



**HAL**  
open science

## Learning the optimal scale for GWAS through hierarchical SNP aggregation

Florent Guinot, Marie Szafranski, Christophe Ambroise, Franck Samson

► **To cite this version:**

Florent Guinot, Marie Szafranski, Christophe Ambroise, Franck Samson. Learning the optimal scale for GWAS through hierarchical SNP aggregation. 2017. hal-01597553v2

**HAL Id: hal-01597553**

**<https://hal.science/hal-01597553v2>**

Preprint submitted on 24 Oct 2017 (v2), last revised 19 Oct 2018 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Learning the optimal scale for GWAS through hierarchical SNP aggregation

Florent Guinot<sup>1,2,\*</sup>, Marie Szafranski<sup>1</sup>, Christophe Ambroise<sup>1,3</sup>, and Franck Samson<sup>1</sup>

<sup>1</sup>UMR 8071 LaMME – UEVE, CNRS, ENSIIE, USC INRA – 91000 Évry, FRANCE

<sup>2</sup>BIOOptimize – 51000 Reims, FRANCE

<sup>3</sup>UMR MIA-Paris – AgroParisTech, INRA, Université Paris-Saclay – 75005 Paris, FRANCE

\*To whom correspondence should be addressed: florent.guinot@genopole.cnrs.fr

## ABSTRACT

**Motivation:** Genome-Wide Association Studies (GWAS) seek to identify causal genomic variants associated with rare human diseases. The classical statistical approach for detecting these variants is based on univariate hypothesis testing, with healthy individuals being tested against affected individuals at each locus. Given that an individual's genotype is characterized by up to one million SNPs, this approach lacks precision, since it may yield a large number of false positives that can lead to erroneous conclusions about genetic associations with the disease. One way to improve the detection of true genetic associations is to reduce the number of hypotheses to be tested by grouping SNPs.

**Results:** We propose a dimension-reduction approach which can be applied in the context of GWAS by making use of the haplotype structure of the human genome. We compare our method with standard univariate and multivariate approaches on both synthetic and real GWAS data, and we show that reducing the dimension of the predictor matrix by aggregating SNPs gives a greater precision in the detection of associations between the phenotype and genomic regions.

## INTRODUCTION

Recent breakthroughs in microarray technology have meant that hundreds of thousands of single nucleotide polymorphisms (SNPs) can now be densely genotyped at moderate cost. As a result it has become possible to characterize the genome of an individual with up to a million genetic markers. These rapid advances in DNA sequencing technologies have also made it possible to carry out exome and whole-genome sequencing studies of complex diseases. In this context, Genome-Wide Association Studies (GWAS) have been widely used to identify causal genomic variants implied in the expression of different human diseases (rare, Mendelian, or multifactorial diseases).

From a statistical point of view, looking for these variants can be supported by hypothesis testing. The standard approach in GWAS is based on univariate regression (logistic regression in case-control studies), with affected individuals being tested against healthy individuals at one or more loci. Classical testing schemes are subject to false positives (that is to say SNPs that are falsely identified as significant variables). One way around this problem is to apply a correction for the False Discovery Rate (Benjamini and Hochberg, 1995; Dalmasso et al., 2005). Unfortunately, this increases the risk of missing true associations that have only a small effect on the phenotype (which is usually the case in GWAS). Maher (2008) suggested that standard approaches such as multiple hypothesis testing may not be appropriate for the detection of small effects from multiple SNPs. In such cases a significant part of the heritability can be missing and GWAS fails to detect all possible genetic variants associated with a disease.

Furthermore, this kind of standard approach faces other limitations:

1. It does not directly account for correlations among the predictors, whereas these correlations can be very strong as a result of linkage disequilibrium (LD). SNPs can be correlated even where they are not physically linked, because of population structure or epistasis.
2. It does not account for epistasis (gene by gene interactions), i.e. causal effects that are only observed when certain combinations of mutations are present in the genome.

3. It does not directly provide predictive models for estimating the genetic risk of the disease.
4. It focuses on identifying common markers with allele-frequency (MAF) above 5%, although it is likely that analyzing low-frequency (0.5%  $\leq$  MAF  $\leq$  5%) and rare (MAF  $\leq$  0.5%) variants would be able to explain additional disease risks or trait variability (Lee et al., 2014).

Uncovering some of the missing heritability can sometimes be achieved by taking into account correlations among variables, interaction with the environment, and epistasis, but this is rarely feasible in the context of GWAS because of the multiple testing burden and the high computational cost of such analyses (Manolio and Visscher, 2009). Moreover, regarding limitation (4), it may be argued that analyzing rare variants is more complex than analyzing more common variants. A large sample size is needed to observe a rare variant with a high probability.

A number of region- or gene-based multimarker tests have been proposed in the context of rare-variant association analysis. These include burden tests (Asimit et al., 2012), variance-component tests (Wu et al., 2011) and combined burden and variance component tests (Lee et al., 2012). Instead of testing each variant individually, these methods evaluate the cumulative effects of multiple genetic variants in a gene or a region, increasing power when multiple variants in the group are associated with a given disease or trait. They are based on various hypothesis about the underlying genetic model, and the power for each test depends on the true disease model. In comparison to single-variant-based tests, gene-based and region-based tests can represent a loss of power in cases where only a very small number of the variants in a gene are associated with the trait, where many variants have no effect, and where causal variants are low-frequency variants (Lee et al., 2014).

Although classical GWAS have limitations that prevent a full understanding of the heritability of genetic and/or multifactorial diseases, there are nevertheless ways of overcoming these limitations to some degree. For instance, it is possible to take into account the structure of the genome to define groups of genetic variants. It is well known that the human genome is structured into haplotype blocks, i.e. sizable regions over which there is little evidence for historical recombination and within which only a few common haplotypes may be observed (Ardlie et al., 2002). The boundaries of blocks and the specific haplotypes that they contain are highly correlated across populations (Gabriel et al., 2002). With this property of the human genome in mind, Huang et al. (2007) developed a method for detecting haplotype-disease associations in genome-wide studies, based on sliding windows of adjacent SNPs, along with a Monte Carlo procedure to adjust for multiple testing.

Group-based methods require an appropriate group definition, the usual approach is to group SNPs which are included in the same gene but this limits the analysis to coding regions. In Wu et al. (2010), the authors proposed to group SNPs into sets on the basis of their proximity to genomic features such as genes or haplotype blocks and then to identify the joint effect of each set via a logistic kernel-machine-based test. This approach lays the foundation for the Sequence Kernel Association Test method (Wu et al., 2011, SKAT). Llinares-Lopez et al. (2015) proposed an algorithm for genome-wide detection of contiguous intervals that may exhibit genetic heterogeneity with respect to a given binary phenotype. This algorithm automatically finds the start and end positions of these intervals, while properly correcting for multiple hypothesis testing using an approach developed by Tarone (1990). In Listgarten et al. (2013), the authors introduced a likelihood ratio-based set test that accounts for confounding structure. The model is based on the linear mixed model and uses two random effects, one to capture the set association signal and one to capture confounders. They demonstrate a control of type I error as well as an improved power over more traditionally used score test. Dehman et al. (2015) also took into account the structure of the genome in haplotype blocks and showed that a hierarchical clustering that takes the LD between SNPs as the similarity measure is able to recover this structure. To improve the detection power of relevant variables in a high-dimensional setting, Meinshausen (2008) proposed a hierarchical testing approach which considers the influence of clusters of highly correlated variables rather than individual variables. The statistical power of this method to detect relevant variables at single SNPs level was comparable to that of the Bonferroni-Holm procedure, but the detection rate was much higher for small clusters, and it increased further at coarser levels of resolution.

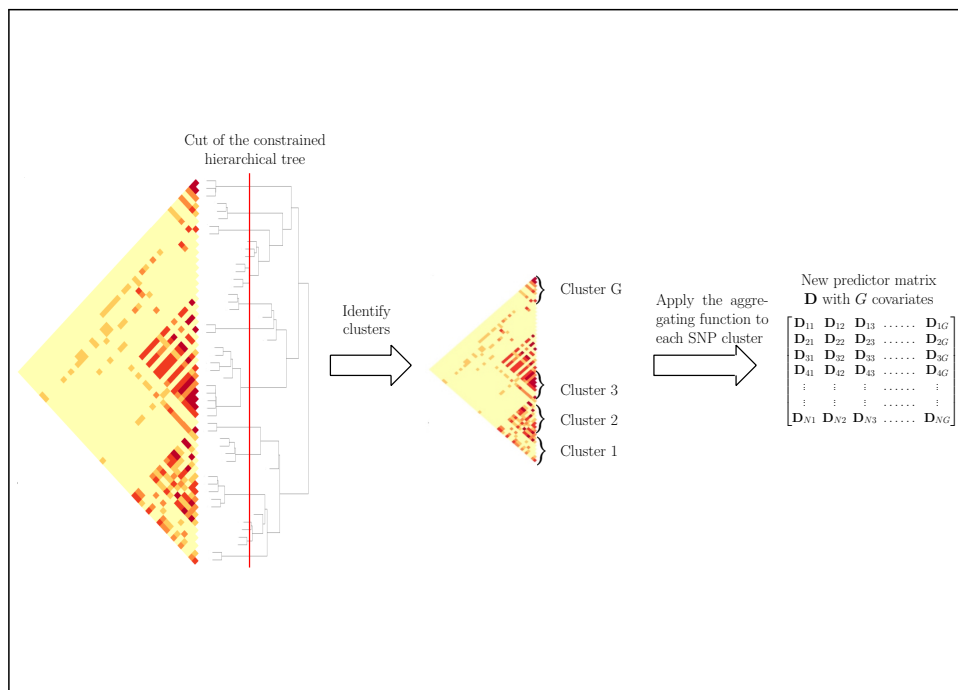
The present paper proposes a block-wise approach for GWAS analysis which leverages the LD structure among the genomic variants to reduce the number of hypotheses testing. We aggregate the SNPs into different clusters according to their LD levels and use a supervised learning approach to identify the clusters of SNPs related to a case-control phenotype. Our algorithm provides a group structure for the variables, enabling us to define a function that aggregates these clusters into new variables to be used in the GWAS machinery. The advantage of this method is that aggregating clusters of several SNPs into a single variable reduces the dimension of the data without loss of information, since we are grouping variables that are highly correlated (in strong LD). Performing multiple hypothesis tests on these new variables reduces the False Discovery Rate, and

consequently the risk of missing true genetic associations. We compare our method with the baseline approach, i.e. univariate hypothesis testing (Purcell et al., 2007), and two state-of-the-art methods, the logistic kernel machine method developed by Wu et al. (2010) and the approach of Llinares-Lopez et al. (2015) known as Fast Automated Interval Search (FAIS) on both synthetic and real datasets from the Wellcome Trust Case Control Consortium (WTCCC, 2007) and on ankylosing spondylitis data (International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al., 2013).

## METHOD

In this section we describe a new method for performing GWAS using a four-step method that combines unsupervised and supervised learning techniques. This method improves the detection power of genomic regions implied in a disease while maintaining a good interpretability. This method consists in:

1. Performing a spatially constrained Hierarchical Agglomerative Clustering (constrained-HAC) of the SNPs matrix  $X$  using the algorithm developed by Dehman et al. (2015).
2. Applying a function to reduce the dimension of  $X$  using the group definition from the constrained-HAC. This step is described and illustrated in Figure 1.
3. Estimating the optimal number of groups using a supervised learning approach to find the best cut into the hierarchical tree (cut level algorithm). This algorithm combines Steps 1 and 2 into an iterative process.
4. Applying the function defined in Step 2 to each group identified in Step 3 to construct a new covariate matrix and perform multiple hypothesis testing on each new covariate to find significant associations with a disease phenotype  $y$ .



**Figure 1.** Schematic view of Step 2 of the algorithm to calculate the matrix of predictors  $D$ .

### Step 1. Constrained-HAC

In GWAS, the covariates are ordinal and correspond to SNP genotypes such that  $X_{ij} \in \{0, 1, 2\}$  corresponds to the number of minor alleles at locus  $j \in [1, \dots, J]$  for observation  $i \in [1, \dots, N]$ .

To take into account the structure of the genome in haplotype blocks, we group the predictors (SNPs) according to their linkage disequilibrium in order to create a new predictor matrix which reflects the structure of the genome. We first use the algorithm developed by Dehman et al. (2015), which clusters SNPs into adjacent blocks. The clustering method is a spatially constrained hierarchical clustering based on Ward's incremental sum-of-squares algorithm (Ward, 1963), in which the measure of dissimilarity is not based on the Euclidean

distance but rather on the linkage disequilibrium between two SNPs:  $1 - r^2(j, j')$ . The algorithm also makes use of the fact that the LD matrix can be modeled as block-diagonal by allowing only groups of variables that are adjacent on the genome to be merged, which significantly reduces the computation cost. This algorithm is available via the R package called BALD on <http://www.math-evry.cnrs.fr/logiciels/bald>.

## Step 2. Dimension reduction function

One way of addressing issues related to high-dimensional statistics (and in particular the multiple testing burden that we mentioned above) is to reduce the dimensionality of the predictor matrix  $X \in \mathbb{R}^{N \times P}$  by creating a reduced matrix  $D$  with new covariates that nevertheless remain representative of the initial matrix. This means reducing the number of predictors  $P$  to  $G \ll P$ , with row  $D_i$  the  $G$ -dimensional vector of new predictors for observation  $i$ . In this study we use a blockwise approach to construct a matrix of new uncorrelated predictors  $D \in \mathbb{R}^{N \times G}$ , with  $G$  the number of groups in linkage disequilibrium identified via the constrained agglomerative hierarchical clustering described in Step .

While classical methods use the initial set of covariates to predict a phenotype, we propose combining a clustering model with a dimension reduction approach in order to predict  $y$ . For each group identified with the constrained-HAC, we apply a function to obtain a single variable defined as the number of minor alleles present in the group. For each observation  $i$  and in each cluster  $g \in [1, \dots, G]$ , the variable is defined as:

$$D_{ig} = \sum_{j \in g} X_{ij}. \quad (1)$$

In order that the values for the different groups are comparable, we eliminate the effect of group size by scaling the matrix  $D$  to unit variance and centering it. In the remainder of the paper we will refer to the covariates in  $D$  as *aggregated-SNP* variables.

## Step 3. Optimal number of groups estimation

---

### Algorithm 1: Supervised learning cut level algorithm

---

**input** : Covariates matrix  $X$   
**output** : Matrix  $D^{best}$  of aggregated-SNPs at best cut level

- 1 Define training and test set;
- 2 hierarchy  $\leftarrow$  Constrained-HAC on  $X^{train}$
- 3 cutlevel  $\leftarrow$  Initialize levels where to cut hierarchy
- 4 **for**  $i \leftarrow$  Sequence (cutlevel) **do**
- 5      $D^{train} \leftarrow$  Aggregating ( $X^{train}$ , hierarchy, cutlevel[ $i$ ]);
- 6      $D^{test} \leftarrow$  Aggregating ( $X^{test}$ , hierarchy, cutlevel[ $i$ ]);
- 7     ridgecoef  $\leftarrow$  RidgeRegression ( $Y^{train} \sim D^{train}$ );
- 8      $Y^{pred} \leftarrow$  Predict ( $X^{test}$ , ridgecoef);
- 9     AUC[ $i$ ]  $\leftarrow$  ROC ( $Y^{test}$ ,  $Y^{pred}$ );
- 10 **end**
- 11 bestlevel  $\leftarrow$  Which (cutlevel, Max (AUC) );
- 12  $D^{best} \leftarrow$  Aggregating ( $X$ , hierarchy, bestlevel);

---

Estimating the optimal number of groups to select, i.e. the level at which the hierarchical clustering tree should be cut, is a fundamental matter which impacts the relevance of the association analysis. It is known that the human genome is structured into haplotype blocks with little or no within-block recombination (Gabriel et al., 2002), but it is not easy to determine how these blocks are allocated throughout the genome for a given set of SNPs.

In the literature, in an unsupervised learning context, a number of models have been proposed for determining this optimal number of groups (Tibshirani et al., 2001; Hartigan, 1975; Caliński and Harabasz, 1974; Krzanowski and Lai, 1988). These methods are all based on the measure of within-group dispersion  $W_G$  with  $G \in [1, \dots, P]$ . Since GWAS consist in evaluating the likelihood of the disease from genetic markers, we propose using the phenotype  $y$  as a way of determining the optimal number of clusters.

We propose here a supervised validation set approach to find optimum. First we apply the constrained-HAC on a subset  $X^{train} \subset X$ , and for a given level of the hierarchy we apply the dimension reduction function defined

above (Step 2) to each of the  $G$  clusters. We then fit a ridge regression model using the training set  $D^{train}$  to estimate the coefficients. Ridge regression is a penalized model which shrinks the estimated coefficients towards zero and is known to have a good stability in comparison to other penalized-regression models such as lasso regression (Bousquet and Elisseeff, 2002). Moreover, a link can be established between the ridge regression model and the mixed linear model used in the estimation of the heritability in a high-dimensional setting (Bonnet et al., 2014). Once the coefficients are estimated, we predict the phenotypic values on the test set and calculate either the mean test set error when  $y$  is quantitative or the Area Under the ROC curve (AUC-ROC) when  $y$  is binary. The procedure, summarized in Algorithm 1, is then repeated for different levels in the hierarchy and the optimal cut level in the tree is defined as the level which maximizes the prediction accuracy criterion.

#### Step 4. Multiple testing on aggregated-SNP variables

Once the optimal number of groups has been determined, we apply the function (1) to each selected group and construct the matrix of aggregated-SNP. Here we use a standard Single Marker Analysis (SMA) to find associations with  $y$ , but instead of calculating  $p$ -value for each SNPs in  $X$ , we calculate  $p$ -value for each aggregated-SNP variables in  $D$ .

In standard SMA, a univariate generalized linear model (Nelder and Wedderburn, 1972) is fitted for each variable  $X_{.j}$ :  $f(\mu_i) = X_{ij}\beta$ , where  $\mu_i \equiv \mathbb{E}(Y_i|X_i)$  ( $Y_i \sim$  some exponential family distribution),  $f$  is a smooth monotonic 'link function',  $X_{ij}$  is the  $i^{th}$  row of the model matrix  $X_{.j}$  and  $\beta$  is a vector of 2 unknown coefficients with  $\beta_0$  for the intercept and  $\beta_1$  for the predictor  $j$ . Where the response variable is a binary trait (i.e. case-control phenotype), we use the logit function as the 'link function'  $f$  and  $Y_i \sim$  Bernoulli distribution. This model is known as the logistic regression model. Then, for each single-predictor model, we perform a Likelihood Ratio Test where we compare the intercept-only model against the single-predictor model and get for each predictor a  $p$ -value using the  $\chi^2$  distribution.

Given that a large number of covariates are being tested, we need to compute an appropriate significance threshold to control the family-wise error rate,  $FWER = \mathbb{P}(FP > 1)$  with  $FP$  being the False Positive, since keeping the threshold at the conventional value of  $\alpha = 0.05$  would yield numerous false positives. Several approaches, including the Bonferroni correction, have been proposed in the context of genetic studies for controlling the FWER (Sham and Purcell, 2014). An alternative approach, developed by Benjamini and Hochberg (1995), seeks to control the False Discovery Rate (FDR) which is the expectation of ratio between the number of false positives and the total positive outcomes:  $FDR = \mathbb{E}\left(\frac{FP}{FP + TP}\right)$ , with  $TP$  being the True Positive. The Bonferroni correction reduces the significance level according to the number of tests carried out in the study. However, in the context of GWAS, where hundreds of thousands of tests have to be performed, the Bonferroni correction is too strong, and will often decrease the significance threshold to a level where almost nothing is significant. Controlling FDR is therefore preferable. It is an approach that is less stringent but nonetheless powerful. The method for controlling FDR does not directly set a significance threshold, but rather identifies the largest  $p$ -value that is substantially smaller than its expected value (by a factor of at least  $1/\phi$  where  $\phi$  is the desired FDR level), given that all the tests follow  $H_0$ . The  $p$ -value thus identified and all smaller  $p$ -values are deemed to be significant. We use this method as the significance level control for all the multiple hypothesis tests described in this paper.

## NUMERICAL SIMULATIONS

The performance evaluation described below was designed to assess the ability of our method to retrieve causal SNPs or causal clusters of SNPs under different simulation scenarios. For each scenario, we use a matrix  $X_{HAPGEN}$  of SNPs generated by the HAPGEN2 software (Su et al., 2011) with a sample size of 1000 individuals. This software allows to simulate an entire chromosome conditionally on a reference set of population haplotypes (from HapMap3) and an estimate of the fine-scale recombination rate across the region, so that the simulated data share similar patterns with the reference data. We generate the chromosome 1 (103 457 SNPs) using the haplotype structure of CEU population (Utah residents with Northern and Western European ancestry from the CEPH collection) as reference set. The HAPGEN2 software allows to generate a controls-only matrix of SNPs (no disease allele). We generate a posteriori the phenotype using the logit model with a given set of causal SNPs. The main difference between the different scenarios is to be found in the way that the case-control phenotype  $y$  is simulated.

## Simulation of the case-control phenotype

For all simulation scenarios, we simulated a case-control phenotype  $y$  under a logistic regression model. We chose a matrix  $\tilde{X}$  with  $\ell \in [1, 3, 5, 7, 10]$  causal variables to create the case-control phenotype  $y$  under the logit model:

$$\mathbb{P}(y_i = 1 | \tilde{X}_i) = \frac{\exp(\beta_0 + \beta \tilde{X}_i)}{1 + \exp(\beta_0 + \beta \tilde{X}_i)},$$

where  $\beta = [\beta_1, \dots, \beta_\ell]$  is the vector of coefficients corresponding to the  $\ell$  predictors  $[\tilde{X}_{.1}, \dots, \tilde{X}_{.\ell}]$  and  $\beta_0$  is the intercept defined as  $\ln\left(\frac{\pi}{1-\pi}\right)$ , with  $\pi$  the true prevalence of the disease in the population. The predictors are centered to have zero-mean before generating the vector of probability.

The logit function gives a vector of probabilities for the phenotype equals to 1, conditionally on the  $\ell$  predictors. The case-control phenotype is then generated following a Bernoulli distribution function with a probability equal to  $\mathbb{P}(y_i = 1 | \tilde{X}_i)$ .

The  $\beta$  coefficients are chosen such that the phenotype can be predicted correctly using the causal predictors in a logistic regression model, in other words such that the Mean Square Error of this model (MSE =  $\frac{1}{n} \sum_{i=1}^n (y_i - \beta \tilde{X}_i)^2$ ) is equal to zero. Concretely, we are testing different values for the  $\beta$  coefficients until we get an MSE equals to 0.

In our simulations, the difficulty of the problem is directly linked to the number of causal predictors used to generate  $y$ . The higher the number of causal predictors used, the harder it is to retrieve the true causal variables when performing multiple hypothesis testing. Since the same  $\beta$  coefficients are used for all causal predictors, each of them has the same effect on the phenotype, which is smaller when there are more of them. As previously mentioned, we generated the phenotype using causal SNPs simulated with the HAPGEN2 software. However, as commercial genechips such as Affymetrix and Illumina arrays do not genotype the full sequence of the genome, some SNPs are thereby unmapped and the marker density is in general lower than the HapMap marker density. That is why we chose, in our numerical simulation, to generate the phenotype with causal SNPs chosen from the HAPGEN matrix and to assess the performance of the methods using only those SNPs which are mapped on a standard Affymetrix genechip (about 41 000 mapped SNPs). By doing so, some causal SNPs are not mapped on the commercial SNP set and thus simulations are more similar to real genome-wide analysis conditions.

There are two simulation scenarios:

- **singleSNP.** The phenotype was simulated using randomly sampled SNPs from the simulated matrix  $X_{\text{HAPGEN}}$  using  $\ell = 1, 3, 5, 7$  and 10 causal single SNPs.
- **clusSNP.** The same methodology was used to simulate the response phenotype  $y$ , but instead of using single SNPs as causal variants, we used aggregated-SNPs as the causal variables. We first performed a constrained-HAC on  $X_{\text{HAPGEN}}$  and chose a cut level in the hierarchy to get clusters with an average size of about 20 SNPs. We chose this value of 20 SNPs since we know from Ardlie et al. (2002) that the human genome is structured in LD blocks with size varying from 10 to 30 kb (for European populations), which corresponds to about 20 SNPs in our simulated matrix. Having identified the number of clusters, we created a new predictor matrix  $D$  as described in Step . We then chose  $\ell = 1, 3, 5, 7$  and 10 causal SNP clusters from  $D$  to generate the case-control phenotype  $y$  using the logit model.

In total, we simulated 5 phenotype vectors for each simulation scenario, our aim being to assess the ability of each method to retrieve the true causal SNPs and/or aggregated-SNPs.

## Performance evaluation

For each scenario, we compare our method with the SKAT model, the FAIS algorithm and the classical Single Marker Analysis (SMA). Our approach is similar to SMA, but instead of testing how the phenotype is associated to each SNP, we are testing how it is associated to each aggregated-SNP. We will refer to this approach as Single Aggregated-SNP Analysis (SASA). Regarding the group definition for the SKAT model, we used a fixed size of 20 SNPs to define the set since Wu et al. (2010) claimed that their logistic kernel-machine testing approach is statistically valid irrespective of the grouping scheme. For FAIS, no predefined group identification is necessary since this method looks at all the possible SNP sets using a sliding windows approach. However, since the FAIS algorithm works only on binary-coded SNPs (0 if no minor allele, 1 otherwise), we assessed this method using a phenotype simulated with binary SNPs as causal variables. In order to get results comparable with the other methods, we kept the same matrix  $\tilde{X}$  of causal predictors and coded the SNPs as binary variables to generate a

phenotype with the logit model. For all methods, we are correcting for multiple testing by controlling the False Discovery Rate using the method of Benjamini and Hochberg (1995).

The problem of retrieving true causal associations can be represented as a binary decision problem where the compared methods are considered as classifiers. The decision made by a binary classifier can be summarized using four numbers: True Positives ( $TP$ ), False Positive ( $FP$ ), True Negatives ( $TN$ ) and False Negatives ( $FN$ ). We represent True Positive Rate (Recall or Power =  $TP/(FN + TP)$ ) versus Precision (Precision =  $TP/(FP + TP)$ ). In this context, a True Positive corresponds to a significant  $p$ -value on true causal variables, a False Positive to a significant  $p$ -value on noise variables, and a False Negative to a non-significant  $p$ -value on true causal variables. In the case where one true causal SNP is not mapped on the commercial SNP set, we consider that we are able to retrieve its position if we find a significant association with the closest SNP mapped on the commercial SNP set.

For each method, the following may be considered as a True Positive depending on the simulation scenario:

- **SMA.** A significant true causal variable (*singleSNP*) or a significant variable included in a true causal cluster (*clusSNP*).
- **SASA.** A significant aggregated-SNP containing at least one true causal variable (*singleSNP*) or at least a part of a true causal cluster (*clusSNP*).
- **SKAT and FAIS.** A significant region including at least one true causal single SNP (*singleSNP*) or containing a part of the genomic region defining a true causal cluster. (*clusSNP*).

We also consider two kinds of evaluation on SNPs and clusters for each scenario, as explained in Figure 3.

## Results and discussions of the numerical simulations

### Area Under the ROC Curve

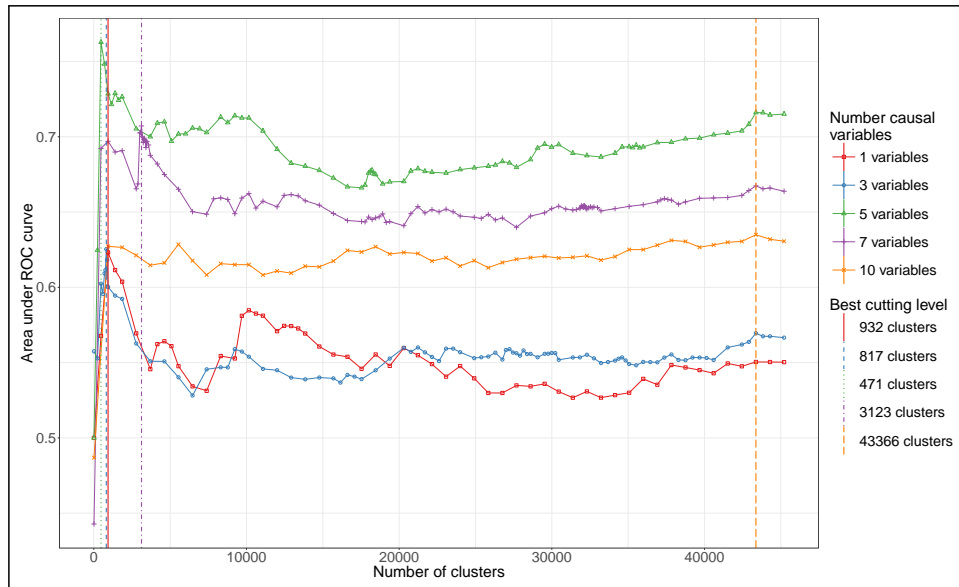
For each simulation, the cut level algorithm was applied. We recall that this algorithm calculates a prediction error on a test set for several levels in a constrained-HAC tree with a ridge regression model and chooses the level for which this error is the smallest. The AUC-ROC is plotted for the different levels, and the best cut level corresponds to the level for which AUC-ROC is the greatest. The results from the simulation scenario *clusSNP* described in Section 12 are shown in Figure 2. Our algorithm cuts the hierarchy either at a fairly high level (few large clusters) or at a low level (many small clusters), depending on the number of causal variables we used to generate the phenotype. Notwithstanding the scenario with 10 causal variables, the algorithm is able to increase the predictive power by aggregating SNPs with the function (1). We are thus able to build a matrix of uncorrelated aggregated-SNP predictors that are representative of the initial SNP matrix and strongly linked to the phenotype. Regarding the scenario with 10 causal variables, our algorithm does not take benefit of aggregating the SNPs with a predictive power which does not increase substantially at a higher level in the hierarchy. Nonetheless, since one of our goals is to reduce the dimension of the predictor matrix (to decrease the amount of hypothesis testing to be performed), even under this scenario we may choose a higher level in the hierarchy in order to reduce the dimension, without greatly affecting the predictive power. For instance, in the AUC-ROC curve for 10 causal variables (Figure 2), we could decide to reduce the number of selected clusters from 43666 to 5000 without losing too much information.

### Performance results for simulated data

As described in Section 12, we evaluate and compare the methods using two metrics, namely *Recall* and *Precision*. We compute these metrics by considering either single SNPs or aggregated-SNPs as true causal variables, with 5 repetitions for each simulation scenario. The averaged results are shown in Figure 3. Since the evaluation based on clusters requires a proper group definition, we only compare group-based approaches SKAT and FAIS to our method. For the evaluation based on SNPs, we also consider the baseline approach SMA.

Compared to other group-based approaches, SASA is fairly competitive. The FAIS algorithm performs poorly both in terms of precision and power, the sliding windows approach does not seem to be able to efficiently retrieve true causal clusters of SNPs. Compared to SKAT, our method shows a better precision when the evaluation is based on clusters, with a comparable power in the *singleSNP* scenario while SKAT exhibits a good power in the *clusSNP* scenario. As shown in Figure 2 and discussed above, the size of the selected clusters in SASA are in average larger than the true causal clusters which may involve more False Negatives than expected and decrease the power in the *clusSNP* scenario. On the other side, when the evaluation is based on SNPs, SASA exhibits a better precision and a much better power than SKAT in the *clusSNP* scenario. Compared to the baseline approach SMA, the performances of SASA are in line in both scenario in term of precision, with a slight advantage for SMA with 3, 5 and 7 causal clusters when evaluating on SNPs. However, SASA performs with a higher power.





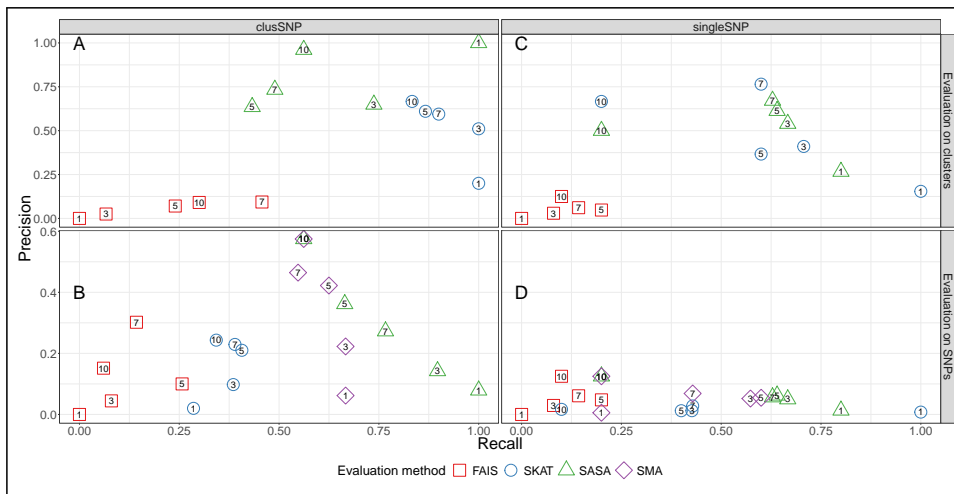
**Figure 2.** AUC-ROC according to the number of clusters in the *clusSNP* scenario: the vertical lines indicate the number of aggregated-SNPs (clusters) obtained with Algorithm 1, i.e. the level where the prediction error is minimized (AUC-ROC at its maximum).

Note that in the *singleSNP* scenario, when the evaluation is based on SNPs, the precision is very low for all the methods, which indicates a high number of False Positives. In GWAS, having a method with a good precision is as important, or even more important, than having a good recall. It is better to spot a few significant associations with a high certainty than to spot numerous significant associations but with only a low level of certainty for most of them. For this reason, we believe that our method represents an improvement in terms of precision without loss of power insofar as SASA seems able to detect significant genomic regions associated with the phenotype with a higher degree of certainty than standard approaches.

## APPLICATION ON REAL DATASETS

To evaluate the performance of our method on real data, we performed GWAS analysis on datasets made available by (WTCCC, 2007). The WTCCC data collection contains 17000 genotypes, composed of 3000 shared controls and 14000 cases representing 7 common diseases of major public health concern: inflammatory bowel disease (IBD), bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), and Type I (T1D) and Type II (T2D) diabetes. Individuals were genotyped with the Affymetrix GeneChip 500K Mapping Array Set and are represented by about 500,000 SNPs (before the application of quality control filters). In parallel to the analysis of the WTCC data, we decided to assess our method on another dataset from a different study. The ankylosing spondylitis (AS) dataset consists of the French subset of the large study of the International Genetics of Ankylosing Spondylitis (IGAS) study (International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al., 2013). For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The dataset contains 408 cases and 358 controls, and each individual was genotyped for 116,513 SNPs with Immunochip technology.

To remove the bias induced by population stratification in Genome-Wide analysis, we added the first 5 genomic principal components into the regression model as described in (Price et al., 2006). Since the methods evaluated here do not deal with missing values, we chose to impute the missing genotypes with the most frequent genotypic value,  $h_j$  observed for each  $j$  SNP. We applied our cut level algorithm to find relevant clusters of SNPs and, as in Section 12, we performed single marker analysis on single SNPs (SMA) and on groups of SNPs (SASA, SKAT, FAIS). We then compared the significant associations detected by the different methods to reveal possible new associations with the phenotype.



**Figure 3.** Recall vs Precision for each method (shape and colors in plot). In columns are the simulation scenarios. In rows, we evaluate performance on clusters (top) and SNPs (bottom) level. The first row illustrates the performance to retrieve the true causal clusters, thus only group-based approaches are considered (SASA, SKAT and FAIS). The numbers inside the points correspond to the number of causal predictors and each point is the average value of 5 repetitions with the simulation parameters. (A) TP = Part or entire cluster as significant. (B) TP = single SNP contained in the causal cluster as significant. (C) TP = cluster containing at least one true causal single SNP as significant. (D) TP = each single SNP as significant.

## Results on real datasets

### AUC-ROC curves

In this section, we compare the AUC-ROC curves generated by our cut level algorithm for each disease (WTCCC and AS data).

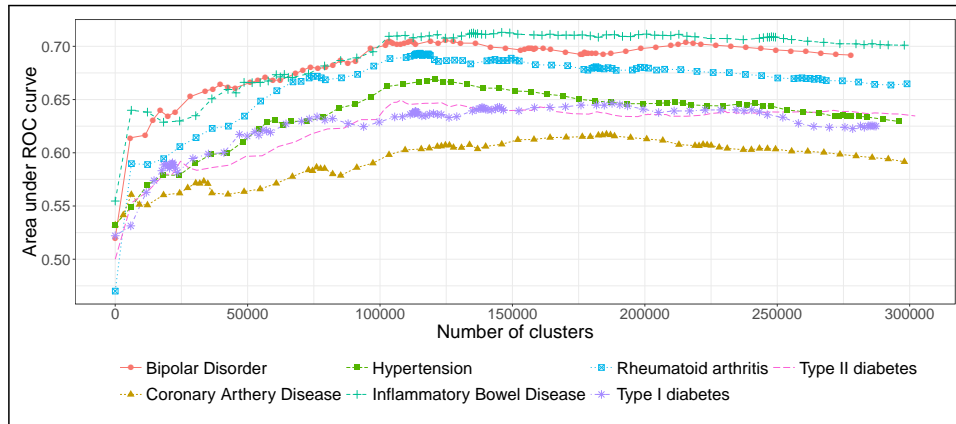
Concerning the WTCCC diseases, given that patients were all genotyped using the same GeneChip, their genotypes have the same LD structure, and therefore the shapes of the AUC-ROC curves should be very similar between the different diseases. As can be observed in Figure 5 (WTCCC diseases), the shape of the AUC-ROC curves are closely similar, with a chosen cut level located around 100 000 clusters of SNPs, suggesting a shared LD pattern among patients.



**Figure 4.** AUC-ROC for different cut levels in a HAC-tree of the spondylitis arthritis disease (ImmunoChip genechip). Each point corresponds to an AUC value computed on a test set from a logistic ridge regression model for a given level in the constrained-HAC tree.

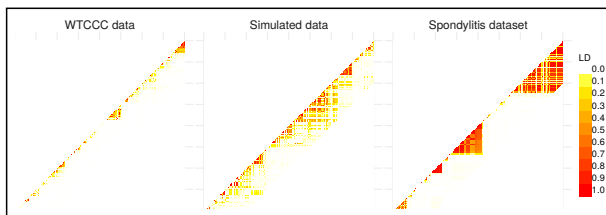
In contrast, the AUC-ROC from the AS data (Figure 4) behaves differently from the WTCCC data. Predictive power is substantially improved if aggregated-SNP predictors are used at a fairly high level in the hierarchical tree

(7478 optimal clusters identified by the cut level algorithm). It is relevant to note that the pattern we observe on this real dataset is similar to the pattern we observed in the numerical simulations, especially under the *clusSNP* scenario.



**Figure 5.** AUC-ROC for different cut levels in a HAC-tree of 7 WTCCC diseases after quality control filters. Each point corresponds to an AUC value computed on a test set from a logistic ridge regression model for a given level in the constrained-HAC tree.

As we remarked concerning the WTCCC results, the algorithm identifies a relatively high number of clusters in relation to AS and simulated data. This difference is certainly due to the LD level among the genetic markers in the Affymetrix GeneChip. The correlation levels among SNPs for a given bandwidth are similar between the simulated and the AS data, but greater than for the WTCCC data (Table 1 and Figure 6). This suggests that there is a stronger LD pattern between blocks of SNPs in AS and simulated data, implying that the optimal number of clusters identified by the algorithm is dependent on the LD level among variables.



**Figure 6.** Comparison of linkage disequilibrium level among SNPs for 3 different types of dataset: WTCCC, simulated and ankylosing spondylitis datasets. LD computation is based on  $R^2$  between SNPs.

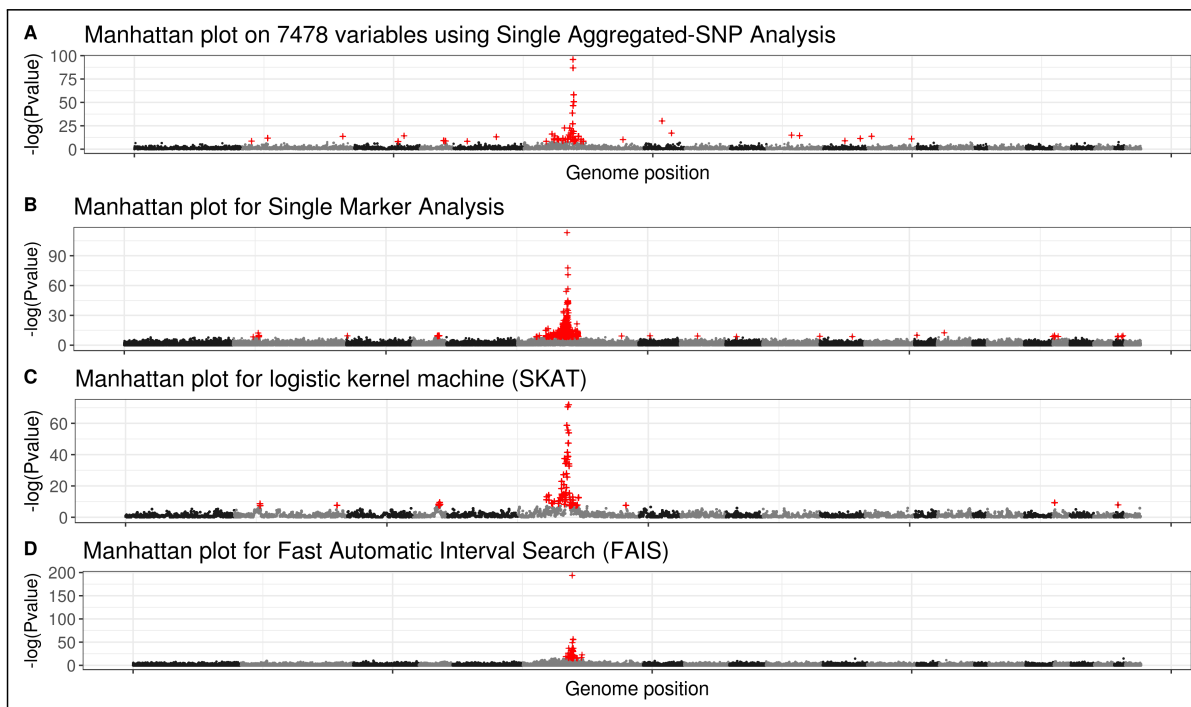
Dataset	SNP/kb	Median	Mean
Simulated data	$1.3 \times 10^{-27}$	$1 \times 10^{-2}$	0.11
WTCCC data	$7 \times 10^{-32}$	$9 \times 10^{-4}$	0.03
AS data	$9 \times 10^{-9}$	$3 \times 10^{-2}$	0.27

**Table 1.** Comparison of marker density and averaged LD level between markers in a region of 300 SNPs for the different datasets

### GWAS analysis on AS dataset

To evaluate the ability of our procedure to discover new associations between SNPs and ankylosing spondylitis, we compare our procedure with the univariate approach (SMA), SKAT model and FAIS algorithm. For SASA, we perform multiple hypothesis testing on the aggregated-SNP predictors in order to unravel significant associations with the phenotype. Figure 7 presents the result of the association analysis. For each method the logarithm of the  $p$ -value of the different predictors is plotted along their position on the genome (this plot is also known as a Manhattan plot).

Both the SMA and the SASA methods highlight a region on chromosome 6 strongly associated with the phenotype. This region corresponds to the Major Histocompatibility Complex (MHC), and Human Leukocyte Antigen (HLA) class I molecules HLA B27 belonging to this region have been identified as a genetic risk factor associated with ankylosing spondylitis (Woodrow and Eastmond, 1978). Our method SASA succeeds in detecting this risk locus with a good precision, 59 aggregated-SNPs variables are significantly associated with the phenotype compared to 602 significantly associated SNPs with the standard SMA approach. SKAT and FAIS identified less association but were still able to retrieve MHC region on chromosome 6.



**Figure 7.** Manhattan plots showing results of GWAS analysis on ankylosing spondylitis data. For each Manhattan plot, the SNPs whose p-values are above the Benjamini-Hochberg threshold are highlighted in red. (A) 59 significantly associated aggregated-SNPs. (B) 602 significantly associated single SNPs. (C) 540 significantly associated groups of SNPs and (D) 106 significantly associated groups of SNPs.

## DISCUSSION

Overall, taking into account the linkage disequilibrium structure of the genome and aggregating highly-correlated SNPs is seen to be a powerful alternative to standard marker analysis in the context of GWAS. In terms of risk prediction, our algorithm proves to be very effective at classifying individuals given their genotype, while in terms of the identification of loci, it shows its ability to identify genomic regions associated with a disease with a higher precision than standard methods.

In this work we propose a four-step method explicitly designed to utilize the linkage disequilibrium in GWAS data. Our method combines, on the one hand, unsupervised learning methods that cluster correlated-SNPs, and on the other hand, supervised learning techniques that identify the optimal number of clusters and reduce the dimension of the predictor matrix. We evaluated the method on numerical simulations and real datasets and compared the results with standard single-marker analysis and group-based approaches (FAIS and SKAT). In our simulations we remarked that the combination of our aggregating function with a ridge regression model leads to a major improvement in terms of predictive power when the linkage disequilibrium structure is strong enough, hence suggesting the existence of multivariate effects due to the combination of several SNPs. These results remained consistent across a wide range of real datasets (WTCCC and ankylosing spondylitis datasets).

In terms of the identification of associated loci in different simulation scenarios, our method demonstrates its ability to retrieve true causal SNPs and/or clusters of SNPs with substantially higher precision coupled with a good power. On real GWAS data, our method has been able to recover a genomic region associated with ankylosing spondylitis (HLA region on chromosome 6) with a higher precision than standard single-marker analysis.

To improve our method further, while taking into account structured input variables in GWAS, there are different avenues that may be explored. One avenue would involve highlighting potential non-linear relationships between aggregated-SNPs and a response phenotype. This could be done by making use of the continuous nature of aggregated-SNPs variables (in contrast to the ordinal nature of single SNP variables), by using generalized additive models (Breiman, 1993), and by performing non-linear regression using natural polynomial splines. In addition, whereas we evaluated our method for binary traits (case-control phenotype), a possible extension might include quantitative non-binary traits (i.e., using a ridge regression model instead of logistic ridge regression).

## ACKNOWLEDGEMENT

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from [www.wtccc.org.uk](http://www.wtccc.org.uk). Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

## FUNDING

UMR 8071 LaMME - UEVE, CNRS, ENSIIE, USC INRA, 23 bd de France, 91000 Evry, France.  
BIOptimize, 51000 Reims, France.

## AVAILABILITY OF DATA AND MATERIAL

A webtool to run the proposed approach have been developed and an access is available on demand. If you wish to use it, please send an email to the corresponding author.

## AUTHORS' CONTRIBUTION

FG, CA and MS developed the approach and the design. FG conducted the simulation study, applied the approach to the real data and drafted the manuscript. MS and CA supervised the implementation process and critically read and edit the manuscript. FS developed the webserver and made it available online. The work of FG fulfills part of the requirements of his PhD. All authors read and approved the final manuscript.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The dataset regarding ankylosing spondylitis consists of the French subset of the large study of the International Genetics of Ankylosing Spondylitis (IGAS) study International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al. (2013). For this subset, unrelated cases were recruited through the Rheumatology clinic of Ambroise Paré Hospital (Boulogne-Billancourt, France) or through the national self-help patients' association: "Association Française des Spondylarthritiques". Population-matched unrelated controls were obtained from the "Centre d'Etude du Polymorphisme Humain", or were recruited as healthy spouses of cases. The protocol was reviewed and approved by the Ethics committee of the Ambroise Paré hospital. All participants gave their informed consent to the study.

## REFERENCES

- Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309.
- Asimit, J. L., Day-Williams, A. G., Morris, A. P., and Zeggini, E. (2012). ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, 73(2):84–94.
- 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(1):289–300.
- Bonnet, A., Gassiat, E., and Lévy-Leduc, C. (2014). Heritability estimation in high dimensional linear mixed models. *ArXiv e-prints*.
- Bousquet, O. and Elisseeff, A. (2002). Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526.
- Breiman, L. (1993). Fitting additive models to regression data. diagnostics and alternative views. *Computational Statistics and Data Analysis*, 15:13–46.
- Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.
- Dalmasco, C., Broët, P., and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics*, 21:660–668.
- Dehman, A., Ambroise, C., and Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16:148.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., L.-C., S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.

- Hartigan, J. A. (1975). *Clustering algorithms*. Books on Demand.
- Huang, B. E., Amos, C. I., and Lin, D. Y. (2007). Detecting haplotype effects in genomewide association studies. *Genetic Epidemiology*, 31(8):803–812.
- International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes, A., Hadler, J., Pointon, J. P., Robinson, P. C., et al. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature Genetics*, 45(7):730–738.
- Krzanowski, W. J. and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1):23–34.
- Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics*, 95(1):5–23.
- Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.
- Listgarten, J., Lippert, C., Kang, E. Y., Xiang, J., Kadie, C. M., and Heckerman, D. (2013). A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, 29(12):1526–1533.
- Llinares-Lopez, F., Grimm, D. G., Bodenham, D. A., Gieraths, U., Sugiyama, M., Rowan, B., and Borgwardt, K. (2015). Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics*, 31(12):i240–i249.
- Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21.
- Manolio, T. A. and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.
- Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384.
- 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 B., P. I. W., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.
- Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346.
- Su, Z., Marchini, J., and Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304.
- Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522.
- Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423.
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.
- Woodrow, J. C. and Eastmond, C. J. (1978). HLA B27 and the genetics of ankylosing spondylitis. *Annals of the Rheumatic Diseases*, 37(6):504–509.
- WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.
- Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *American Journal of Human Genetics*, 86(6):929–942.
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93.