assume that 10 SNP are selected, 9 in one bin and 1 in another bin. With a FDR level of 10%, due to LD, the SNP in the 1 SNP bin is likely to be the false positive. Therefore, the 9 SNP bin is a true positive and the 1 SNP bin a false-positive. The FDR on bins is 50%). The comparison of both methods at a threshold of 5% is in table 5. For one collection designs, the number of bins selected is similar, but all bins selected by the one test per SNP method are in the MHC region.

6. Discussion

We have developed a new method to practically analyze data coming from genome-wide association studies. Our algorithm is based on a bin partitionning of the genome, takes advantage of studying several collections simultaneously and takes into account genotyping errors and local genomic structure (LD) while staying computationally tractable. The method has been applied to analyze around 216 million genotypes from three association studies in Multiple Sclerosis, leading to the discovery of novel potentially associated genes with FDR estimation.

The underlying model relies on definition of ge-

[1] Unpublished methods and results.

TABLE 4

NUMBER OF ASSOCIATED BINS AT FDR 50% AFTER EXCLUSION OF MHC REGION BINS.

Collection(s)	L_3	L_2	Common
A	2	0	0
B	1	1	1
C	0	0	0
$A+B+C$	8	10	7

MHC region bins are assumed to be true-positives. One collection designs select very few bins.

TABLE 5

NUMBER OF ASSOCIATED BINS AT FDR 5% THRESHOLD.

Collection(s)	L_3	L_2	one test per SNP
A	3	2	6
B	3	6	2
C	2	2	3
$A+B+C$	4	6	N/A

For the one test per SNP method, a bin is selected if it contains at least one selected SNP, leading to an underestimation of FDR.

nomic regions called the bins. This bin definition is purposefully gene-based and has some drawbacks. Firstly, if a new gene is discovered in a desert bin, one has to redefine this bin as well as its neighbors. Secondly, a frontier might fall inside an haplotype block, therefore neighbor bins might not be independent. An improvement would be to refine bin limits using haplotype blocks as soon as a consensus will be found on them. But, for several collection studies, bins must be identical for the all collections and it is still difficult to define haplotype blocks across collections from different populations International_HapMap_Consortium (2005).

The algorithm assigns p-values to bins in order to assess the significance of association which is supposed to be linked by design to the studied disease. However, other mechanisms like stratification could also create spurious association of bins. In our studies, the collections have been checked for stratification using a set of 200 unrelated markers Pritchard & Rosenberg (1999) (data not shown). Our method may also be used to a posteriori detect such stratification through the slope of the FDR against the p-value for medium to high p-value (> 0.01): Assuming no computation artifacts, a slow increase reflects a large number of mild associations.

We have chosen to take into account genotyping errors occuring more often for heterozygotous genotypes and chromosomal structure by using a Bayesian model. In our model, we can integrate LD (L_2 two-marker likelihood) or ignore SNP dependencies (L_3 naive model). L_2 is better on the three collection design whereas it is unclear for one collection designs. Indeed, as two neighbor markers are studied together with L_2, the sample size must be higher. This restriction to neighbor markers is not optimal, especially if one of the marker have a low diversity of genotypes. In this case, considering non neighbor pairs may be more efficient because low-diversity markers contain very few information on associations. Instead of a linear chain, the dependence between SNPs could be modeled as a DAG (Direct Acyclic Graph). This DAG could be inferred from the Minimum Spanning Tree obtained using a measure of the LD of all pairs of SNP in a region as a distance.

Despite its biases, the bin approach has the significant advantage to reduce dramatically the number of tests performed, here from 116 204 (number of SNPs in Affymetrix GeneChip® human mapping 100K) to 11 264 (number of bins). It minimizes the multiple-testing problem to the expense of diluting localized associations in large bins. It happens when a bin contains

many haplotype blocks, among which only one is associated. As a result of this trade-off, in one collection designs, the number of bins selected is similar with our method and with a one test per bin method. However, any increase in the number of SNP tested will balance results in favor of bin approaches. Our results clearly show that it is still required to control *p*-value threshold by FDR: FDR is over 80% with the classical 5% type I error rate threshold, see figure 4. Here again, having the haplotype block structure would be helpful to build more efficiently the bins.

The FDR threshold is chosen according to the desired application. To conduct expensive downstream experiments with putatively associated genes, a very low rate of false-positives is required, with the risk of missing true positives. A FDR threshold of 5% seems reasonable. On the contrary, if one wants to minimize the false-negative rate, a FDR of 50% or more is acceptable but the comparison with other sets of independent data (such as bibliography or expression data) is necessary to drag out signal out of noise.

Applying the method to experimental genome-wide association data on three collections permits *(i)* to assess the algorithm and evaluate the different parameters and design and *(ii)* to identify genes potentially associated to Multiple Sclerosis.

We have evidenced that the three collection design outperforms the one-study design in terms of expected number of true-positives, despite differences between the studied collections, especially on the severity of the disease. Indeed, using the collection as a co-variable in a bin approach (contrary to a one test per SNP approach) does not hypothesize that a specific marker of the bin should be found associated in all studies (or even a specific allele like in the Mantel-Haenszel test) but it still assumes that the phenotype is identical over all collections. Furthermore, with this three collection design, the two-marker likelihood L_2 seems to be more efficient thanks to the additional information used.

With this configuration, a FDR threshold of 5% gives 6 associated bins. Four of them are located in the MHC region on chromosome 6 already known to be linked to Multiple Sclerosis Dyment et al. (2004). It is a validation of the method. The two others are bins containing olfactory receptor genes *OR2T2* and *OR4A47*. The biological meaning of such association is unclear but the extended MHC regions contain many other olfactory genes Horton et al. (2004) and olfactory dysfunction has already been reported in Multiple Sclerosis Doty et al. (1998); Zivadinov et al. (1999). At FDR threshold of 50% and after exclusion of bins from MHC, the method selects ten bins. Excluding the ones containing *OR2T2* and *OR4A47*, most of them are desert bins or contain uncharacterized genes. One of them is located on chromosome X despite the smallest number of individuals in collections (only females). These novel potentially associated bins open the perspective of new gene discovery or new gene functionalization to explain Multiple Sclerosis . They are good targets for genotyping validation in a candidate gene approach.

Acknowledgments

We are grateful to the Serono Genetics Institute banking, genotyping and genetic analysis team for producing high-quality data. The work has been significantly made easier by the Serono Genetics Institute Research Knowledge Management team and we acknowledge them particularly. We also thank Pierre-Yves Bourguignon for the idea of the hidden Markov chain and Jean Duchon for the distribution of the distance between two *p*-value. This article has also been greatly improved thanks to the comments of many reviewers.

REFERENCES

Agresti, A. 2002, Categorical Data Analysis, 2nd Edition, Wiley Series in Probability and Mathematical Statistics (New-York: Wiley)

Benjamini, Y., & Hochberg, Y. 1995, J. Roy. Statist. Soc. Ser. B, 57, 289

Benjamini, Y., & Yekutieli, D. 2001, Annals of Statistics, 29, 1165

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Gräf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kähäri, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Hubbard, T. J. P. 2006, Nucleic Acids Res, 34, D556

Cardon, L. R., & Bell, J. I. 2001, Nat Rev Genet, 2, 91

Cohen, D. 2005, Serono Identifies 80 Genes Involved in Multiple Sclerosis Using 100000 SNPs, Interview in Affymetrix microarray bulletin

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., & Weissenbach, J. 1996, Nature, 380, 152

Doty, R. L., Li, C., Mannon, L. J., & Yousem, D. M. 1998, Ann N Y Acad Sci, 855, 781

Dyment, D. A., Ebers, G. C., & Sadovnick, A. D. 2004, Lancet Neurol, 3, 104

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush, M. J., Povey, S., Talbot, C. C., Wright, M. W., Wain, H. M., Trowsdale, J., Ziegler, A., & Beck, S. 2004, Nat Rev Genet, 5, 889

International_HapMap_Consortium. 2005, Nature, 437, 1299

Kennedy, G. C., Matsuzaki, H., Dong, S., min Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M. S., Boyce-Jacino, M. T., Fodor, S. P. A., & Jones, K. W. 2003, Nat Biotechnol, 21, 1233

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. 2005, Science, 308, 385

Lewis, C. M. 2002, Brief Bioinform, 3, 146

Liu, W., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T. B., Webster, T. A., Dong, S., Liu, G., Jones, K. W., Kennedy, G. C., & Kulp, D. 2003, Bioinformatics, 19, 2397

Naïm, P., Wuillemin, P.-H., Leray, P., Pourret, O., & Becker, A. 2004, Réseaux bayésiens (Paris: Eyrolles)

Pounds, S., & Cheng, C. 2004, Bioinformatics, 20, 1737

Pounds, S. B. 2006, Brief Bioinform, 7, 25

Pritchard, J. K., & Rosenberg, N. A. 1999, Am J Hum Genet, 65, 220

Rabbee, N., & Speed, T. P. 2006, Bioinformatics, 22, 7

Rajagopalan, D., & Agarwal, P. 2005, Bioinformatics, 21, 788

Storey, J. D., & Tibshirani, R. 2003, Proc Natl Acad Sci U S A, 100, 9440

Thomas, D. C., & Witte, J. S. 2002, Cancer Epidemiol Biomarkers Prev, 11, 505

Zivadinov, R., Zorzon, M., Bragadin, L. M., Pagliaro, G., & Cazzato, G. 1999, J Neurol Sci, 168, 127

This 2-column preprint was prepared with the AAS LaTeX macros v5.2.

A. Optimization of the Monte-Carlo procedure

The goal of the procedure is to sort B bins by empirical p-value estimation $\hat{\pi_b}$ so as to select only the most likely associated bins for further analysis. The questions are: What is the number of permutations needed sort confidently the bins? How to use most of the computing power on estimations of lower p-values ($\pi_b \leq \theta$) which are likelier to be true positives?

Firstly, for a given bin b, the number of permutations N_b^+ leading to likelihood ratio higher than the observed one out of the N_b permutations realized leads to the following estimator of π_b: $\hat{\pi_b} = N_b^+/N_b$. Indeed, N_b^+ follows a binomial law of parameter the p-value π_b which is asymptotically (in N_b) normal of mean $N_b\pi_b$ and variance $\sqrt{N_b\pi_b(1-\pi_b)}$. For instance, a 90% confidence lower bound can be estimated as $N_b\pi_b - \gamma\sqrt{N_b\pi_b(1-\pi_b)}$ with $\gamma = 1.3$. Moreover, as the difference of two consecutive p-values π_{b1} and π_{b2} is a linear combination of two independent normal distributions, the corresponding N_b^+ difference is normally distributed of mean $N_b(\pi_{b2} - \pi_{b1})$ and of variance $\sqrt{N_b(\pi_{b1}(1-\pi_{b1}) + \pi_{b2}(1-\pi_{b2}))}$ which is approximatively equal to $\sqrt{2N_b\pi_b(1-\pi_b)}$ if $\pi_{b1} \approx \pi_{b2} \approx \pi_b$. Assuming that $\pi_{b1} \leq \pi_{b2}$, sorting consecutive p-values requires that the lower bound of the difference is positive, *i.e.*:

$$\gamma\sqrt{2N_b\pi_b(1-\pi_b)} \leq N_b(\pi_{b2} - \pi_{b1}) \tag{A1}$$

Secondly, if no bin is associated, the asymptotic (in sample size I) distribution of a p-value is assumed to be uniform between 0 and 1. Consequently, the difference between two consecutive p-values is asymptotically (in B) characterized by a Poisson process of intensity B. Therefore, this difference follows an exponential law of density Be^{-Bx}, so $\pi_{b2} - \pi_{b1} \geq \delta/B$ with probability $e^{-\delta}$. Combined with equation (A1), we obtain that the probability to be able to order with a confidence level γ two consecutive p-values approximately equal to π_b is $e^{-\delta}$ if $\gamma\sqrt{2N_b\pi_b(1-\pi_b)} \leq \delta N_b/B \leq N_b(\pi_{b2} - \pi_{b1})$, i.e. if N_b is high enough such that:

$$N_b \geq \left(\frac{\sqrt{2}\gamma}{\delta}B\right)^2 \pi_b(1-\pi_b) \tag{A2}$$

The right term of this inequality is increasing for $\pi_b \in [0, 1/2]$. Therefore, if an upper bound of the threshold that will be used to select p-values is $\theta \leq 1/2$ and $\pi_b \leq \theta$, it is enough that N_b satisfy:

$$\pi_b \leq \theta \Rightarrow N_b \geq \left(\frac{\sqrt{2}\gamma}{\delta}B\right)^2 \theta(1-\theta) \tag{A3}$$

However, it is useless to spent computation time to estimate p-values above θ with so many permutations, because they almost surely concern not associated bins. Only a lower bound of such p-values is required. For coherence, the confidence on the lower bound is chosen to be γ as previously. Given a tolerance β and a bin b, it is enough to have a relative confidence interval of $\beta\pi_b$ that tend to 0 for lower p-values: $\beta N_b\pi_b \geq \gamma\sqrt{N_b\pi_b(1-\pi_b)}$. It yields a new bound on the number of permutations needed to compute p-values$\pi_b \geq \theta$:

$$\pi_b \geq \theta \Rightarrow N_b \geq \left(\frac{\gamma}{\beta}\right)^2 \frac{1-\pi_b}{\pi_b} \tag{A4}$$

To ensure coherence of inequalities (A3) and (A4) for $\pi_b = \theta$, β must be defined as:

$$\beta = \frac{\delta}{\sqrt{2}B\theta} \tag{A5}$$

Given that the right term of inequality (A4) decreases with p-value, the 2 inequalities sum up in:

$$N_b \geq \left(B\theta\frac{\sqrt{2}\gamma}{\delta}\right)^2 \min\left(\frac{1-\theta}{\theta}, \frac{1-\pi_b}{\pi_b}\right) \tag{A6}$$

This inequality draw the lines of a Monte-Carlo procedure: for each bin b compute likelihood ratios for new permutations of phenotypes (labels) until the number of permutations realized N_b satisfies it, replacing π_b by its estimation N_b^+/N_b. The inequality also shows that 2 parameters control the quality of the method: θ which is an upper bound of the threshold that will be used and $\sqrt{2}\gamma/\delta$ which controls the error due to the randomness of the process. The right term of the inequality is quadratic with the number of tests B, illustrating a computational difficulty of the multi-test problem.

For the computations of this article, $B = 11\,264$, $\theta = 0.001$, $\sqrt{2}\gamma/\delta = 2$ thus $N_b \leq 507\,003$.