Modeling Genetical Data with Forests of Latent Trees for Applications in Association Genetics at a Large Scale
Which Clustering Method should Be Chosen?

D.-T. Phan[1], P. Leray[1] and C. Sinoquet[2]
[1]*LINA / UMR CNRS 6241, Polytech/ University of Nantes, rue Christian Pauc, BP 50609, 44306 Nantes, France*
[2]*LINA / UMR CNRS 6241, Faculty of Sciences, University of Nantes, 2 rue de la Houssinière, 44322 Nantes, France*

Keywords: Linkage Disequilibrium, Genome-wide Association Study, Multilocus Association Study, Data Dimension Reduction, Probabilistic Graphical Model, Bayesian Network.

Abstract: Association genetics, and in particular genome-wide association studies (GWASs), aim at elucidating the etiology of complex genetic diseases. In the domain of association genetics, machine learning provides an appealing alternative framework to standard statistical approaches. Pioneering works (Mourad et al., 2011) have proposed the forest of latent trees (FLTM) to model genetical data at the genome scale. The FLTM is a hierarchical Bayesian network with latent variables. A key to FLTM construction is the recursive clustering of variables, in a bottom up subsuming process. In this paper, we study the impact of the choice of the clustering method to be plugged in the FLTM learning algorithm, in a GWAS context. Using a real GWAS data set describing 41400 variables for each of 3004 controls and 2005 individuals affected by Crohn's disease, we compare the influence of three clustering methods. Data dimension reduction and ability to split or group putative causal SNPs in agreement with the underlying biological reality are analyzed. To assess the risk of missing significant association results through subsumption, we also compare the methods through the corresponding FLTM-driven GWASs. In the GWAS context and in this framework, the choice of the clustering method does not impact the satisfying performance of the downstream application, both in power and detection of false positive associations.

1 INTRODUCTION

With the finalization of the Human Genome Project in 2003, it was confirmed that any two individuals share, on average, 99.9% of their genome with one another. It is then the sole 0.1% genetic variations that may explain why individuals are physically different or should inherit a greater risk of contracting genetic disorders, such as coronary heart disease, diabetes, autism, some cancers. As a consequence, identifying the genetic factors underlying these diseases potentially plays a crucial role in prediction, monitoring subjects with risks, as well as developing new treatments. Deciphering the putative causes of complex genetic diseases has been one of the main focuses of human genetics research during the last thirty years. Among different approaches that have been proposed, association studies stand out as one of the most successful paths, even though their potential is yet to be fully tapped.

The HapMap Project (Gibbs et al., 2003) and its successor, the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), were launched with the hope to establish a catalogue of human genome regions in which people of different populations have differences.

When no clue is available about the genome regions likely to contain one of the putative causes for a studied disease, geneticists are compelled to resort to genome-wide association studies (GWASs). Genetic markers are used for this purpose, as well as a population of affected and unaffected individuals. Genetic markers represent as many DNA sequences, spread over the whole genome, with a known location, where the DNA variations within a given population may be observed. In a nutshell, GWASs seek to identify genetic markers whose variants vary systematically between affected (cases) and unaffected (controls) individuals (Balding, 2006). The standard GWAS consists in comparing variant frequencies in cases and controls, on massive genotypic datasets (tens of thousands of individuals each described by

Phan D., Leray P. and Sinoquet C..
Modeling Genetical Data with Forests of Latent Trees for Applications in Association Genetics at a Large Scale - Which Clustering Method should Be Chosen?.
DOI: 10.5220/0005179800050016
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 5-16
ISBN: 978-989-758-070-3

hundreds of thousands up to a few millions of genetic markers). The goal is to identify the loci on the genome for which the distributions of variants are significantly different between cases and controls, using dependence - namely *association* - tests (e.g. the Chi^2 test). The unit variants, called single nucleotide polymorphisms (SNPs), which refer to single base pair changes in the DNA sequence, represent the most abundant type of variants in human; they are very often used as markers in GWASs.

The key to GWASs lies in this interesting phenomenon known as "linkage disequilibrium" (LD) where variants for different SNPs tend to co-occur non-randomly (Pritchard and Przeworski, 2001) (the corresponding SNPs are said to be in LD). The case would be exceptional if a genetic marker, which is observed in the population, coincided with a genetic causal factor. Nevertheless, thanks to LD, a dependence exists between the non observed causal factor and a genetic marker nearby the former. On the other hand, by definition, a dependence exists between the causal factor and the disease of interest. Therefore, it is likely that a dependence will be detected between the nearby genetic marker and the disease.

In the human genome, the HapMap project confirmed evidence of the linkage disequilibrium, this latent structure organized in the so-called "haplotype blocks". Therein, regions showing high dependences between contiguous markers (blocks) alternate with shorter regions characterized by low statistical dependences. In general, LD exhibited among physically close loci is stronger than LD between SNPs that are farther apart. In other words, LD decays with distance.

However, standard GWASs do not fully exploit LD. Some authors proposed to test combinations of SNPs - haplotype blocks - against the disease, rather than merely each SNP against the disease: this is the principle of multilocus approaches. First, if the causal SNP has low frequency and is not in high LD with any one of the genotyped SNPs, then the multilocus test will tend to be more powerful. Besides, the advantage to the GWAS is that the LD is likely to reveal an excess of haplotype sharing around a causative locus, amongst cases. Third, testing haplotypes instead of SNPs is a way to implement data dimension reduction. In this context, fine LD modeling at genome scale is required.

Few works have focused on LD modeling at genome scale, which is a challenging task. The proposals of (Abel and Thomas, 2011) and (Verzilli et al., 2006) both rely on the use of Markov random fields, a popular kind of probabilistic graphical models. Two scalable models designed for the specific purpose of multilocus GWASs have been described by (Browning and Browning, 2007) and (Mourad et al., 2011). The approach in (Browning and Browning, 2007) relies on a variable length Markov chain (VLMC), a Markov model where the size of the memory conditioning the prediction of the variant at a given location is flexible. In contrast with this block-based method, the works in (Mourad et al., 2011) seek to subsume clusters of SNPs through latent variables. SNPs within the same cluster are not necessarily contiguous. Such latent variables are intented to be tested against the disease. Both methods account for the fuzzy nature of LD since block boundaries are not accurately defined over the genome. However, being blocked-based, the method in (Browning and Browning, 2007) cannot take into account long-range dependences. Moreover, LD is intrinsically hierarchical, with clusters of SNPs recursively structured in clusters of lower and lower correlated SNPs. To attempt a faithful representation of LD upstream of a GWAS, hierarchical clustering is one of the key ingredients of the learning algorithm of the Bayesian model used in (Mourad et al., 2011). Since clustering is central to learning the model in (Mourad et al., 2011), namely the forest of latent tree models (FLTM), this paper analyses the impact of the choice of the clustering method in a GWAS context.

2 OBJECTIVES AND ORGANIZATION OF THE PAPER

In the remainder of this paper, data partitioning - or clustering - denotes the generation of a set of non overlapping clusters. Such a task is highly complex. Though, choosing a clustering method to learn an FLTM must comply with the scalability goal. This paper compares the native clustering method used in (Mourad et al., 2011) ($CAST_{bin}$) with a relaxed version ($CAST_{real}$) and another clustering method (DBSCAN). In this framework, two aims of the paper are to evaluate whether FLTM learning is robust to the choice of the clustering method and how close a clustering method approximates the underlying biological reality. To fulfill the first goal, a protocol is used that relies on assessing how much two partitions agree. The second objective is met by applying the previous protocol to compare each clustering method to a reference partition supposed to be close to biological reality. The Haploview software program is the tool chosen to derive such a reference partition. Focusing on the data dimension reduction aspect, a third objec-

tive of the paper is to analyze the impact of the choice of the clustering method on data subsumption quality. By construction, an FLTM-based GWAS processes data subsumed through latent variables, to hopefully pinpoint the interesting regions of a genome without testing each SNP for association. Thus, the third objective of this paper is to assess whether the choice of the clustering method impacts the risk of missing significant association results through subsumption. FLTM-driven GWASs are run to study this impact.

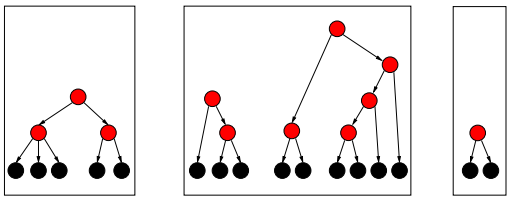
The remainder of the paper is organized in five sections. Section 3 first offers a brief introduction to Bayesian networks, the kind of probabilistic graphical models FLTM is based upon. Then section 3 provides a broad brush description of the FLTM learning algorithm together with a sketch of a GWAS strategy based on FLTM. Section 4 briefly refers to the native clustering method used in FLTM ($CAST_{bin}$) and to its relaxed version ($CAST_{real}$); it then motivates the choice of the alternative clustering method (DBSCAN) plugged in the FLTM learning algorithm. Then, section 5 explains the design of the protocols and methods used in our work. First, we discuss the protocol used to assess how much two partitions agree. Second, we justify the use of the Haploview software program to derive the reference partition, supposedly the closest representation of the underlying reality. In section 6, we describe the Crohn's disease GWAS data used in our study. Section 7 is devoted to the presentation and discussion of the results observed.

3 FRAMEWORK AND FLTM MODEL

The FLTM model is a tree-structured Bayesian network (BN). Therefore this section first briefly introduces Bayesian networks, to further focus on the FLTM model. The principle of the FLTM learning algorithm is then presented. Finally, the principle for a multilocus GWAS based on the FLTM model is sketched.

3.1 A Brief Reminder about Probabilistic Graphical Models

When probabilistic graphical models are learnt from scratch, one has to learn their two fundamental components from a data matrix. In this matrix, the lines correspond to the observations and the columns correspond to the variables X_i ($1 \leq i \leq n$). For example, in the case of genetical data, the observations



(a) (b) (c)

Figure 1: (a) Latent tree model (LTM). (b) Forest of latent tree model (FLTM). (c) Latent class model. Observed and latent variables are represented in dark and light color shades, respectively.

are the individuals (cases and controls) in the population studied, and the variables are the SNPs. The qualitative component of a BN is a graph where the variables are represented as nodes. The connections between the nodes represent the direct dependences between the variables. More specifically, the qualitative component of a BN is a directed acyclic graph. The quantitative component of a BN is a collection of probability distributions, denoted as "the parameters θ". If the variable X_i has no parent in the graph, then θ_i is merely an *a priori* distribution $(\theta_i = \mathbb{P}(X_i))$. If the variable X_i has a set of parents Pa_{X_i}, then θ_i is the conditional distribution $\theta_i = \mathbb{P}(X_i \mid Pa_{X_i})$. In particular, Bayesian networks offer a practicable framework: exploiting the network structure, this framework allows to compute the joint probability of the variables, $\mathbb{P}(X)$, as a product of low-dimensional functions.

It may happen that the data observed is thought to embed a latent structure, depicted through latent variables and their connections in the learnt BN. In this case, learning the Bayesian network encompasses the task of inferring the latent variables, and their connections within the BN.

The FLTM model is a forest of latent tree models (LTMs). The Figure 1(a) shows that an LTM is characterized by a hierarchical structure organized in layers. The first layer is composed of the observed variables. The other layers are composed of latent variables. The learning algorithm of the FLTM (see Figure 1(b)) relies on the simplest LTM that may be described, the latent class model (LCM). A latent class model connects a single latent variable to child variables ; no connections are allowed between the latter (see Figure 1(c)).

3.2 Sketch of the FLTM Learning Algorithm

Learning a BN is a hard task that consists in inferring both the graph structure and the parameters. Learning a BN with latent variables is far more complicated. First, one does not even know how many latent

variables have to be inferred. Second, in a BN without latent variables, the parameters are estimated to maximize the likelihood, that is the probability of the (observed) data given the parameters. In contrast to this rapid algorithm, a slow procedure has to be employed for BNs with latent variables, the expectation-maximization algorithm dedicated to learn parameters in the case of missing data. Prior knowledge (the hierarchical LD structure) is used by the specific procedure described in (Mourad et al., 2011), to provide a scalable learning algorithm. Figure 2 depicts the principle of this iterative algorithm, based on an ascending hierarchical clustering procedure.

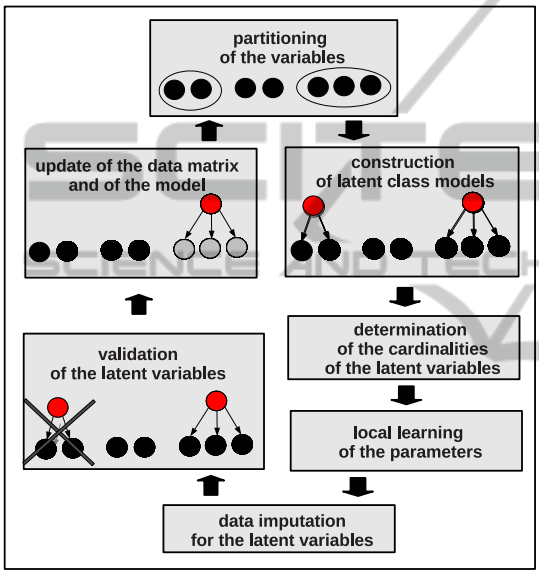


Figure 2: Sketch of the iterative FLTM learning algorithm.

In the case of LD modeling, the observed variables are the SNPs. The cardinality of these observed variables is equal to 3, which codes for minor homozygosity, heterozygosity and major homozygosity. The first iteration starts with the partitioning of the observed variables into non overlapping clusters of pairwise highly dependent SNPs. No two variables are allowed in the same cluster if their physical distance on the genome is above a given threshold, δ. For each cluster, an LCM is constructed whose child variables are the variables in the cluster and whose latent variable is created. The approach of (Mourad et al., 2011) considers discrete latent variables whose cardinalities may be different from one another; a heuristic is used to determine the specific cardinality of each latent variable. This specific cardinality is computed as an affine function of the number of variables in the cluster. Then, the parameters of each LCM are estimated through the standard expectation-maximization procedure. Knowing the parameters of each LCM further allows to impute the data corresponding to its latent variable. Such imputed data are used by a validation step; this step relies on a normalized mutual information criterion to examine whether each novel latent variable is sufficiently informative to subsume its child variables. The data is updated with the validated latent variables replacing their child variables. The validated latent variables can thus be considered as observed variables and an iteration begins anew.

This ascending process is iterated until no valid cluster can be identified, or a single cluster of maximal size is obtained. Among other parameters of the learning procedure, the clustering procedure plugged in the algorithm is likely to impact the quality of the LD modeling, and therefore the quality of the GWAS performed downstream.

3.3 Performing a GWAS Guided by the FLTM Model

Central to the use of the FLTM model for a GWAS purpose are the request for data dimension reduction and the motivation for a multilocus strategy. In the study described here, we have implemented a multilocus GWAS strategy as follows: in the lowest layers, we traverse the forest top down, following a best-first search strategy which only tests all child nodes for the nodes whose association significances (i.e. p-values) are below a threshold. The way to compute this threshold is specific to the layer the variable belongs to. These child nodes are selected in turn with respect to the appropriate threshold. The standard Chi^2 test is applied to test a variable against the disease and to provide a pointwise p-value. We now explain how to compute the thresholds for the variables in the lowest layers. To test statistical significance, we consider a global threshold, say $\alpha = 5\%$. The pointwise p-value of a variable is corrected based on permutations applied on all the variables in this variable's latent layer. Thus, the correction is layer specific. Statistical significance is assessed by testing the condition ($\text{p-value}_{corrected} \leq \alpha$). The details and justification for this correction will be provided in an extended version of the article. On the other hand, due to dimension reduction, the highest layers have a low number of variables. No correction is applied for these layers, from which we systematically select the top most associated variables (e.g. $\beta = 10\%$).

4 THE CAST AND DBSCAN PARTITIONING METHODS

Performing an optimal clustering is NP-hard (Ackerman and Ben-David, 2009). Therefore, heuristics must be designed instead. In this paper, we focus on two partitioning methods, CAST and DBSCAN, to study how they impact LD modeling and a further downstream GWAS analysis.

Since we address high-dimensional data, we could not envisage the use of ascending hierarchical clustering, whose complexity scales in $O(n^3)$ where n is the number of objects to be asssigned to clusters. Moreover, in contrast to AHC and k-means, another well known partitioning method, CAST and DBSCAN do not request the tuning of the number of clusters.

Partitioning objects into clusters relies on pairwise distances (alternatively pairwise similarities). Storing a pairwise similarity matrix at the genome scale is intractable. Thus, following (Mourad et al., 2011), we acknowledge a physical constraint, δ, expressed in *kbp* (kilobase pairs), in both implementations of the CAST and DBSCAN methods. This constraint δ represents the physical distance on the genome beyond which two objects (in our case two variables) are not allowed in the same cluster. Additional calculus is required to estimate the distance between two variables one of which at least is a latent variable.

The CAST (Cluster Affinity Search Technique) algorithm was proposed in (Ben-Dor et al., 1999) and is depicted in (Cahill, 2002). Its theoretical runtime complexity scales in $O(n^2\ (log(n))^c)$, with c some constant, and its empirical complexity allows to handle high-dimensional data. CAST is the native clustering method used in the FLTM learning algorithm depicted in (Mourad et al., 2011).

To decide cluster membership, CAST relies on an affinity measure. In the implementation of CAST adapted to FLTM learning, the binary similarity measure is assessed as the thresholded mutual information (MI). A parameter $q_{pairwise}$ (e.g. 50%) allows to compute the MI quantile (e.g. median) over the pairs of variables whose physical distance is below δ. This quantile allows to assign a binary similarity $(0/1)$, as in the native FLTM learning algorithm. In this study, we also consider the unthresholded version. These two CAST versions are denoted CAST_{bin} and CAST_{real}.

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was proposed in (Ester et al., 1996). Its theoretical runtime complexity is $O(n^2)$, where n is the number of objects to be assigned to clusters. However, its empirical complexity is known to be lower. To grow clusters, DBSCAN relies on the merging of objects' neighborhoods. This method requires two parameters: R, the maximum radius of the neighborhood to be considered, and N_{min}, the minimum number of neighbors needed for a cluster.

We have chosen DBSCAN as it is resistant to noise. Besides, DBSCAN is known to be able to identify a cluster embedded in another cluster. On the genome line, long-range LD corresponds to this situation.

5 METHODS

In this section, we first present the protocol used to evaluate how much two partitions agree when focusing on the top most associated SNPs found by a GWAS. Then, we motivate how we derived the so-called reference partition (to be further defined). This methodological section ends with the presentation of the protocol used to compare the impact of the choice of the partitioning method on the subsequent GWAS. For this purpose, we rely on GWAS results published in the literature.

5.1 Comparing Two Partitions

To cope with the genome scale, we were compelled to select a simple method: we focused on the comparison of the partitions respectively obtained for the first layer (SNPs) by two partitioning methods, and we examined how the top most associated SNPs identified by a GWAS are distributed among the clusters.

The methods dedicated to the comparison of two partitions may be categorized into three main groups (Meila, 2005). Two groups attempt to map a partition onto the other, either from set matching functions, or from information theory-centered methods. The third category relies on counting for how many pairs of elements two partitions agree or disagree. The FLTM-driven GWAS strategy is a multilocus strategy by definition. In this multilocus GWAS framework, it is relevant to analyze pair agreement between two partitioning methods, for a selection of top most associated SNPs. A counting method fits well this purpose of focusing on a subset of SNPs.

Given two partitions over the same set of objects, and a pair of objects (in our case, a pair of variables), an agreement means that the two partitions both group the two variables in a cluster or both assign two different clusters to the two variables. Given two partitions P_1 and P_2, let

- N_{11}, the number of pairs both partitions assign to one cluster,

- N_{00}, the number of pairs both partitions assign to different clusters,
- N_{10}, the number of pairs kept in the same cluster by P_1 but splitted by P_2,
- N_{01}, the symmetric case of the latter.

From here, a large set of comparison measures is available. We selected three measures to perform the following comparisons: CAST_{bin} versus CAST_{real}, CAST_{bin} versus DBSCAN, CAST_{real} versus DBSCAN, and each of the three methods CAST_{bin}, CAST_{real} and DBSCAN versus the reference partition (to be defined in section 5.2). The comparison measures selected are:

- the Rand index (Rand, 1971):

$$RI = \frac{N_{11} + N_{00}}{N_{11} + N_{00} + N_{10} + N_{01}} \quad (1)$$

for which we used instead an adjusted corrected-for-chance version ($ARI = \frac{RI - expected\ RI}{maximum\ RI - expected\ RI}$) (for the detailed description, see (Hubert and Arabie, 1985));

- the Mirkin distance (Mirkin, 1998):

$$MI = \frac{S_{P_1} + S_{P_2} - 2S_{P_1P_2}}{n^2} \quad (2)$$

with

$$S_{P_j} = \sum_{cluster_i \in P_j} |\, cluster_i \,|^2, j = 1,2$$

$$S_{P_1P_2} = \sum_{cluster_i \in P_1,\ cluster_j \in P_2} |\, cluster_i \,|\, |\, cluster_j \,|$$

and n the number of objects to be assigned to clusters and $|\, S \,|$ the size of set S;

- the Fowlkes-Mallows index (Fowlkes and Mallows, 1983):

$$FM = \sqrt{\frac{N_{11}}{N_{11} + N_{10}} \cdot \frac{N_{11}}{N_{11} + N_{01}}}. \quad (3)$$

5.2 Deriving the Reference Partition

The reference partition intends to be the closest representation of the underlying reality, that is the haplotype blocks. We used the Haploview software program (Gabriel et al., 2002) for this purpose. This application allows to select commonly used block definitions to partition the genome into regions of strong LD (Gabriel et al., 2002; Wang et al., 2002). As this block generation is dedicated to handle genetical data, Haploview can only be used for the first layer (observed variables). This reason explains why the partitioning method of the Haploview application has not been plugged in the FLTM learning algorithm.

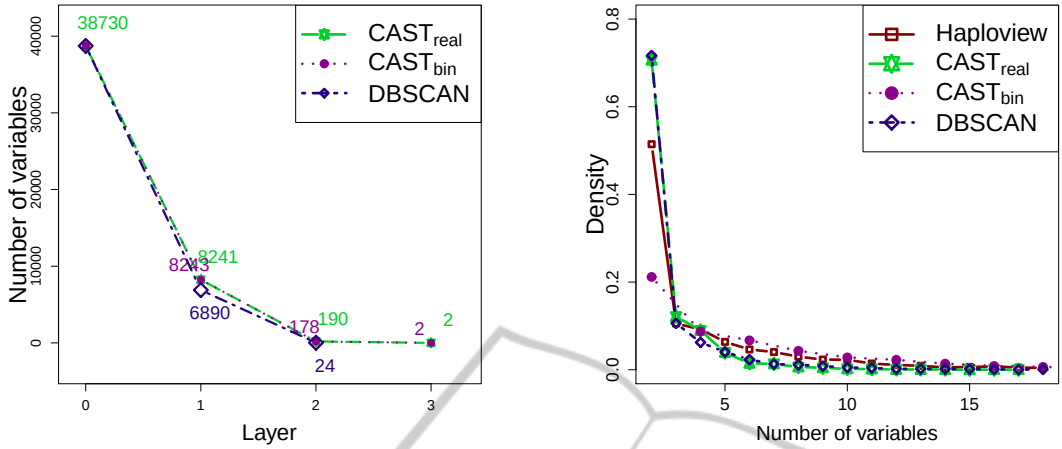
6 CROHN'S DISEASE GWAS DATA

The Crohn's disease data set we used is made available by the WTCCC Consortium (http://www.wtccc.org.uk/); it consists of 5009 individuals genotyped using the Affymetrix GeneChip 500K Mapping Array Set (3004 controls, 2005 cases). We performed the same data quality control as the WTCCC. We excluded individuals, using exactly the same criteria as the WTCCC ((WTCCC, 2007), page 26) (e.g. individuals with more than 3% missing data across all SNPs; individuals sharing more than 86% of identity with other ones). The rules to exclude SNPs were also modelled after those of the WTCCC (e.g. missing rate over 5%; if MAF (minor allele frequency) under 5%, missing rate threshold decreased to 1%) ((WTCCC, 2007), page 27).

In this paper, we focus on chromosome 2, known to harbour SNPs with susceptibility towards Crohn's disease. The initial WTCCC data set describes 41400 SNPs. After the quality control step, our data consisted of 38730 SNPs.

7 RESULTS AND DISCUSSION

The parameter t_{CAST} (see details in (Cahill, 2002)) specific to the CAST method, whatever the version (*bin* or *real*), was empirically set to 0.50. The parameter $q_{pairwise}$ specific to the CAST_{bin} clustering method was empirically chosen to be 50%. The N_{min} and R parameters specific to DBSCAN were tuned to 2 and 0.2 respectively. The FLTM learning algorithm requires the setting of six parameters. We systematically evaluated the coefficients of the affine function used to determine the cardinality of each latent variable, ℓ_1 and ℓ_2, in $[0.2,\ 0.3,\ 0.4,\ 0.5] \times [1,\ 2]$. We observed no differences between the eight settings, with regard to the sizes and contents of the clusters. Thus, ℓ_1 and ℓ_2 were set 0.5 and 1. Following (Mourad et al., 2011), we fixed the maximum cardinality as 20, the physical distance constraint δ as 45 *kbp* and the number of restarts of the stochastic expectation-maximization procedure as 10. The threshold for the quality control of the candidate latent variables was set to a low value, 0.01. The GWAS thresholds α and β were fixed to 5% and 10%. The study was conducted using a 3.3 GHz processor. We had to adapt the generic versions of the CAST and DBSCAN algorithms, to store a sparse similarity matrix instead of a pairwise similarity matrix (see section 4).



(a) (b)

Figure 3: Impact of the choice of the partitioning methods $CAST_{bin}$, $CAST_{real}$ and DBSCAN on the structure of the FLTM model. (a) Impact on the number of variables per layer. (b) Impact on the sizes of the clusters for the first layer (observed layer).

7.1 FLTM Architectures

On average, the running time observed for FLTM learning with each clustering method is in the order of 60 hours. A closer examination shows that clustering and other operations only required at most 1 minute for each layer, and that practically all the running time was spent in the expectation-maximization procedure (see section 3.2). Moreover, it is likely that the presence of a few clusters of large size (size up to 50) severely increases running times for the expectation-maximization procedure.

We first analyze the impact of the partitioning method on the structure of the FLTM model constructed prior to a GWAS. Figure 3 (a) compares the impacts of the three partitioning methods on the data dimension reduction. We observe that for any layer, the total number of latent variables created using $CAST_{real}$ is always greater than that created using DBSCAN. Moreover, layer 3 does not exist for DBSCAN whereas it exists for $CAST_{bin}$ and $CAST_{real}$. Indeed, for DBSCAN, no more variables can be grouped in layer 2: all candidate clusters are singletons. The numbers of variables in layers 1 and 3 are either very close or similar between $CAST_{bin}$ and $CAST_{real}$. Again, among the three methods, the numbers of variables in layer 2 are the closest for $CAST_{bin}$ and $CAST_{real}$.

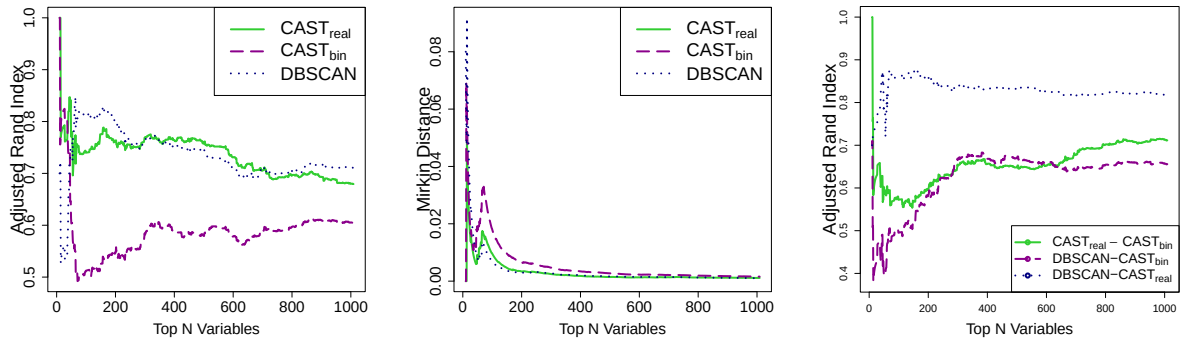
Figure 3 (b) provides the histogram for the sizes of the clusters in the first (observed) layer, for each of the three partitioning methods, together with the histogram of the reference Haploview partitioning. It has to be mentioned that, for reasons of presentation, the histograms have been truncated. Very few clusters of large sizes are observed: the maximum sizes observed are 18, 45 and 50 for $CAST_{real}$, DBSCAN and $CAST_{bin}$, respectively. Such clusters would normally appear far apart on the right section of Figure 3 (b).

First, we observe that from size 3, the $CAST_{bin}$ curve is slightly above the $CAST_{real}$ and DBSCAN curves. Besides, the latter curves are nearly superimposed. Finally, we note that from size 3, the curve relative to the reference partitioning is located slightly below that of $CAST_{bin}$, on the one hand, and slightly above the quasi superimposed curves of $CAST_{real}$ and DBSCAN, on the other hand.

Therefore, the general conclusion to draw for this section is the propensity for DBSCAN to produce a lower number of variables than $CAST_{bin}$ and $CAST_{real}$, but with no clear impact on the differences between the cluster size histograms.

7.2 Comparison of the Partitioning Methods in a GWAS Context

In a GWAS context, we wish to focus in priority on pairs of SNPs selected among the top SNPs found most associated with the studied disease. The standard tool PLINK was used to identify these top SNPs (Purcell et al., 2007) (http://pngu.mgh.harvard.edu/purcell/plink/). Relying on PLINK, we performed a single-SNP GWAS on the WTCCC data set relative to chromosome 2. The association test used was the Chi^2. We have extended the agreement analysis of two partitions to embedded



(a) (b) (c)

Figure 4: Agreement of two partitioning methods, in a GWAS context. (a) and (b) Agreement of a partitioning method with the reference block partitioning method used by Haploview. Comparison for the partitioning methods CAST$_{bin}$, CAST$_{real}$ and DBSCAN. Impact of the number of top SNPs considered on the agreement. The top SNPs considered are those found most significantly associated by a standard single-SNP GWAS. (a) Adjusted Rand index. (b) Mirkin distance. (c) Pairwise comparison of the partitioning methods CAST$_{bin}$, CAST$_{real}$ and DBSCAN. Impact of the number of top SNPs considered on the agreement. Adjusted Rand index.

sets of associated SNPs, increasing the size of the set of top associated SNPs up to 1000.

Figures 4 (a) and (b) compare the partioning methods CAST$_{bin}$, CAST$_{real}$ and DBSCAN together with Haploview, following two of the three comparison criteria described in section 5.1.

The adjusted Rand index is all the higher as the agreement between two partitioning methods is high. Thus, we observe that CAST$_{bin}$ does not agree with the reference (Haploview) partitioning as well as CAST$_{real}$ and DBSCAN. This specificity of CAST$_{bin}$ is explained by the conversion of real mutual information values into binary values (see the role of parameter $q_{pairwise}$ in section 4). This discretization therefore entails slightly larger cluster sizes for CAST$_{bin}$, as seen in section 7.1.

On the left section of Figure 4 (a), the index is computed from few top SNPs. We observe that CAST$_{bin}$ and CAST$_{real}$ show a high Rand index in contrast to DBSCAN. However, in a GWAS context, we do not wish to examine only, say, the 20 top significantly associated SNPs. Thus, the most relevant section to focus on is around 50-100 top SNPs. In this latter section of Figure 4 (a), we observe that the CAST$_{real}$ and DBSCAN curves are comparatively close and clearly located higher than the CAST$_{bin}$ curve. This trend is observed up to the 1000 top most associated SNPs.

In Figure 4 (b), a low Mirkin distance indicates a high agreement between two partitioning methods. The observations in Figure 4 (b) confirm that CAST$_{bin}$'s agreement with Haploview clustering is always worse than the other two methods'. We have not shown the results for the Fowlkes-Mallows index as the curves obtained are quasi superimposable with those plotted for the adjusted Rand index.

The first general conclusion to draw from this first series of agreement comparisons on the Crohn's disease data set is that DBSCAN and CAST$_{real}$ show a high level agreement with Haploview partitioning, both being quite clearly better than CAST$_{bin}$.

Figure 4 (c) displays the results for pairwise comparisons: CAST$_{real}$ versus CAST$_{bin}$, DBSCAN versus CAST$_{bin}$ and DBSCAN versus CAST$_{real}$. According to the adjusted Rand index, DBSCAN and CAST$_{real}$ show a high agreement. Given our previous observations, we expected that CAST$_{bin}$ and CAST$_{real}$ would show a low level agreement, which is confirmed. DBSCAN and CAST$_{bin}$ yield partitions that almost always disagree more than for the two former couples of partitioning methods. This trend is confirmed with the Mirkin distance and the Fowlkes-Mallows index (results not shown).

As a second general conclusion of this section, we cross-confirm one of our previous observations: DBSCAN and CAST$_{real}$ each show a high agreement with Haploview. This fact is therefore also reflected by a high agreement between DBSCAN and CAST$_{real}$.

7.3 FLTM-driven GWASs

In Figure 5, the comparison of plots (a) to (c) and plot (d) shows how the dimension reduction allows to pinpoint the potentially most interesting regions on the genome. Thus, "sparse" association profiles are produced, as opposed to the dense output of the standard single-SNP GWAS.

The two putative causal SNPs located on chromosome 2 respectively reported in the WTCCC study (WTCCC, 2007) and in (Barrett et al., 2008)

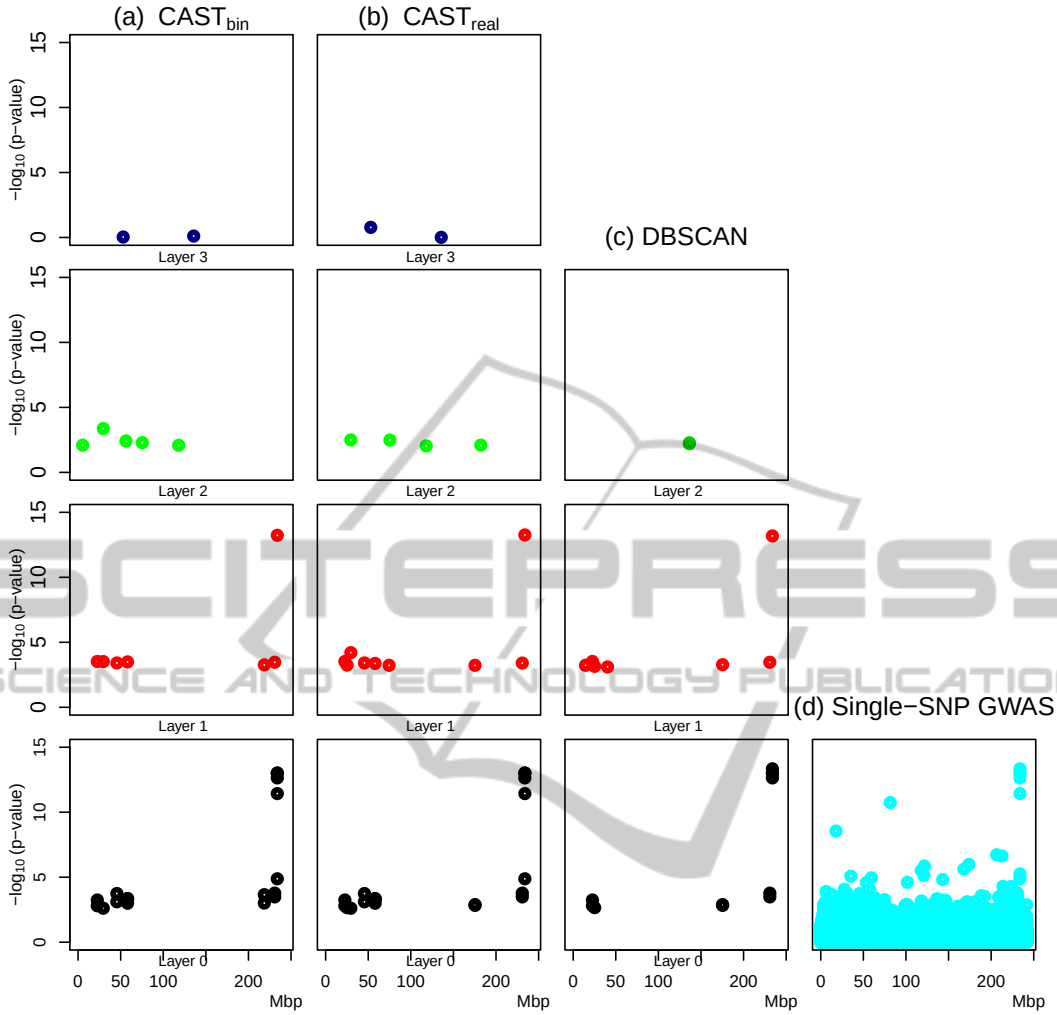


Figure 5: Impact of the choice of the partitioning method on the multilocus GWAS results. For the FLTM-based GWASs ((a) to (c)), one "sparse" association profile is displayed for each layer, as not all variables in a layer are examined. The single-SNP GWAS in (d) was performed using the gold standard PLINK (Purcell et al., 2007). Its output only deals with variables in layer 0 (observed variables). All plots show initial (i.e. non corrected) p-values.

are identified by the three FLTM-driven GWASs. Given that we used the same data set as in (WTCCC, 2007), one of the two results was expected. However, this result was not guaranteed, because of the data dimension reduction and of the subsumption involved in an FLTM-driven GWAS. Besides, it must be highlighted that the study in (Barrett et al., 2008) analyzed 8059 individuals (3230 cases and 4829 controls), whereas the WTCCC data set describes a population of size 5009. Table 1 shows that CAST$_{bin}$ and CAST$_{real}$ capture exactly the same four highly associated SNPs through the latent variables L_1 and L_2, belonging to layer 1. These variables are the right-most latent variables in layer 1, on the plots (a) and (b) of Figure 5. The virtual location of a latent variable is computed as the average of the locations of its child variables. Thus, the location of L_1 (or L_2) is 233837691 bp. The p-values computed for L_1 and L_2 differ since the data imputed for these latent variables differ. For either CAST$_{bin}$ or CAST$_{real}$, the SNP published in (WTCCC, 2007) is not grouped with other SNPs into a cluster, in contrast to DBSCAN. Table 1 shows that for DBSCAN, the latent variable L_3 subsumes SNPs among which are the two already published putative causal SNPs. L_1 and L_3 share three highly associated SNPs, including the putative causal SNP published in (Barrett et al., 2008). The virtual location of L_3 is 233830355 bp. We can see that L_1 captures LD on a slightly wider range than L_3, since the regions encompassed by the former and the latter variables spread over 29292 and 21571 bp, respectively.

A more thorough analysis of the Affymetrix ar-

Table 1: Analysis of the latent variables in layer 1 found significantly associated with Crohn's disease, by the three FLTM-driven GWASs with plug-in $CAST_{bin}$, $CAST_{real}$ and DBSCAN, respectively. For each clustering method, the latent variable is described on the first line. On the following lines, the highly associated SNPs subsumed by this latent variable are depicted. The identifier of each SNP is provided (rsXXXXXXX). The • character highlights the SNPs which are common children of latent variables L_1 (or L_2) and L_3. $*$ Note that the association tests used may differ between studies.

Clustering method	Variable	Location	p-value	p-value reported in another study*
$CAST_{bin}$	latent L_1	233837691 (1)	5.86×10^{-14}	
	rs6752107	233826187 • (2)	9.65×10^{-14} (3)	
	rs6431654	233826508 • (2)	9.96×10^{-14} (4)	
	rs3828309	**233845149** • (2)	$\mathbf{2.30 \times 10^{-13}}$ (5)	$\mathbf{2 \times 10^{-32}}$ (Barrett et al., 2008)
	rs3792106	233855479 (2)	3.70×10^{-12}	
$CAST_{real}$	latent L_2	see (1)	5.52×10^{-14}	
		see (2)		
DBSCAN	latent L_3	233830355	6.58×10^{-14}	
	rs10210302	**233823578**	$\mathbf{4.60 \times 10^{-14}}$	$\mathbf{7 \times 10^{-14}}$ (WTCCC, 2007)
	rs6752107	233826187 •	see (3)	
	rs6431654	233826508 •	see (4)	
	rs3828309	**233845149** •	see (5)	$\mathbf{2 \times 10^{-32}}$ (Barrett et al., 2008)

ray indicates that the region encompassed by L_1, [233826187, 233855479], contains four highly associated SNPs, interspersed with three non associated SNPs. Similarly, the interval covered by L_3, [233823578, 233845149], contains eight SNPs, including four non associated SNPs. Clearly, among the four highly associated SNPs pinpointed by each of L_1 and L_3, respectively three and two SNPs are highly associated with the disease because they are in LD with a putative causal SNP (see Table 1). However, not every SNP close to a putative causal SNP has been incorporated in the cluster subsumed by L_1, L_2 or L_3. To confirm the relevance of the clustering performed, an in-depth examination shows that these former close SNPs that are not in LD with putative causal SNPs are found poorly associated with the disease (in the order of 10^{-1}). Importantly, even the SNP flanking on the left the causal putative SNP published in (Barrett et al., 2008) and having a p-value equal to 1.32×10^{-5}, was not retained in L_1 or L_3's cluster. This observation shows that a fine-grain clustering is achieved for each of the three partitioning methods.

Therefore, a first remarkable result is that the subsumption process does not hinder the informativeness of L_1, L_2 and L_3: L_1, L_2 and L_3 are still found highly associated with the disease ($5,86 \times 10^{-14}$, $5,52 \times 10^{-14}$, 6.58×10^{-14} respectively).

Moreover, a second remarkable result is obtained. The standard GWAS (Figure 5 (d)) identifies two SNPs with a high statistical significance (rs13394205, located at around 18 Mbp (17849508), and rs11887827, located at around 81 Mbp (81519665)). The p-values of these two SNPs are respectively 2.28×10^{-9} and 1.81×10^{-11}. None of these SNPs were reported in former studies (WTCCC, 2007) and (Barrett et al., 2008), which identified them as false positives. In the layers 0 of the plots (a) to (c) of Figure 5, none of these two SNPs either appears. The reason lies in that during the top down traversal of the FLTM, the parents of these SNPs are detected as not significantly associated with the studied disease. Consequently, the descendants of these latent variables are not examined (and not displayed in the sparse outputs). Therefore, the FLTM-driven GWAS strategy exerts an efficient control of the number of false positives. Furthermore, all layers potentially allow to exert such a control, with a pruning effect on the forest structure guiding the GWAS.

In the context of this study, the general conclusion to draw from this section is that the three FLTM-driven GWASs capture the SNPs reported associated by two other studies and correctly detect false positive associations. Second, the differences reported in sections 7.1 and 7.2 between $CAST_{bin}$ and the two other clustering methods do not impact the quality of the corresponding FLTM-driven GWAS.

8 CONCLUSION AND PERSPECTIVES

In this paper, we have studied the impact of the choice of the clustering method to be plugged in the FLTM learning algorithm, for the purpose of a GWAS application. We have started analyzing this impact focusing on two scalable clustering methods, adding a relaxed variant of one of them. For this purpose, a methodological framework has been designed, which allows to compare the three clustering methods according to the following viewpoints: (1) effective ability to split or group the top associated SNPs, according to the underlying linkage disequilibrium structure; (2) data dimension reduction and associated risk of missing significant results through subsump-

tion; (3) relevance of the partitioning method to guide an FLTM-based GWAS pinpointing regions with significantly associated SNPs. The $CAST_{bin}$ clustering method was shown slightly different from $CAST_{real}$ and DBSCAN, from the clustering viewpoint. However, this difference was not reflected by a difference in GWASs' performances. Therefore, to the initial question "Which clustering method should be chosen", the answer for the Crohn's disease WTCCC data set relative to chromosome 2 would rather prioritize easiness in tuning parameters. In our experiments so far, the FLTM learning algorithm seems robust to the choice of the clustering method, provided that the intrinsic parameters of the latter are appropriately set. Further works include extending the current analysis to other chromosomes, for the WTCCC data set, as well as to other diseases, and extending our analysis to other clustering methods.

It was the first time that the FLTM learning algorithm was run on real GWAS data. It is questionable whether the present study should be complemented by intensive experiments run on simulated GWAS data sets. Given the high processing times required as soon as GWASs are addressed, and the recurring question of generating sufficiently realistic GWAS data, a less systematic approach, encompassing more diseases, seems wholly relevant.

Finally, to return to the multilocus aspect of the type of GWAS addressed here, one of our next tasks is to compare the FLTM-based GWAS strategy with the few other scalable multilocus approaches existing, including BEAGLE (Browning and Browning, 2007).

ACKNOWLEDGEMENTS

The project SAMOGWAS (Specific Advanced MOdels for Genome Wide Association Studies) is supported by the French National Research Agency (Agence Nationale de la Recherche, ANR). The authors are also grateful to the Wellcome Trust Case Control Consortium for providing the GWAS data used in this study.

REFERENCES

Abel, H. and Thomas, A. (2011). Accuracy and Computational Efficiency of a Graphical Modeling Approach to Linkage Disequilibrium Estimation. *Statistical Applications in Genetics and Molecular Biology*, 10(1):Article 5.

Ackerman, M. and Ben-David, S. (2009). Clusterability: a Theoretical Study. In Dyk, D. and Welling, M., editors, *Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS09), Journal of Machine Learning Research, Proceedings Track*, volume 5, pages 1–8.

Balding, D. (2006). A Tutorial on Statistical Methods for Population Association Studies. *Nature Reviews Genetics*, 7(10):781–791.

Barrett, J., Hansoul, S., Nicolae, et al. (2008). Genome-wide Association Defines more than 30 Distinct Susceptibility Loci for Crohn's Disease. *Nature Genetics*, 40(8):955–962.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering Gene Expression Patterns. In *Third Annual International Conference on Research in Computational Molecular Biology (RECOMB99)*, pages 33–42.

Browning, B. and Browning, S. (2007). Efficient Multilocus Association Testing for Whole Genome Association Studies Using Localized Haplotype Clustering. *Genetic Epidemiology*, 31:365–375.

Cahill, J. (2002). Error-Tolerant Clustering of Gene Microarray Data. Bachelors Honors Thesis, Boston College, Massachusetts.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining (KDD96)*, pages 226–231.

Fowlkes, E. and Mallows, C. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Gabriel, S., Schaffner, S., Moore, J., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science*, 296(5576):2225–2229.

Gibbs, R., Belmont, J., Hardenbol, P., et al. (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1):193–218.

Meila, M. (2005). Comparing Clusterings: an Axiomatic View. In *Twenty-second International Conference on Machine Learning (CML05), ACM*, pages 577–584.

Mirkin, B. (1998). Mathematical Classification and Clustering: from How to What and Why. *Classification, Data Analysis, and Data Highways*, 690:172–181.

Mourad, R., Sinoquet, C., and Leray, P. (2011). A Hierarchical Bayesian Network Approach for Linkage Disequilibrium Modeling and Data-dimensionality Reduction prior to Genome-wide Association Studies. *BMC Bioinformatics*, 12:16+.

Pritchard, J. and Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics*, 69(1):1–14.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: a Toolset for Whole-genome Association and Population-based Linkage Analysis. *The American Journal of Human Genetics*, 81(3):559–575.

Rand, W. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

The 1000 Genomes Project Consortium (2010). A Map of Human Genome Variation from Population-scale Sequencing. *Nature*, 467(7319):1061–1073.

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

Wang, N., Akey, J., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: the Interplay of Population History, Recombination, and Mutation. *The American Journal of Human Genetics*, 71(5):1227–1234.

WTCCC (2007). Wellcome Trust Case Control Consortium. Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature*, 447(7145):661–678.