Submitted to the Annals of Applied Statistics

A BAYESIAN GRAPHICAL MODEL FOR GENOME-WIDE ASSOCIATION STUDIES (GWAS)*

By Laurent Briollais[1,2†], Adrian Dobra[3], Jinnan Liu[1], Matt Friedlander[1,4], Hilmi Ozcelik[1] and Hélène Massam[4]

(1) *Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Canada*
(2) *Dalla Lana School of Public Health, University of Toronto, Canada*
(3) *Department of Statistics, University of Washington, USA*
(4) *Department of Mathematics and Statistics, York University, Canada*

The analysis of GWAS data has long been restricted to simple models that cannot fully capture the genetic architecture of complex human diseases. As a shift from standard approaches, we propose here a general statistical framework for multi-SNP analysis of GWAS data based on a Bayesian graphical model. Our goal is to develop a general approach applicable to a wide range of genetic association problems, including GWAS and fine-mapping studies and more specifically able to: 1) Assess the joint effect of multiple SNPs that can be linked or unlinked and interact or not; 2) Explore the multi-SNP model space efficiently using the Mode Oriented Stochastic Search (MOSS) algorithm and determine the best models. We illustrate our new methodology with an application to the CGEM breast cancer GWAS data. Our algorithm selected several SNPs embedded in multi-locus models with high posterior probabilities. Most of the SNPs selected have a biological relevance. Interestingly, several of them have never been detected in standard single-SNP analyses. Finally, our approach has been implemented in the open source *R* package genMOSS.

1. Introduction. The emergence of high-throughput technologies for SNP genotyping and their application to large scale genome-wide association studies (GWASs) have generated promises that the genetic basis of many common human diseases could be elucidated (Risch and Merikangas, 1996; Risch, 2000; Hirschhorn and Daly, 2005; Kingsmore et al., 2008; Kruglyak, 2008; McCarthy and Hirschhorn, 2008). These GWASs have identified hundreds of genetic variants implicated in various human diseases and complex traits, providing valuable insights into their genetic mechanisms (Hindorff et al., 2009a,b). The rationale underlying GWAS is that com-

*Research supported by a MITACS seed project
†Corresponding author
Keywords and phrases: Keywords: Graphical model, Bayesian, Stochastic Search, GWAS, SNP, Breast Cancer

mon genetic variants (i.e. present in more than 1-5% of the population) can explain most of the attributable risk of common human diseases, also referred to as the common disease, common variant (CDCV) hypothesis. The present paradigm for GWAS involves the collection of more than 1,000 cases and 1,000 controls and an exhaustive search among $> 500K$ SNPs of those associated with the disease outcome using simple univariate test statistics. Despite its relative merits at identifying new genetic variants, GWASs have also given rise to criticisms. For example, the SNPs selected through univariate statistics have generally a low predictive value, explain a fairly modest proportion of the genetic variability of the disease and maybe more importantly, do not usually provide much understanding of the underlying biological process.

A few alternative approaches have been proposed to the usual GWAS paradigm. After a pioneer paper demonstrating the feasibility of the exhaustive testing of two genetic markers (Marchini, Donnelly and Cardon, 2005), several papers emphasized the power of multi-SNP approaches (Zhang and Liu, 2007; Schwartz, Ziegler and Konig, 2008; Wu and Zhao, 2009). Two general classes of methods emerged: penalized regression and Bayesian selection methods. The most popular penalized regression approach, the LASSO (Tibshirani, 1996), has been further extended to GWAS analysis (Hoggart et al., 2008). The Bayesian framework offers various competing approaches for multi-SNP analysis, where usually a regression model for the response is defined as well as a prior for the regression coefficients associated with the SNPs. In order to deal with the high-dimensional model space, efficient stochastic search algorithms such as MCMC are needed to perform the model selection. While these approaches provide a step forward compared to univariate statistics, they also have limitations. They are sometimes restricted to low-dimensional models with only two SNPs (Zhang and Liu, 2007), might only consider those SNPs that are in linkage disequilibrium (Verzilli, Stallard and Whittaker, 2006) or could be more specifically designed for continuous outcomes (Guan and Stephens, 2011). They also often require a very "aggressive" first step selection to reduce the model space (Wilson et al., 2010). Our goal here is to propose a more general framework for multi-SNP analysis of GWAS data based on Bayesian graphical models.

The ability to model complex dependency structures makes graphical models an attractive approach for GWAS analysis. The application of graphical models to discrete genetic data such as SNPs remains relatively rare. Among the few examples, Thomas and Camp (2004) proposed the use of graphical models to study the patterns of allelic association between genetic markers in a small chromosomal region. Their work focused on decomposable

graphical models in the frequentist framework and used simulated annealing for model fitting. A more recent approach is based on a fully Bayesian approach where prior knowledge about linkage disequilibrium around each marker can be incorporated (Verzilli, Stallard and Whittaker, 2006). The model fitting used a MCMC algorithm that yields samples from the posterior probability and where inference is based on model averaging. They used decomposable graphical models where their clique definition was restricted to SNPs physically close to each other, ignoring the complex nature of association patterns in GWAS. Additional work on the application of probabilistic graphical models to genetic associations was also reported using either the Bayesian (Xing et al., 2011; Ungvari et al., 2012) or the frequentist framework (Jiang, Barmada and S., 2010; Han, Park and Chen, 2010).

Our goal in this paper is to develop a general approach, applicable to a wide range of genetic association problems and more specifically able to: 1) Assess the joint effect of multiple SNPs that can be linked or unlinked and can interact or not; 2) Explore the model space efficiently using the Mode Oriented Stochastic Search (MOSS) algorithm (Dobra and Massam, 2010) and determine the best multi-SNPs models. We illustrate the interest of our new methodology through an application to the CGEM breast cancer GWAS data.

2. Discrete Bayesian Graphical models for Modelling the Joint Effect of SNPs in GWAS.

2.1. *Overview of the approach.* In GWAS, we are interested in modelling the response variable (i.e. case control status) as a function of the SNP variables. Let $X = \{X_1, \ldots, X_r\}$ be a vector of random discrete variables with $Y = X_r$, $r \in V$ be a response variable and X_A, $A \subset V \setminus \{r\}$ be the set of SNPs. A typical GWAS dataset can include several thousands of SNPs with the aim of finding a small subset associated with the case-control status. Our goal is therefore to search for sets A such that the probability of the regression $[Y|X_A]$ is highest. This probability can be expressed as the ratio between the marginal likelihood of the saturated model for $(n)_{A\cup\{r\}}$ and for $(n)_A$ (Dobra and Massam, 2010), where $(n)_{A\cup\{r\}}$ and $(n)_A$ are cross-classifications involving $X_{A\cup\{r\}}$ and X_A, respectively.

$$P(Y|X_A) = \frac{P(Y, X_A)}{P(X_A)} \tag{1}$$

Because of the complex dependence structure among the SNPs in a GWAS, the marginal likelihood of the models is expressed using graphical model

methodology and the search for the best regression models in this high-dimensional setting is conducted using the mode oriented stochastic search (MOSS) algorithm (Dobra and Massam, 2010) (see section 3).

2.2. *Graphical models.* In this paper, we assume that the variables $X_i, i = 1, \ldots, r$, that include the SNPs and case-control status, take a finite number of values. Practically, the case-control status is binary with values 0 (controls) and 1 (cases) whereas the SNP variables can take up to three values. For ease of notation, we recall the theory below for binary variables only. The reader is referred to Massam, Liu and Dobra (2009) for general notations. We consider a fixed number N of individuals that we classify in a contingency table according to these r criteria. Let $\mathcal{E}$ denote the collection of all non empty subsets of $V$and $\mathcal{E}_0$ the collection of possible subsets of V including $\emptyset$. The elements F in $\mathcal{E}_0$ are in 1-1 correspondence with the cells in the contingency table and we can use p_F to denote the cell probability

$$p_F = P(X_v = 1, v \in F, X_v = 0, v \notin F). \tag{2}$$

i.e. the probability that, for a given individual, the variables $X_v, v \in F$ are all equal to 1 while the variables $X_v, v \in V \setminus F$ are all equal to 0.

Since N is fixed, the cell counts $n_F, F \in \mathcal{E}_0$ follow a multinomial distribution with the well-known density function

$$f((n), p) = \binom{N}{(n)} p_\emptyset^{N - \sum_{F \in \mathcal{E}} n_F} \prod_{F \in \mathcal{E}} p_F^{n_F} \tag{3}$$

where the parameters are the cell probabilities $p_F, F \in \mathcal{E}_0$.

An alternative representation of the multinomial distribution is to write it in a natural exponential family form using loglinear parameters instead of cell probabilities. We use the following loglinear parameters

$$\theta_E = \log \prod_{F \subseteq E, F \in \mathcal{E}_0} p_F^{(-1)^{|E \setminus F|}} \text{ with } \theta_\emptyset = \log p_\emptyset \, . \tag{4}$$

where θ_E can be interpreted as the generalized log odds ratio.

Using Moebius inversion lemma, we can show that (4) is equivalent to

$$\log p_E = \sum_{F \subseteq E, F \in \mathcal{E}_0} \theta_F \text{ with } \log p_\emptyset = \theta_\emptyset \, . \tag{5}$$

After the change of variable $(n_F, F \in \mathcal{E}) \mapsto (y_F, F \in \mathcal{E})$ where

$$y_F = \sum_{D \supseteq F, D \in \mathcal{E}_0} n_D \tag{6}$$

and $y_\emptyset = N$ are the marginal F-cell counts and total counts respectively, the multinomial density for $(n) = (n_F, F \in \mathcal{E}_0)$ becomes the following density for $y = (y_F, F \in \mathcal{E})$

$$f(y;\theta) = \exp\Big(\sum_{E\in\mathcal{E}} \theta_E y_E - N\log(1+\sum_{E\in\mathcal{E}} \exp(\sum_{D\subseteq E} \theta_D))\Big). \tag{7}$$

Let us now consider the case of interest in this paper, i.e., the case where the model for X is a graphical model, which we will now define.

An undirected graph G is a pair (V, E) where $V = \{1, 2, \ldots, r\}$ is a finite set of vertices, and E, the set of edges, is a subset of the set $V \times V$ of unordered pairs of distinct vertices $\{i, j\}, i \in V, j \in V$. Let $X = \{X_1, \ldots, X_r\}$ be a vector of random variables. Each variable X_i is represented by the vertex i of G. For $A \subseteq V$, X_A indicates the collection of random variables $\{X_i, i \in A\}$. In GWAS, the vertices represent the disease status, the SNPs, and occasionally confounding variables (e.g. that control for population stratification).

For G given, a probability distribution for X is said to be Markov with respect to G if for any two non-adjacent vertices $i, j \in V$, X_i is independent of X_j given $X_{V\setminus\{i,j\}}$. Therefore no edge between two variables means conditional independence between these variables given all the other variables while an edge between two variables is an indication of association between these variables. A graphical model is a family of probability distributions for X Markov with respect to a given graph G. A discrete graphical model is a graphical model where each random variables $X_i, i = 1, \ldots, r$ is discrete.

For a given model with underlying graph G, let

$$\mathcal{D} = \{D \in \mathcal{E} \mid D \quad \text{is complete in} \quad G\} \tag{8}$$

be the clique set of G. For E and F in $\mathcal{E}_0$, we will use the notation

$$E \subseteq_G F$$

to mean that $E \subseteq F$ and $E \in \mathcal{D}$. Following Massam, Liu and Dobra (2009), it can be shown that for a graphical model Markov w.r.t. the graph G

$$\theta_E = 0, \quad E \notin \mathcal{D}. \tag{9}$$

Then (5) and (7) become respectively

$$\log p_E = \sum_{F\subseteq_G E, F\in\mathcal{E}_0} \theta_F \tag{10}$$

$$f(y;\theta) = \exp\Big(\sum_{E\in\mathcal{D}} \theta_E y_E - N\log(1+\sum_{E\in\mathcal{E}} \exp(\sum_{D\subseteq_G E} \theta_D))\Big) \tag{11}$$

From (10), it is immediate to derive the conditional distribution of X_v given $X_{V\setminus\{v\}}$, $v \in V$ and to show that $P(X_v = 1|X_{V\setminus\{v\}})$ is a function of $(\theta_D, D \in \mathcal{D}, v \in D)$ only and therefore, in the logistic regression of X_r where X_r represents the disease status. If the parameter $\theta_{\{r,u\}} = 0, u \in V \setminus \{r\}$, we can conclude that X_r is conditionally independent of X_u given the other variables and that therefore there is no edge between r and u in G.

Thus, we see that graphical models together with MOSS allow us to select the best SNPs jointly associated with the response variable, including marginal and interaction SNP effects (See section 3).

2.3. *Bayesian Graphical Model (BGM).* Let us assume we perform a model search in the family of models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. We write the models as

$$\mathcal{M}_j = \{p(x|\vartheta), \vartheta \in \Theta_j\}, j = 1, \ldots, k \tag{12}$$

where ϑ is a parameter in the parameter set Θ_j and $p(x|\vartheta)$ is a probability density function. In the particular case where the model is a graphical model, the parameter space is defined by the underlying graph G and we identify models $\mathcal{M}_j$ with their underlying graph G_j.

In a Bayesian framework we assume a prior probability $P(\mathcal{M}_j), j = 1, \ldots, k$ on the set of models $(\mathcal{M}_1, \ldots, \mathcal{M}_k)$ and a prior probability on the parameters ϑ, and want to derive the posterior model probabilities $P(\mathcal{M}_j|\mathbf{x})$ for each one of the models $\mathcal{M}_1, \ldots, \mathcal{M}_k$, that is, the conditional distribution of $\mathcal{M}_j$ given the data.

The Bayesian solution is to choose the model with the highest posterior probability. According to Bayes' theorem, the posterior probability for $\mathcal{M}_j$ is

$$P(\mathcal{M}_j|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{M}_j)P(\mathcal{M}_j)}{\sum_{i=1}^k P(\mathbf{x}|\mathcal{M}_i)P(\mathcal{M}_i)} \tag{13}$$

The term $\sum_{i=1}^k P(\mathbf{x}|\mathcal{M}_i)P(\mathcal{M}_i)$ in (13) is a constant. Therefore we can write

$$P(\mathcal{M}_j|\mathbf{x}) \propto \underbrace{P(\mathbf{x}|\mathcal{M}_j)}_{(the\ \ Marginal\ \ Likelihood)} \quad \underbrace{P(\mathcal{M}_j)}_{(the\ \ Model\ \ Prior)} \tag{14}$$

In our problem, $p(x|\vartheta)$ is given by (11) and therefore $\vartheta = \theta = (\theta_D, D \in \mathcal{D})$ and since (11) is a member of a natural exponential family, the conjugate priors for θ have density of the form

(15)
$$\pi_G(\theta|s,\alpha) = I_G(s,\alpha)^{-1} \exp\{\sum_{D\in\mathcal{D}} \theta_D s_D - \alpha \log\Big(1 + \sum_{E\in\mathcal{E}} \exp(\sum_{D\subseteq_G E} \theta_D)\Big)\} \ ,$$

where $s = (s_D, D \in \mathcal{D}) \in \Re^{|\mathcal{D}|}$ and $\alpha \in \Re$ are hyperparameters and $I_G(s,\alpha)$ is the normalizing constant.

2.4. *Specification of the prior.* A method to construct hyperparameters of a proper prior $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|(s,\alpha))$ is to start with a fictive prior contingency table with all cell counts ν_F positive, not necessarily integers. With α denoting the total count in the given fictive contingency table, γ_D denoting the marginal cell counts, we can take as hyperparameters $\alpha = N$ and $s_D = \gamma_D,\ D \in \mathcal{D}$. Lack of prior information can be expressed through what is sometimes called a flat prior by taking all the fictive cell entries to be equal and equal to $\frac{\alpha}{|\mathcal{I}|}$. We used this latter prior specification in our simulations and real data application.

2.5. *Posterior of a model.* The posterior of G is proportional to the ratio of the two normalizing constants:

$$P(G \mid Y) \propto I_G(y+s, n+\alpha)/I_G(s,\alpha). \tag{16}$$

For G decomposable, the prior $\pi(\theta|\alpha,s)$ is identical to the hyper Dirichlet (see Massam, Liu and Dobra (2009)). It therefore follows that the normalizing constants I_G can be computed analytically when the graph G is decomposable. When G is non decomposable, I_G needs to be computed numerically.

3. SNP selection with the MOSS algorithm. The mode oriented stochastic search (MOSS) algorithm is a two-stage Bayesian variable selection procedure that aims at identifying combinations of SNPs (rather than single SNPs) that are associated with a response variable. The first stage of MOSS consists of identifying the best saturated graphical models including the response variable and a small subset of SNPs (typically between 2 to 6 SNPs in a GWAS). The second stage is used to search the space of log-linear models to identify the most relevant interactions among the variables in each of the top models. By using the generalized hyper Dirichlet prior of Massam, Liu and Dobra (2009), the computations in both steps is done

efficiently. The principle of MOSS is the following:

Let $\mathcal{M}$ denote a set of possible regression models. We associate with each candidate model $m \in \mathcal{M}$ a neighbourhood $\text{nbd}(m) \subset \mathcal{M}$. Any two models $m, m' \in \mathcal{M}$ are connected through a path $m = m_1, m_2, ..., m_l = m'$ such that $m_j \in \text{nbd}(m_{j-1})$ for $j = 2, ..., l$. The neighbourhood of $m = [Y|X_A]$ is obtained by addition moves, deletion moves, and replacement moves. For details see Edwards and Havranek (1985) and Dellaportas and Forster (1999). In an addition move, we include in A any variable in $V \backslash A$, one at the time. In a deletion move, we delete any variable that belongs to A, one at the time. For a replacement move, we replace any one variable in A with any one variable in $V \backslash A$. The first stage of the MOSS procedure is as follows:

We make use of a current list of regressions $\mathcal{M}$ that is updated during the search. Define

$$\mathcal{M}(c) = \left\{ m \in \mathcal{M} : P(m) \geq c \max_{m' \in \mathcal{M}} P(m') \right\}$$

where $c \in (0, 1)$. A regression $m \in \mathcal{M}$ is called explored if all of its neighbours $m' \in \text{nbd}(m)$ have been visited.

1. Initialize a starting list S of regressions. For each $m \in S$, calculate and record its marginal likelihood $P(m)$. Mark m as unexplored.
2. Let L be the set of unexplored regressions in S. Sample an $m \in L$ according to probabilities proportional with $P(m)$ normalized within L. Mark m as unexplored.
3. For each $m' \in \text{nbd}(m)$, check if m' is currently in S. If it is not, evaluate and record its marginal likelihood $P(m)$. Eliminate the regressions $S \backslash S(c')$ for some pre-chosen value $0 < c' < c$.
4. With probability q, eliminate from S the regressions in $S \backslash S(c)$.
5. If all the regressions in S are explored STOP. Otherwise return to step 2.

The role of the parameters c, c', and q is to limit the number of regressions that need to be visited to a manageable number. In our simulations and real data application, the values c, c', and q were set to control the false discovery rate (FDR) at a given level.

At the end of the first stage, we have a set of top regressions $[Y|X_A]$, each involving a small number of SNPs. At this point, we relax the assumption that the saturated model holds for all the variables V. In the second stage,

we search the space of log-linear models $[Y, X_A]$ to identify the most relevant interactions among the SNPs and between the SNPs and the response variable in each regression. We do a separate search for each regression identified in the first stage, looking for the log-linear model $m = [Y, X_A]$ with the highest marginal likelihood. To do this, we once again begin by defining the concept of the neighbourhood of a model and for a given set of variables, the algorithm tries to find m that maximizes $P(m)$ in an analogous way to the algorithm described above.

At the end of the second stage of MOSS, we also add a pruning procedure where any of the variables X_A that is not interacting with the response variable Y in the log-linear model is removed from the list of SNPs selected. In this second stage, we use a small α (i.e. $\alpha = 0.01$) to favour sparser models to be selected (Letac and Massam, 2012) and found this strategy to perform well in all our simulation scenarios.

4. Risk Estimation and Prediction Based on Bayesian Model Averaging. Once MOSS has identified a set of regression models $\mathcal{S}$ for some $c \in (0, 1)$, one can estimate the risk associated with the selected SNPs and perform risk prediction. This is done using Bayesian Model Averaging. Let us consider a regression model $m_j \in \mathcal{S}(c)$, which is $[Y|X_{A_j}]$ with $A_j \subset V \setminus \{r\}$ and $j \in B$. Here B is a set of indices for the collection of models over which we are averaging. The regression model of Y on the selected variables X_{A_j} (i.e. SNPs) is a weighted average of the regression models in $\mathcal{S}$, where the weights represent the posterior probability of each regression model (see for example Yeung, Bumgarner and Raftery (2005)):

$$\Pr(Y = y|(n)) = \sum_{j \in B} \Pr(Y = y|(n)_{A_j}) \cdot \Pr(m_j|(n)). \tag{17}$$

Since we assumed that all models are a priori equally likely, the posterior probability of each regression is equal to its marginal likelihood normalized over all the models in $\mathcal{S}$:

$$\Pr(m_j|(n)) \quad = \quad \frac{\Pr(r|A_j)}{\sum_{l \in B} \Pr(r|A_l)}.$$

It is shown in Madigan and Raftery (1994) that the weighted average of regressions in (17) has a better predictive performance than any individual model in $\mathcal{S}$. The revelance of each predictor X_j can be quantified by its posterior inclusion probability defined as the sum of the posterior probabilities of all the models that include X_j.

To estimate the parameters of a given graphical model, one can use the algorithm described in Dobra and Massam (2010), called the Bayesian iterative proportional fitting algorithm. Alternatively, we can estimate parameters at the mode of the posterior distribution, as implemented in our *R* package genMOSS.

5. Simulation Study.

5.1. *Simulation scenarios.* To assess the performance of this novel BGM, we simulated datasets that mimic real GWAS data. In scenarios 1 to 3, we simulated GWAS data according to the breast cancer study analyzed in Section 7, which includes 1,145 cases and 1,142 controls. This represents a typical GWAS problem where the SNPs are tag SNPs with low linkage disequilibrium (LD). In addition, we also simulated $53K$ SNPs taken from the original breast cancer GWAS in section 7 and assumed them to be independent from the disease status. This represents approximately 10% of the total number of SNPs available for this study. Since these SNPs were extracted from a real SNP array, they have a realistic genome-wide correlation structure. Scenarios 4 and 5 correspond to a fine mapping problem where the SNPs are in high LD and are extracted from a small chromosomal region.

In our **Scenario 1**, we consider a 5-SNP main effects model with an interaction between SNP2 and SNP3. For each individual, the case-control status was generated from a Bernouilli trial with probability p of being a case given by

$$\text{logit}(p) = \beta_0 + \sum_{i=1..5} \beta_i SNP_i + \beta_{23} SNP_2 \times SNP_3.$$

We chose the β's to reflect the range of SNP effects found in our real GWAS data (section 7), i.e. $\beta_1 = 0.405$, $\beta_2 = 0.916$, $\beta_3 = 0.182$, $\beta_4 = -0.405$, $\beta_5 = 1.386$ and $\beta_{23} = \beta_2 \times \beta_3$, corresponding to odds-ratio for the genetic association of 1.5, 3.0, 1.2, 0.67, 5.0, 1.92, respectively. The parameter β_0 was determined to get $1,145$ cases and $1,142$ controls as in our real dataset.

For each SNP, we generated two genotypes with probability $f_i, 1 - f_i, i = 1, \ldots, 5$ from a Bernoulli trial with f_i equal to 0.10, 0.10, 0.36, 0.36 and 0.02 for the five SNPs. These genotype frequencies correspond to a minor allele frequency (MAF) of 0.05, 0.05, 0.2, 0.2 and 0.025, respectively, under a dominant genetic model, and thus represent a wide spectrum of uncommon,

common and rare SNPs.

Scenario 2 corresponds to a SNP-SNP interaction model with three 2-way interactions without main effects. For each individual, the case-control status was generated from a Bernoulli trial with probability p of being a case given by

$$\text{logit}(p) = \beta_0 + \beta_{23} SNP_2 \times SNP_3 + \beta_{34} SNP_3 \times SNP_4 + \beta_{45} SNP_4 \times SNP_5.$$

The MAF of the SNPs was 0.25, 0.25, 0.36, 0.36 and 0.15, respectively for the five SNPs, and the regression coefficients for the interactions were $\beta_{23} = 1.099$, $\beta_{34} = 0.916$ and $\beta_{45} = 1.609$.

The analyses were performed with main effects models only under this scenario since some of the methods used in our simulations cannot model specifically the interactions.

Scenario 3 corresponds to a fine mapping problem where a causal SNP (SNP_2) is associated with a disease status Y, and where SNP_2 is in linkage disequilibrium (LD) with two other SNPs, SNP_1 and SNP_3. These two other SNPs are conditionally independent of the disease status given SNP_2, so that only one causal locus in the region is observed. The 3 SNPs constitute a cluster of SNPs and are denoted X_1, X_2, and X_3 for simplicity. The distribution of the four discrete variables in the graph was generated from a multinomial distribution and can be represented by a 4-way contingency table with joint cell probabilities given by the log-linear model

$$logP_{ijkl} = \theta_i^Y + \theta_j^{X_1} + \theta_k^{X_2} + \theta_l^{X_3} + \theta_{ik}^{YX_2} + \theta_{jk}^{X_1X_2} + \theta_{kl}^{X_2X_3}$$

where the subscripts $i, j, k, l \in \{0, 1\}$ index the levels of the variables Y, X_1, X_2, and X_3, respectively. We chose as parameters $\theta_1^Y = 0$, $\theta_1^{X_1} = \theta_1^{X_2} = \theta_l^{X_3} = \ln(0.2) = -1.6$, $\theta_{11}^{YX_2} = 0.717$, $\theta_{11}^{X_1X_2} = \theta_{11}^{X_2X_3} = 1.792$. All the other parameters were set to 0. The association among SNPs has a level of LD of $Q = 0.71$ and D' close to 0.55 (Devlin and Risch, 1995), which is a moderate/strong level, as generally observed for the SNPs in a same haplotype block. The MAF was 0.20 for all three SNPs and we assumed a dominant model for SNP_2.

Scenarios 4 & 5: After a GWAS, the main loci discovered are often followed-up by a fine-mapping study with the goal to refine the location

of the causal genetic variants. These two scenarios correspond to a fine-mapping problem and are motivated by a real data analysis on prostate cancer (PCa). We used real data from an ongoing study focusing on SNPs located in the Kallikrein (KLK) region on chromosome 19 and we simulated the fine-mapping data with a similar LD structure. This study included 772 cases and 1,052 controls and the KLK region was originally composed of 308 SNPs (Scenario 4). To assess the sensitivity of the multiple-SNP models to the SNP density in the region, we also imputed an additional 590 SNPs from the 1,000 genomes reference data to reach a total of 898 SNPs (Scenario 5). The KLK family consists of 15 genes clustered in a region that spans about 261,558 bp on chromosome 19q13.3-4 and display significant homology to each other (Diamandis and Youssef, 2002). PSA is a member of the KLK family, a very important gene family in PCa diagnosis. The KLK region was partitioned into 51 haplotype blocks based on the Haploview software (Barrett et al., 2005). We assumed 3 causal SNPs located in the haplotype blocks 7, 19 and 49, associated with the outcome with an OR of 2.0, all with a dominant effect and a MAF of 19.2%, 15.6% and 18.6%, respectively. The non causal SNPs (303 SNPs and 896 SNPs in scenarios 4 & 5, respectively) were taken from the original data and analyzed with 3 genotype categories. Their MAFs were all $> 2\%$.

5.2. *Method comparison.* For our method comparison, we used the penalized regression method LASSO (Tibshirani, 1996), two Bayesian approaches: BVRS (Guan and Stephens, 2011) and BEAM3 (Zhang, 2012), and a simple test statistic based on χ^2 statistic with either Bonferonni or FDR correction for multiple testing and applied all these methods to our simulated datasets.

For **LASSO**, we used the HyperLASSO formulation proposed for GWAS data by Hoggart et al. (2008), which is based on shrinkage priors. Each regression coefficient is assigned an independent shrinkage prior with a density function that is sharply peaked at zero. The prior density function can be defined either as a double exponential (DE) or a normal exponential gamma (NEG) distribution. Parameter estimates are obtained by maximizing the posterior density $p(\beta|\mathbf{X}, \mathbf{y})$ over β, where $\mathbf{X}$ is the normalized genotype data and y the response variable (i.e. the case-control status). Taking logarithms in Bayes theorem, the problem can be thought of as maximizing the penalized log-likelihood function:

$$\log p(\beta|y, \mathbf{X}) = L(\beta) - f(\beta) + const$$

where L is the log-likelihood for the logistic regression model and f is the log-

prior density with a minus sign to allow f to be interpreted as a penalty function. With the DE prior, the maximization of the penalized log-likelihood is equivalent to the LASSO procedure. A SNP j is included in the final regression whenever:

$$|L'(\beta_j = 0)| > f'(\beta_j = 0^+)$$

We used the NEG distribution prior in our simulations. To control the FDR at a given level, we changed the hyper-parameters of the prior NEG distribution.

The **BVSR** approach is based on a linear multi-SNP regression model and defines some normal priors for the regression coefficients associated with the SNPs as well as for the probability of each regression coefficient to be zero in the model, which controls the sparsity of the model. An important feature of this approach is to to have the normal prior distribution for the regression coefficients that depend on a parameter that controls the proportion of genetic variability explained (PVE) by the selected SNPs (which itself is function of the sparsity parameter). This prevents the risk that more complex models explained substantially higher PVE. The induced prior for PVE given the sparsity parameter is a "flat" prior in the range (0,1). BVSR was initially developed for continuous outcomes but was then extended to binary responses, using a probit link function. The inference is based on MCMC and models with highest posterior PVE are selected. In our simulations, we allowed the model size for BVSR to vary between 1 and 5 and the hyper-parameters were chosen so that the control of FDR was similar to that of the other approaches whenever possible.

BEAM3 is a Bayesian graphical method recently developed for large-scale association mapping (Zhang, 2012). BEAM3 can simultaneously detect single-SNP and SNP-SNP interactions in genetic association studies. It was described as a powerful method for analyzing a large number of SNPs even in the context where the SNPs are in strong LD (Zhang, 2012). The rationale behind BEAM3 is to define two sets of SNPs, i.e. those associated with the response and those not associated. The SNPs within these two sets are embedded into two distinct cliques of a graph that account for SNP dependency and for which a joint probability distribution is specified. The method also requires to define a prior inclusion probability for the SNPs to be included in the associated and non-associated sets as well as a prior distribution for the cliques partition and clique interactions for the associated set. The inference is done with MCMC and can be summarized through a

posterior probability of inclusion for the associated SNPs.

MOSS algorithm is described above. We fitted 3-SNP models in all simulation scenarios with the following settings: α=1, c=0.005, c'=0.0005, q=0.1, replicates=10 in the first stage of MOSS and α=0.01, c=0.005, c'=0.0005, q=0.1 in the second stage. These settings gave us the best performance of MOSS in all the simulation scenarios.

Univariate test statistics: We also calculated a χ^2 statistic for testing single SNP associations. Since the SNPs had 2 genotypes in simulation scenarios 1 to 3 and 3 genotypes in scenarios 4 and 5, the number of degrees of freedom for the χ^2 statistics was respectively 1 and 2 in these situations. A correction for multiple testing was performed using either a Bonferroni or an FDR adjustment.

5.3. *Control of False and True Discovery Rate.* For our different simulation scenarios, we estimated the False Discovery Rate (FDR) (see Benjamini and Hochberg (1995)) by the proportion of non causal SNPs among all the SNPs discovered by a particular approach. The main effect FDR (FDRm) for a SNP j is defined for a particular method as

$$FDR_m = \frac{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k \cap \text{ SNP } j \notin \text{causal SNPs })}{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k)}$$

where N_d is the number of simulated datasets and $I(.)$ is the indicator function.

In scenarios 4 and 5 of the simulations, we also computed the cluster FDR (FDR_C) where the cluster C corresponds to the haplotype block and the lenient cluster FDR (FDR_{LC}) defined as:

$$FDR_C = \frac{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k \cap \text{ SNP } j \notin \text{cluster } C)}{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k)}$$

and

$$FDR_{LC} = \frac{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k \cap \text{ SNP } j \notin \text{clusters } \{C-1, C, C+1\})}{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k)}$$

We also computed the true discovery rate (TDR) for each individual SNP j associated with the outcome as

$$TDR = \frac{\sum_{k=1\cdots N_d} I(\text{ SNP } j \text{ is discovered in dataset } k)}{N_d}$$

and an overall TDR as

$$TDR_{all} = \frac{\sum_{k=1\cdots N_d} \sum_{j=1\cdots N_s} I(\text{ SNP } j \text{ is discovered in dataset } k)}{N_d \times N_s}$$

where N_s the number of causal SNPs associated with the outcome.

In scenarios 1 and 2 of our simulations, we computed a TDR for each specific pair of SNPs (j, j') corresponding to the interaction term in our simulated model as

$$TDR_{pair} = \frac{\sum_{k=1\cdots N_d} I(\text{ SNPs } j \text{ and } j' \text{ are discovered in dataset } k)}{N_d}$$

In scenarios 4 and 5, we also computed a TDR for the cluster (haplotype block) and a lenient cluster TDR defined as:

$$TDR_C = \frac{\sum_{k=1\cdots N_d} \sum_{j=1\cdots N_s} I(\text{ SNP } j \text{ is discovered in dataset } k \cap \text{ SNP } j \in \text{cluster } C\ )}{N_d \times N_s}$$

and

$$\text{TDR}_{LC} = \frac{\sum_{k=1\cdots N_d} \sum_{j=1\cdots N_s} I(\text{ SNP } j \text{ is discovered in dataset } k \cap \text{ SNP } j \in \text{cluster } \{C-1, C, C+1\}\ )}{N_d \times N_s}$$

5.4. *Control of FDR.* We tried to control FDRm at the same level with all the methods compared to get a fair comparison of the TDR statistics. This was achieved by varying the tuning parameter of the NEG distribution with the HyperLASSO and the SNP inclusion probability with the three

Bayesian approaches MOSS, BVSR and BEAM3. However, the control of FDRm could not always be achieved for all the methods in certain situations. Along with the FDR and TDR statistics, we also computed the rank of each SNP based on a χ^2 statistic with 1 *df* (Scenarios 1 to 3) or 2 *df*'s (Scenarios 4 and 5).

6. Simulation results.

6.1. *GWAS simulation results (Scenarios 1 to 3).* FDR and TDR results are presented in Table 1 and in Supplement B.

In **Scenario 1**, all methods control FDRm at a level $< 20\%$ except BEAM3 for which FDRm is much higher (30%). Under very similar simulation scenarios, previous results reported FDR levels very close to ours (Hoggart et al. (2008), He and Lin (2011)). TDR estimates vary substantially across methods with the best results obtained with BVSR (62.2%) and MOSS (56.8%). The pairwise SNP effects are better detected by BVSR and MOSS with TDR_{pair} of 71% and 30%, respectively. The AUC values are all close to each other, from 58.0 to 61.1%. In **Scenario 2**, all methods control FDRm at a very low level (i.e <6.1%). MOSS yields the best TDR results with 73.7% while BVSR performs the worst (TDR=18.5%). The AUCs vary between 53.2 to 59.8%. MOSS performs also very well to detect pairwise SNP interactions. The **Scenario 3** is the most complex since the goal is to find one single causal variant among a group of 3 SNPs in strong linkage disequilibrium. This complex situation is reflected by an overall higher level of FDRm compared to scenarios 1 and 2 and a larger difference across methods (i.e. with FDRm varying from 1.0% to 47.0%). In that situation, BVSR and LASSO were the only methods to not find the causal variant in all the simulated datasets. MOSS has the lowest FDRm in that scenario, i.e 1%, while all the other methods have much larger FDRm statistics (varying from 16.7% to 47.0%). In all three scenarios, MOSS has better performance than the univariate chi-square statistic with either Bonferroni or FDR adjustment for multiple testing.

[Table 1 about here.]

6.2. *Fine mapping simulations (Scenarios 4 and 5).* FDR and TDR results are presented in Table 2.

In **Scenario 4**, it was not possible to control FDR_m at a similar level with all the different methods. The best results are obtained with MOSS

both in terms of FDR (all values below 10.0%) and TDR (> 96.3%). The LASSO also performs well under this scenario but with inflated FDR values compared to MOSS (FDR_m=18.0%) and TDR values > 89.0%. The single SNP analyses BONFmain and FDRmain gave very poor results. The AUCs vary between 67.8% (with BVSR) and 70.7% (with BONFmain). In **Scenario 5**, all the multi-SNP methods control FDRm at a level < 20.9% while the two single SNP analyses showed huge inflation of this statistic: BONFmain (91.3%) and FDRmain (38.7%). MOSS has the highest TDR statistics (based on TDR_m, TDR_c and TDR_{tc}) and reaches levels > 93.7%. LASSO has also levels of TDR > 93% but to the price of increased level of FDR_c and FDR_{lc}. The AUC values vary from 52.5% (with FDRmain) to 70.5% (with BONFmain). Additional simulation results are given in Supplement C.

[Table 2 about here.]

6.3. *Computation time.* In simulation scenarios 1 to 3, the median computation time to fit one simulated dataset was about 6 hours for a 2-SNPs model and 17.9 hours for a 3-SNPs model with MOSS, 13 mins with the LASSO and 5-6mins with BEAM3 and BVSR. For the simulation scenarios 4 and 5, the computation time was about 4-5 mins for a 2-SNPs model and about 6 mins for a 3-SNPs model with MOSS, 1 min with the LASSO, about 10 secs with BEAM3 and 5 mins with BVSR. The longer computation time required by MOSS could be explained by the extensive model search performed by this algorithm compared to the other methods.

6.4. *Sensitivity to prior specification with MOSS.* We noticed that the performance of MOSS in terms of FDR and TDR remain unchanged for various specifications of the priors (results not shown). In particular, defining the prior cell counts to be all 1 or proportional to the sample size of the observed cell counts with various possible proportions, did not change our main conclusions (Supplement D).

7. Analysis of the CGEM breast cancer GWAS data.

7.1. *The breast cancer paradigm.* In most Western populations, approximately one in ten women develop breast cancer. Epidemiological studies have shown that women who have first-degree relatives with a history of breast cancer have a two-fold increase in risk of the disease (Collaborative Group on Hormonal Factors in Breast Cancer, 2002). The risk ratio increases with increasing the number of affected first-degree relatives. Twin studies

have indicated that most of the excess familial risk is due to inherited predisposition (Peto and Mack, 2000). Particularly, BRCA1 and BRCA2 are the most important susceptibility genes conferring, when mutated, high lifetime risks of breast cancer (Thompson and Easton, 2002; The Breast Cancer Linkage Consortium, 1999). Mutations in BRCA1 and BRCA2 account for about 16% of the familial risk of breast cancer (Anglian Breast Cancer Study Group, 2000). Mutations in other genes (TP53, PTEN, STK11, CDH) are also associated with elevated risks but it is unlikely that mutations in these six genes account for more than 20% of the familial risk of the disease. Therefore the remaining 80% of the familial risk remains to be explained. The search for this missing heritability has led to the identification of other high-penetrant mutations in candidate genes such as CHEK2, ATM, BRIP1 and PALB2. However, they still confer a small contribution to the familial risk of breast cancer (Thompson and Easton, 2004). Alternatively, common low-penetrant alleles have been sought through GWAS. So far, only a small number of such variants have been identified and confirmed in different populations and they just modestly improved the performance of risk models for breast cancer (Wacholder et al., 2010; Gail, 2008). The bulk of breast cancer genetic susceptibility thus remains to be determined.

7.2. *The CGEM study.* The CGEM genome-wide association studies (GWAS) for breast cancer has been completed in the Nurses' Health Study (NHS) with nearly 550,000 SNPs genotyped. The analysis includes 1,145 individuals who developed breast cancer during the observational period and 1,142 age-matched individuals who did not develop breast cancer during the same time period. Both the genotype data and the pre-computed analyses based on the genotype data were retrieved from the following website (http://cgems.cancer.gov/). The first GWAS study using the CGEM breast cancer data identified several SNPs within the gene $FGFR2$ (Hunter et al., 2007) and this result has been replicated in many independent studies. A SNP close to the gene BUB3 was also very significant in the initial study but has not been replicated yet.

7.3. *Data pre-processing.* Our initial dataset included 555,341 SNPs and 2,287 observations (1,145 affected individuals and 1,142 controls). After exclusion of SNPs with a high rate of missing genotypes (missing rate $\geq 10\%$), we had 546,540 SNPs left. For the remaining SNPs, we imputed the missing genotype values using the program MACH (Li et al., 2010). Our final number of SNPs after imputation was 546,253. We assessed the presence of population stratification using the program EIGENSTRAT (Price et al., 2006), which is based on principal component analysis (PCA). Using pro-

jections on the two first principal components, we found 20 individuals (9 cases and 11 controls) who appear to be outliers and were removed from our analysis. We also estimated the identity by state (IBS) matrices in cases and controls separately based on all the SNPs and compared the mean IBS values between the 2 groups using permutation testing as implemented in the software PLINK (Purcell et al., 2007). Because we did not find any significant difference, no adjustment for population stratification was performed in our analysis. We did not filter out SNPs neither based on their MAF nor on Hardy-Weinberg disequilibrium test since there was no evidence of deviation of this test in our data (Hunter et al., 2007).

7.4. *Analyses with MOSS.* We searched for regression models containing at most 2 and 3 SNPs but since the selection of the best SNPs was very similar under these 2 models, we only present the simple 2 SNP models in Table 3. The total number of possible 2-SNP regressions was 1.49197×10^{11}. The number of models evaluated by MOSS in each of the $1,000$ instances was considerably smaller and varied between $209,445$ and $837,778$, with a mean of $497,065$. The eight regressions in the resulting $\mathcal{S}(0.5)$ involve twelve SNPs embedded or very close to known genes (Table 3).

MOSS selected 12 SNPs with MAF varying between 0% and 42% in the European population. Three SNPs have a MAF lower than 5%. In general, frequentist approaches applied to GWAS would not be able to perform a test statistic for these SNPs. Most of the SNPs detected by MOSS have a high rank when using the more conventional univariate $p-$value criteria. The two SNPs in the gene $FGFR2$ were previously identified from univariate analysis of the CGEM data (Hunter et al., 2007) and have been replicated in multiple studies. The SNP in the gene $BUB3$ was also identified in the initial analysis of the CGEM data but not further replicated. It is noteworthy that MOSS was able to replicate some initial findings from the CGEM study. Additionally, several novel SNPs emerge from our analysis. An example, is the SNP $rs3130544$ associated with the highest posterior probability. To our knowledge, this SNP has never been identified in previous breast cancer GWAS. We also noticed that the SNP $rs1882619$ in the gene APC which has a very low rank based on univariate analysis, would have never been selected with a standard approach. This SNP has been selected by MOSS because it has a joint effect with the SNP in the gene $BUB3$. While MOSS is able to detect more SNPs associated with the disease of interest in GWAS, the question remains to know whether these results have any biological validity. In the next tables and Supplement E, we give more insights into the biological in-

terpretation of our results.

[Table 3 about here.]

Interestingly, 4 out of 17 genes in our list have been previously implicated in breast cancer, including $BUB3$, $NTSR1$, $FGFR2$ and APC. Furthermore, eight genes have previous relation to cancers, which suggests an enrichment of cancer genes in the MOSS selection. The most interesting gene found by MOSS is the gene $C6orf15$. This gene is located in the HLA region and does not have a very clear function. However it is located in a region characterized by a dense cluster of genes which has been found overexpressed in many cancer types. This is therefore a region that would be worth sequencing to find potential causal variants associated with breast cancer or other cancers.

Table 4 displays the best eight two-SNP models identified by MOSS and their associated marginal likelihood and Bayes Factor (BF). We first notice that the BFs for these models are much higher (from 18.32 to 17.18) than any of the BF for the single SNP models in Table 3 (i.e. the maximum value was 4.50 for the SNP $rs10510126$ close to gene $BUB3$). There is also a certain level of internal replication. Indeed, two pairs of models $(1, 3)$ and $(5, 7)$ appear almost identical since they involve the same two genes but different pairs of SNPs. The two SNPs that belonged to the same gene were in linkage disequilibrium (LD) in both cases. It is therefore remarkable that MOSS was able to identify SNPs strongly in LD through the selection of the best models. In some models, the interaction term between the two SNPs was not included. In most instances, the best models include one strong marginal SNP effect (log odds > 1) and a weaker one ($|\log \text{odds}| < 1$), the sign of the coefficient for this latter being either positive (risk effect) or negative (protective effect). In terms of allele frequency, a rare, uncommom and common SNPs correspond to a MAF of $< 5\%$, $\geq 5\%$ and $< 10\%$ and $\geq 10\%$, respectively. Among the eight models detected by MOSS, four of them involve two common SNPs, two include one common and one uncommon SNPs and the last two models entail one rare and one common SNPs. It is therefore of interest that MOSS was able to select these latter two models since most common approaches for GWAS are limited to common SNPs.

[Table 4 about here.]

7.5. *Risk prediction with Bayesian Model Averaging.* The model prediction was obtained by Bayesian model averaging of the eight regression models using 500 iterations of two-fold cross-validation. The area under the ROC

curve (AUC) was estimated to be 63.5%. By comparison, the prediction obtained from the set of best seven common SNPs identified through previous GWAS on breast cancer (based on univariate analysis) was only 57.4% (Gail, 2008). A selection of the best 10 SNPs combined with the major known epidemiological risk factors for breast cancer resulted in an AUC of 61.3% (Wacholder et al., 2010). Therefore, MOSS improves substantially the AUC and the addition of known epidemiological and clinical factors (which were not available for this study) to our model could provide even better predictive ability. MOSS yielded an AUC estimate very similar to that given by other modelling approaches, i.e 63.7%, 64.3% and 63.6% with BEAM3, BVSR and HyperLASSO, respectively.

7.6. *The R package genMOSS.* To run MOSS on the example dataset simuCC dataset we use the function `MOSS_GWAS`:

```
R>MOSS_GWAS(alpha = 1, c = 0.1, cPrime = 0.0001, q = 0.1,
replicates = 5, maxVars = 3, data, dimens, confVars = NULL, k = NULL)
```

The parameters alpha, c, cPrime, and q, have been described in Section 2. Replicates is the number of instances the first stage of the MOSS procedure is run. The top regressions are culled from the results of all the replicates. The parameter maxVars is the maximum number of variables allowed in a regression (including the response). Data is a data frame containing the genotype information for a set of SNPs. It must be organized such that each row refers to a subject and each column to a SNP; the last column in data is interpreted as the binary response. Rows with missing values (i.e., NA's) are ignored. Dimens is the number of possible values for each column in the dataset. In our example, this is three except for the case-control status which is binary. The parameter k is the fold for the cross validation. If k is NULL then only the first stage of MOSS is carried out. Finally, confVars determines the number of confounding variables to be forced to be in every regression (e.g. population stratification variables). In this example, we used the default values for all the parameters (except for k, which is NULL by default, and the parameters data and dimens which, of course, are based on the dataset). A complex *R* code to simulate and analyze genetic data is given in Supplement A.

8. Summary and Discussion. GWAS has emerged as one of the most spectacular advances in genetic research with thousands of novel genetic variants discovered and implicated in many complex human diseases (Hindorff et al., 2009a). Despite this success, the clinical and biological rele-

vance of these findings still remains to be determined. The current challenge in GWAS goes beyond the identification of SNPs that have main effect but also entails the elucidation of more complex genetic mechanisms including SNP by SNP interactions and LD patterns in fine mapping studies. The ultimate goal is to improve the biological relevance of the genetic discoveries. To answer some of these challenges, we proposed a Bayesian graphical model to search for multi-SNP models in the context of GWAS analysis.

Our simulation studies and real data application demonstrate the versatility of MOSS for analyzing complex GWAS data. We showed that MOSS was able to identify genetic variants associated with a binary response in a wide range of association studies where the SNPs could be linked or unlinked, could have main effects and/or interaction effects on the response variable. MOSS can also be applied to fine mapping problems where it can reveal more complex patterns of association with the response. Our simulations showed that MOSS has the best performances overall when compared to more standard approaches for multi-SNP analyses.

Our real application to a breast cancer GWAS data confirms the interest of our novel approach and its relevance for genetic research. We found 12 SNPs embedded in 8 two-SNPs models associated with breast cancer. These two-SNP models included both common and rare variants. We replicated some known associations, e.g. with SNPs in the FGFR2 gene, but also discovered new ones that are biologically very promising. Many of these genetic associations would not have been discovered by conventional approaches, which are generally limited to single SNP analyses or simple multi-SNP models. This is the case of the two SNPs we found associated with the genes APC and BUB3. The association with the SNP in APC and breast cancer has never been reported in the original paper because it is a rare SNP (Hunter et al., 2007). Biological information about BUB3 shows that it interacts physically with APC, thus validating biologically one of the two-SNP models we discovered.

Some future extension of MOSS could include the discovery of complex gene networks. While our results suggest that MOSS can find simple SNP-SNP interactions, further work is needed to infer these more complex networks.

Acknowlegments. We would like to thank Olia Vesselova for her part in the development of the *R* package genMOSS as well as the Associate Editor and the referees for their very constructive comments. This paper is published in memoriam of Dr. Hilmi Ozcelik who inspired a lot this work.

SUPPLEMENTARY MATERIAL

Supplement A: Example of R code
(; Rcode.pdf). This is a simple example of code to run our R package genMOSS.

Supplement B: Complete table 1 results
(; Table1Supp.pdf). This table is similar to table 1 but adds additional FDR results for each of the five SNPs simulated and for the SNP pairwise interactions.

Supplement C: Additional simulation results
(; SimulSupp.pdf). We performed additional simulations to assess the performance of MOSS where it is compared to the standard Bonferroni correction. The R code used to generate the data is given in Supplement A.

Supplement D: Sensitivity analyses
(; SensitivitySupp.pdf). In this section, we assess the sensitivity of the priors to the detection of rare and common genetic variants.

Supplement E: Additional real data analyses
(; RealDataSupp.pdf). This section provides additional results from the real data analysis.

References.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **15** 263-265.

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* **57** 289-300.

Consortium, T. B. C. L. (1999). Cancer risks in BRCA2 mutation carriers. *Journal of the National Cancer Institute* **91** 1310-1316.

Dellaportas, P. and Forster, J. J. (1999). Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical log-linear models. *Biometrika* **86** 615-633.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29** 311-22.

Diamandis, E. P. and Youssef, G. M. (2002). Human Tissue Kallikreins: A Family of New Cancer Biomarkers. *Clinical Chemistry* **48** 1196-1205.

Dobra, A. and Massam, H. (2010). The Mode Oriented Stochastic Search (MOSS) for Log-linear Models With Conjugate Priors. *Statistical Methodology* **7** 240-253.

Edwards, D. E. and Havranek, T. (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika* **72** 339-351.

Gail, M. (2008). Discriminatory accuracy from single nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute* **100** 1037-1041.

Group, A. B. C. S. (2000). Prevalence and penetrance of BRCA1 and BRCA2 in a population based series of breast cancer cases. *The British Journal of Cancer* **83** 1301-1308.

Guan, Y. and Stephens, M. (2011). Bayesian Variable Selection Regression for Genome-wide Association Studies, and other Large-Scale Problems. *Annals of Applied Statistics* To Appear.

Han, B., Park, M. and Chen, X. W. (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics* **11 Suppl.3** S5.

He, Q. and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27** 1-8.

Hindorff, A. L., Junkins, A. H., Hall, N. P., Mehta, P. J. and Manolio, A. T. (2009a). A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.

Hindorff, A. L., Sethupathy, P., Junkins, A. H., Ramos, M. E., Mehta, P. J., Collins, S. F. and Manolio, A. T. (2009b). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. **106** 9362-7.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6** 95-108.

Hoggart, C. J., Whittaker, J. C., De Iorio, M. D. and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* **4** e1000130.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E. and Wacholder, S. e. a. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* **39** 870-874.

Jiang, X., Barmada, M. M. and S., V. (2010). Identifying genetic interactions in genome-wide data using Bayesian networks. *Genet Epidemiol* **34** 575-81.

Kingsmore, S. F., Linquist, I. E., Mudge, J., Gessler, D. D. and Beavis, W. D. (2008). Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews* **7** 221–230.

Kruglyak, L. (2008). The road to genome-wide association studies. *Nature Genetics* **9** 314-318.

Letac, G. and Massam, H. (2012). Bayes regularization and the geometry of discrete hierarchical loglinear models. *The Annals of Statistics* **40** 861-890.

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34** 816-834.

Madigan, D. and Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association* **89** 1535-1546.

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

Massam, H., Liu, J. and Dobra, A. (2009). A Conjugate Prior for Discrete Hierarchical Log-linear Models. *Annals of Statistics* **37** 3431-3467.

McCarthy, M. I. and Hirschhorn, J. N. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Human Molecular Genetics* **17** R156-R165.

on Hormonal Factors in Breast Cancer, C. G. (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. *Lancet* **360** 187-195.

Peto, J. and Mack, T. M. (2000). High constant incidence in twins and other relatives of women with breast cancer. *Nature Genetics* **26** 411-414.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405** 847-856.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273** 1516-1517.

Schwartz, D. F., Ziegler, A. and Konig, I. R. (2008). Beyond the results of genome-wide association studies. *Genet Epidemiol* **32** 671.

Thomas, A. and Camp, N. J. (2004). Graphical Modelling of the joint Distribution of Alleles at Associated Loci. *Am. J. Hum. Genet.* **74** 1088-1101.

Thompson, D. and Easton, D. F. (2002). Cancer incidence in BRCA1 mutation carriers. *Journal of the National Cancer Institute* **94** 1358-1365.

Thompson, D. and Easton, D. F. (2004). The genetic epidemiology of breast cancer genes. *Journal of Mammary Gland Biolology and Neoplasia* **9** 221-236.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267-288.

Ungvari, I., Hullam, G., Antal, P., Kiszel, P. S., Gezsi, A., Hadadi, E., Virg, V., Hajós, G., Millinghoffer, A., Nagy, A., Kiss, A., Semsei, . F., Temesi, G., Melegh, B., Kisfali, P., Széll, M., Bikov, A., Gálffy, G., Tamsi, L., Falus, A. and Szalai, C. (2012). Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. *PLoS One* **7** e33573.

Verzilli, C. J., Stallard, N. and Whittaker, J. C. (2006). Bayesian Graphical Models for Genomewide Association Studies. *American Journal of Human Genetics* **79** 100-112.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M. and Feigelson, e. a. H. S. (2010). Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine* **362** 986-993.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C. and Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *Annals of Applied Statistics* **4** 1342-1364.

Wu, Z. and Zhao, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet.* **5** e1000582.

Xing, H., McDonagh, P. D., Bienkowska, J., Cashorali, T., Runge, K., Miller, R. E., Decaprio, D., Church, B., Roubenoff, R., Khalil, I. G. and Carulli, J. (2011). Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Comput Biol.* **7** e1001105.

Yeung, K. Y., Bumgarner, R. E. and Raftery, A. E. (2005). Bayesian Model Averaging: Development of an Improved Multi-class, Gene Selection and Classification Tool for Microarray Data. *Bioinformatics* **21** 2394-2402.

Zhang, Y. (2012). A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genet Epidemiol* **36** 36-47.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39** 1167-73.

Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute,
60, Murray Street, Toronto, ON, M5T 3L9, Canada
E-mail: laurent@lunenfeld.ca

List of Tables

TABLE 1
Simulation results: False Discovery Rate (FDR) and True Discovery Rate (TDR) (in %) in Simulation Scenarios 1 to 3

		FDR[1]	TDR[2]	AUC
Scenario	**Method**			
1	MOSS	15.7	56.8	61.2
1	LASSO	19.4	29.0	58.0
1	BVSR	16.8	62.2	59.6
1	BEAM3	30.0	31.2	58.3
1	BONFmain[3]	1.0	38.0	58.4
1	FDRmain[4]	15.8	54.4	61.1
2	MOSS	5.7	73.7	59.8
2	LASSO	2.5	39.2	57.7
2	BVSR	5.1	18.5	53.2
2	BEAM3	5.6	46.2	56.1
2	BONFmain	1.8	33.2	57.8
2	FDRmain	6.1	45.7	58.3
3	MOSS	1.0	100.0	56.6
3	LASSO	47.0	36.0	53.9
3	BVSR	28.1	94.0	57.1
3	BEAM3	16.7	100.0	56.6
3	BONFmain	23.6	100.0	56.7
3	FDRmain	30.5	100.0	57.2

[1] FDR_m: FDR for SNP main effects
[2] TDR_m: TDR for SNP main effects; TDR_{pair}: TDR for a pair of SNPs; TDR_{all}: TDR for SNP main effects over all the SNPs
[3] TDR corresponding to a $\chi^2(1)$ statistic and Bonferroni corrected p-value
[4] TDR corresponding to a $\chi^2(1)$ statistic and FDR corrected p-value

TABLE 2
Simulation results: False Discovery Rate (FDR) and True Discovery Rate (TDR) (in %) in Simulation Scenarios 4 & 5

		False Discovery Rate			True Discovery Rate			AUC
		FDR_m^1	FDR_c^1	FDR_{lc}^1	TDR_m^2	TDR_c^2	TDR_{lc}^2	
4	MOSS	10.0	5.8	3.6	96.3	97.0	98.3	68.5
4	LASSO	18.0	13.4	10.9	89.0	90.3	90.7	68.5
4	BVSR	28.8	23.4	17.8	87.0	87.3	87.7	67.8
4	BEAM3	46.2	19.6	15.8	59.7	83.3	86.0	67.9
4	BONFmain	92.2	79.5	57.0	100.0	100.0	100.0	70.7
4	FDRmain	73.6	61.7	48.5	86.6	86.6	86.6	68.3
5	MOSS	16.1	4.9	3.6	93.7	96.0	96.7	68.6
5	LASSO	18.4	13.4	10.3	93.0	94.7	94.7	68.9
5	BVSR	14.9	12.0	6.2	85.3	85.3	85.6	67.5
5	BEAM3	20.9	11.8	10.7	53.0	54.7	55.0	65.1
5	BONFmain[3]	91.3	77.8	57.2	100.0	100.0	100.0	70.5
5	FDRmain[4]	38.7	9.1	4.5	6.3	6.7	6.7	52.5

[1] FDR_m: FDR for SNP main effects; FDR_c: FDR for the cluster of SNPs; FDR_{lc}: FDR for the lenient cluster of SNPs
[2] TDR_{all}: TDR for SNP main effects; TDR_c: TDR for the cluster of SNPs; TDR_{lc}: TDR for the lenient cluster of SNPs
[3] Bonferroni corrected p-value based on a $\chi^2(2)$ statistic for the main effects
[4] FDR corrected p-value based on a $\chi^2(2)$ statistic for the main effects

TABLE 3. *SNPs with the highest posterior probability found by MOSS based on two-locus models from the breast cancer GWAS data*

SNP ID	Chr #	Location (in Kb)	Allele1	Allele2	MAF[1]	p-value[2]	log Bayes factor[3]	MOSS post probability[4]	Rank[5] (p-value)	Rank[6] (MOSS)	Closest gene[7]	Other related genes	Function[8]
rs3130544	6	31167	C	A	0.15	1.7×10^{-4}	2.50	0.349	76	1	C6orf15 (20 kb)	DPCR1, CDNA, TCF19	-
rs10510126	10	124992	C	T	0.14	2.0×10^{-6}	4.50	0.281	1	2	BUB3 (80 kb)	-	-
rs2249938	20	60857	A	G	0.14	1.6×10^{-4}	2.63	0.211	70	3	NTSR1 (0 kb)	-	Yes
rs2274352	10	13742	C	T	0.08	2.2×10^{-3}	1.91	0.192	1214	4	FRMD4A (0 kb)	PRPF18	Yes
rs722936	4	91190	G	T	0.19	9.8×10^{-4}	2.19	0.178	544	5	KIAA1680 (<60 kb)	-	-
rs17344557	5	45613	C	T	0.14	8.8×10^{-3}	0.60	0.178	4844	6	HCN1 (0 kb)	-	Yes
rs2427448	20	60866	C	T	0.11	3.6×10^{-4}	2.28	0.138	184	7	NTSR1 (1 kb)	-	-
rs16910213	8	83400	G	A	0.0	NA	0.85	0.131	NA	8	HNRNPA1P4 (30 kb)	-	-
rs1219648	10	123336	A	G	0.42	1.2×10^{-5}	4.10	0.120	4	9	FGFR2 (0 kb)	-	-
rs1882619	5	112107	T	C	0.07	0.79	0.88	0.083	421682	10	APC (0 kb)	SRP19, REEP5	-
rs2420946	10	123341	C	T	0.40	1.5×10^{-5}	4.00	0.072	6	11	FGFR2 (0 kb)	-	Yes
rs6995588	8	61167	C	T	0.0	NA	0.80	0.067	NA	12	CA8 (100 kb)	-	-

[1]Minor allele frequency (= Allele1) in the Hapmap European population

[2]p-value from single marker test using logistic regression adjusted for age, region, three main population stratification principal components. NA means that the SNP was not analyzed due to MAF=0

[3]Bayes Factor from single marker test

[4]MOSS posterior probability normalized with respect to the models retained in our list of top models (See table 2)

[5]Rank based on single-marker p-value. NA means that the SNP was not analyzed due to MAF=0

[6]Rank based on MOSS posterior probability for each single marker

[7]Closest gene based on physical distance

[8]Function is associated with transcription regulatory mechanisms

TABLE 4. *Best models with the highest posterior probability found by MOSS based on two-locus models from the breast cancer GWAS data*

Model	Log marginal log likelihood	Bayes Factor	SNP1 (MAF)	Gene 1 SNP1	SNP2 (MAF)	Gene 2 SNP2	log odds[2] SNP1	log odds[2] SNP2	log odds[2] SNP1 × SNP2
1	-15239.10	14.01	rs3130544 (0.15)	C6orf15	rs2249938 (0.14)	NTSR1	1.91 [0.94; 3.07]	-0.40 [-0.59; -0.21]	-
2	-15239.27	13.84	rs722936 (0.36)	KIAA1680	rs17344557 (0.14)	HCN1	0.36 [0.17;0.54]	1.73 [0.61; 3.16]	-
3	-15239.52	13.59	rs3130544 (0.15)	C6orf15	rs2427448 (0.11)	NTSR1	1.88 [0.96; 3.08]	-0.38 [-0.57; -0.18]	-
4	-15239.58	13.53	rs16910213 (0.0)	HNRNPA1P4	rs10510126 (0.14)	BUB3	2.58 [0.96; 5.20]	-0.52 [-0.73; -0.32]	-
5	-15239.66	13.45	rs1219648 (0.42)	FGFR2	rs2274352 (0.08)	FRMD4A	1.22 [0.85; 1.62]	0.55 [0.36; 0.73]	-1.26 [-1.78; -0.81]
6	-15240.03	13.08	rs1882619 (0.07)	APC	rs10510126 (0.14)	BUB3	5.28 [1.45; 17.38]	-0.51 [-0.72; -0.29]	-
7	-15240.17	12.94	rs2274352 (0.08)	FRMD4A	rs2420946 (0.40)	FGFR2	1.21 [0.82; 1.61]	0.54 [0.34; 0.72]	-1.25 [-1.71; -0.77]
8	-15240.24	12.87	rs6995588 (0.0)	CA8	rs10510126 (0.14)	BUB3	5.35 [1.97; 15.15]	-0.52 [-0.73; -0.32]	-

[1] Log marginal likelihood normalized across all the models retained in the selected list

[2] Bayesian posterior log odds estimate and 95% interval estimate based on Bayesian Iterative Proportional Fitting