UNIVERSITE DE LIEGE
FACULTE DE MEDECINE VETERINAIRE
DEPARTEMENT DE GESTION VETERINAIRE DES RESSOURCES ANIMALES
SERVICE DE BIOINFORMATIQUE ET BIOSTATISTIQUE

# CONTRIBUTION AUX METHODES DE CARTOGRAPHIE D'EPISTASIE UTILISANT LA STATISTIQUE NON-PARAMETRIQUE

# CONTRIBUTION TO EPISTASY MAPPING METHODS THROUGH THE USE OF NON-PARAMETRIC METHODOLOGY

**Sinan ABO ALCHAMLAT**

**THESE PRESENTEE EN VUE DE L'OBTENTION DU GRADE DE**
**Docteur en Sciences Vétérinaires**

**ANNEE ACADEMIQUE 2017-2018**

# *Acknowledgments*

Many people have influenced, inspired, and helped me throughout my studies.

First and foremost, I would like to express my sincerest gratitude to my supervisor, Professor: FARNIR Frédéric for inspiring my research work and guiding me with endless patience in the past few years.

I would like to thank all the members of the department of biostatistics and bioinformatics especially Nassim Moula and Evelyne Moyse.

I also would like to thank all the PhD students of the department of biostatistics and bioinformatics especially Do Duc Luc and Duy Nguyen.

I also thank the University of Damascus in Syria for its financial support for part of this studying.

I will never forget to thank my friends for providing support and friendship that I needed.

I especially thank my mom, dad, brother, sister and her husband, my parents who have sacrificed their lives for my sister my brother and myself and provided unconditional love and care.

Last but not least, I owe my greatest gratitude my wife Lamees for her love and support. She always encouraged me to pursue what I want and make best efforts to make it happen.

For my children: Jana, Elena and Amir.

# Abbreviations

| | |
|---|---|
| **AIC** | Akaike Information Criterion |
| **ANN** | Artificial Neural Network |
| **BEAM** | Bayesian Epistasis Association Mapping |
| **BHIT** | Bayesian High-Order Interaction Toolkit |
| **BIC** | Bayesian Information Criterion |
| **BNP** | Bayesian Network Prior |
| **BOOST** | BOolean Operation-based Screening and Testing |
| **CART** | Classification And Regression Trees |
| **DT** | Decision Tree |
| **FAM-MDR** | Flexible Family-based Multifactor Dimensionality Reduction |
| **GMDR** | Generalized Multifactor Dimensionality Reduction |
| **GRAMMAR** | Genomewide Rapid Association using Mixed Model And Regression |
| **GWAS** | Genome-Wide Association Study |
| **HSA9** | Human chromosome 9 |
| **IBD** | Identify-By-Descent |
| **ID3** | Iterative Dichotomiser 3 |
| **GEDT** | Grammatical Evolution Decision Trees |
| **GPDTI** | Genetic Programming Decision Tree Induction method to find epistatic effects in common complex diseases |
| **KNN** | K-Nearest Neighbors |
| **LD** | Linkage Disequilibrium |
| **MAF** | Minor Allele Frequency |
| **MB-MDR** | Model-Based Multifactor Dimensionality Reduction |
| **MDR** | Multifactor Dimensionality Reduction |
| **MDR-ER** | Balancing Functions for Adjusting the Ratio in Risk Classes and Classification Errors for Imbalanced Cases and Controls Using Multifactor-Dimensionality Reduction |
| **MCMC** | Markov Chain Monte Carlo |
| **MLP** | MuLtilayer Perceptron |
| **NN** | Neural Networks |
| **OOB** | Out-Of-Bag |
| **PDM** | Parameters Decreasing Method |
| **PGMDR** | Pedigree-based Generalized Multifactor dimensionality Reduction |
| **RF** | Random Forests |
| **RMDR** | Robust Multifactor Dimensionality Reduction |
| **SNP** | Single-Nucleotide Polymorphism |
| **SVM** | Support Vector Machine |
| **SWSFS** | Sliding Window Sequential Forward Feature Selection |
| **WTCCC** | Wellcome Trust Case Control Consortium |

# Table of content

# Summary - Résumé

# *Summary*

## Introduction

These last years have seen the emergence of a wealth of genetic information at the molecular level. Some of the main recent breakthroughs in biology originate from this new knowledge, allowing application of new strategies in many fields of the biological research. Although approaches targeting the association between phenotypic characteristics and DNA variations have been successful, many elements in the genetic landscape of the studied traits are still unknown and uncharacterized. A track to new findings, potentially useful for a better understanding of complex determinisms, is the detection of interactions between genomic regions affecting the traits of interest rather than single locus associations. While the detection of such interactions has been the focus of many methods, and despite some successes of these methods to solve difficult problems and to detect some of these genetic interactions, there is currently no gold standard method able to detect interactions in all situations, and the relative performances of these methods remain largely unclear. This thesis is a contribution to this field of interactions mapping:in the first study, we propose a novel approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) methods for the detection of gene-gene interactions as a possible alternative to existing algorithms, especially in situations where the number of involved determinants is high. In the second study, we propose another strategy based on the principle of the aggregation of experts, where the experts would be a set of popular published methods.

## Results

The results obtained in the first study on both simulated data and real genome-wide data demonstrate some of the features that make KNN-MDR interesting in terms of accuracy and power: in many cases, it significantly outperforms its recent competitors. More specifically, the analyses on a real large dataset demonstrate the feasibility of scans using a large number of markers, as opposed to MDR where the computer burden explodes with the number of markers (when it simply increases linearly with KNN-MDR). This might for example allow highlighting interactions between markers far apart on the genomic map (trans-interactions), while some strategies propose to restrict the scans to close-by markers (cis-interactions) or to markers with significant marginal effects to reduce the amount of computations.

For the second study, we also show that aggregating methods results is a strategy with interesting features for detecting epistatic interactions. Experimental results, based again on simulated and real

genome-wide data, show that the aggregated predictor can produce better performances, in terms of statistical power and false positive rates, than each individual predictor to detect genetic interactions. It is consequently a useful addition to the various methods available to tackle this complicated problem.

## Conclusion and Perspectives

In this dissertation, we focused on investigating and developing non-parametric statistical methods aiming at the detection of genetic interactions. We have shown that our novel methods complement, and sometimes improve, existing approaches used to detect genetic interactions in simulated and real datasets. The presented methodologies (KNN-MDR and aggregation of experts) are valuable in the context of loci and interaction mapping and can enhance the understanding of the biological mechanism underlying traits of interest, including diseases. More precisely, the new knowledge gained using these methodologies can assist in the prediction of clinical diseases and can contribute to provide new therapeutic opportunities.

To take further steps to these appealing perspectives, a first objective could be to implement a better version of the KNN-MDR software. The improvements could be on the overall performance of the software (optimization of the time-consuming parts of the program, parallelization), but also on the improvement of the “user-friendliness” of the program. This would involve an easier (and maybe automated) tuning of the parameters allowing an optimal detection power. These parameters include: the optimal sizes of the windows - which are dependent on the studied population, the markers density, the LD pattern, the optimal size of the neighborhoods to be considered, the pre-selection of markers in the early phase of large dataset analyses, the used distance measure or the adaptative selection scheme for the selection of markers in large studies, among others, the use of other types of genomic variants (microsatellites, copy number variations, sequencing data).
Another potential track would be to use a priori information on the interactions: this could be by using the results of previous studies, or by exploiting the known information on gene networks.

# *Résumé*

## Introduction

Ces dernières années ont vu l'émergence de sources riches d'informations génétiques au niveau moléculaire. Certaines des principales percées récentes en biologie proviennent de ces nouvelles connaissances, permettant l'application de nouvelles stratégies dans de nombreux domaines de la recherche biologique. Bien que les approches ciblant l'association entre les caractéristiques phénotypiques et les variations de l'ADN aient été couronnées de succès, de nombreux éléments dans le paysage génétique des caractères étudiés sont encore inconnus et non caractérisés. Une piste potentielle vers de nouvelles découvertes, qui pourrait aider à mieux comprendre les déterminismes complexes, est de détecter les interactions entre les régions plutôt que les associations avec une région unique. Alors que de nombreuses méthodes ont été proposées pour détecter de telles interactions et malgré le succès de ces méthodes pour résoudre certains problèmes et détecter certaines de ces interactions génétiques, il n'existe actuellement aucune méthode de référence capable de détecter les interactions dans toutes les situations. De plus, les méthodes restent relativement peu efficaces. Cette thèse est une contribution au développement de méthodes dans ce domaine.

Dans la première étude, nous proposons une nouvelle approche combinant les méthodes des K Plus Proches Voisins (KNN) et de Réduction Multidimensionnelle (MDR) pour détecter les interactions entre régions génomiques comme alternative possible aux algorithmes existants, notamment dans les situations où le nombre de déterminants impliqués est plus élevé que deux. Dans la deuxième étude, nous proposons une stratégie basée sur le principe de l'agrégation d'experts, où les experts seraient différentes méthodes de détection d’interactions validées et publiées dans des revues scientifiques.

## Résultats

Les résultats obtenus dans la première étude à la fois sur des données générées par simulation et sur des données réelles à l'échelle du génome démontrent certaines des caractéristiques qui rendent l’application du modèle KNN-MDR potentiellement intéressante en matière de précision et de puissance : dans de nombreux cas, il surclasse nettement ses concurrents. De plus, des analyses sur un large ensemble de données réelles démontrent la faisabilité d'analyses utilisant un grand nombre de marqueurs, par opposition à la méthode MDR où la charge informatique explose avec le nombre de marqueurs (alors qu’elle augmente simplement linéairement avec KNN-MDR). Cela pourrait par exemple permettre de mettre en évidence des interactions entre des marqueurs éloignés sur la carte génomique alors que certaines stratégies proposent de limiter les scans aux marqueurs proches ou à un ensemble de marqueurs préalablement sélectionné pour réduire la quantité de calculs.

Pour la seconde étude, nous montrons aussi que la méthode de l'agrégation des résultats est une stratégie avec des caractéristiques intéressantes pour détecter les interactions épistatiques. Les résultats expérimentaux, basés à nouveau sur des données simulées et réelles à l'échelle du génome, montrent que le prédicteur agrégé peut produire de meilleures performances que chaque prédicteur individuel pour détecter des interactions génétiques, et est donc un complément utile aux diverses méthodes disponibles pour résoudre ce problème compliqué.

## Conclusions et Perspectives

Dans cette thèse, nous nous sommes concentrés sur l'étude et le développement de méthodes statistiques non paramétriques pour la détection des interactions génétiques. Les méthodes que nous proposons sont présentées pour compléter et améliorer les approches existantes utilisées pour détecter les interactions génétiques dans des ensembles de données réelles et simulées. Les méthodologies présentées (KNN-MDR et agrégation d'experts) sont utiles dans le contexte de la cartographie des interactions et peuvent améliorer la compréhension du mécanisme biologique sous-jacent à divers caractères d'intérêt, y compris des maladies. L'acquisition de cette nouvelle connaissance, outre la compréhension fondamentale qu'elle implique, peut par exemple contribuer à la prédiction pronostique ou diagnostique des maladies étudiées, peut offrir de nouvelles possibilités thérapeutiques ou peut conduire à l'amélioration de caractères ayant un intérêt médical, agronomique, zootechnique ou autre.

Pour aller plus loin par rapport à ces perspectives attrayantes, un premier objectif pourrait être de mettre en œuvre une meilleure version du logiciel KNN-MDR. Les améliorations pourraient porter sur la performance globale du logiciel (optimisation des parties chronophages du programme, parallélisation), mais aussi sur l'amélioration de la "convivialité" du programme. Cela impliquerait un réglage plus facile (et peut-être automatisé) des paramètres permettant une puissance de détection optimale. Ces paramètres comprennent: les tailles optimales des fenêtres - qui dépendent de la population étudiée, la densité des marqueurs, le modèle de LD, la taille optimale des voisins à considérer, la présélection des marqueurs dans la première phase des analyses de grands ensemble de données, la mesure de la distance utilisée ou le schéma de sélection adaptatif pour la sélection des marqueurs dans les grandes études, entre autres, l'utilisation d'autres types de variantes génomiques (microsatellites, variations du nombre de copies, données de séquençage).

Une autre piste potentielle serait d'utiliser des informations sur les interactions: cela pourrait être possible en utilisant les résultats d'études antérieures, ou en exploitant les informations connues sur les réseaux de gènes.

# General preamble

Genetics laboratories activities have recently become more familiar in the public audience: forensics and DNA profiling are present in many prime-time shows and series, and genetic diseases research is nowadays largely advertised and sometimes supported through crowdfunding. The recent increase in the public interest for this science somehow reflects the huge advances made in genetics in the recent decades. Genetic technologies have indeed revolutionized our ability to explore the genetic architecture underlying complex traits and generated high (and sometimes exaggerated) hopes to understand the fundamental molecular mechanisms underlying biological processes, such as solving medical problems or improving the efficiency of bio-mechanisms underlying traits of economic importance. One of the disciplines involved to reach these long-term perspectives is positional cloning of genes. The aim of this technique is to identify genomic regions underlying traits of interest based only on the phenotypes and the genotypes of individuals for a panel of molecular markers. In this field, breakthroughs in the genotyping and sequencing technologies - such as DNA markers microarrays and NGS techniques - have made association studies based on the whole genome affordable in many species and populations. This new situation of large molecular data availability was promising and expectations were high that many new insights would readily become available to scientists. Despite many successes in the last two decades, much work remains to be done. As an example of the progresses to be made, the genetic variants identified to date in most genome-wide association studies only explain a small part of the total heritability of the studied traits. Although other explanations are possible, genetic interactions (epistasis) is one potential important source of unexplained variability. Consequently, further investigations in the field of interactions mapping in large-scale studies seems a reasonable avenue of promising research.  Our work is a contribution to this field.

Throughout this thesis work, we have aimed at presenting statistical non-parametric methods for identifying potentially epistatic interactions from genomic (and sometimes genome-wide) data. We have assessed the mechanisms and the main characteristics of these new methods and we have tried to provide some evidence for the utility of these methods over simulated and real data.

More specifically, in the first study, we propose a novel approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) methods for detecting gene-gene interactions. This method is an extension of the well-known MDR methodology. It increases the span of the possible situations where MDR can be useful to situations with large number of markers and when the number of underlying genetic determinants is potentially higher than two. The way we use the data in KNN-MDR is shown to have a positive impact on the computational burden, making accessible situations that could not be tackled using the classical MDR techniques. Furthermore, and as a side effect, the approach we propose is also shown to be more powerful and accurate in difficult situations where individual genes have only minor (or no) marginal effect and where genetic heterogeneity - i.e. different genotypic configurations leading to the same phenotype, and the same genotype leading to distinct phenotypes - is present. A comparison of our method (KNN-MDR) to a set of the other most performing methods has been carried on to detect interactions using simulated data as well as real genome-wide data. Experimental results on both simulated data and real genome-wide data show that KNN-MDR has, as mentioned, interesting properties in terms of accuracy and power, and that, in many cases, it performs better than its recent competitors.

In a second study, we propose using a method based on the principle of the aggregation of experts, where the experts would be a set of popular published methods. The rationale of the aggregation strategy we propose is to benefit from the synergistic work of known methods, each with different strengths and weaknesses, to produce more reliable results than each of these individual methods. Our work shows that this strategy might lead to increases in both detection power and accuracy in the genetic interactions problem, while properly controlling for false discoveries.

In summary, our contribution to the hunt for genomic interactions underlying phenotypic traits is to provide one non-parametric method and one strategy allowing to improve the detection characteristics, and to show how these approaches could be used on today large real datasets.

# Introduction

## 1. General introduction

These last years have seen the emergence of a wealth of biological information and a steep increase in the rate of development of genomic and other basic biological research. Facilitated access to the genome sequence, along with massive data on genes expression and on proteins have revolutionized the research in many fields of biology (Visscher et al. 2012). The development of efficient genomic tools has allowed unraveling a large share of the molecular variation in many species, paving the way for studies aiming at associating genomic polymorphisms to phenotypic variation. An instance of this process is the use of panels of single nucleotide polymorphisms (SNPs) in large scales studies to track genes potentially involved in complex traits such as human, animal or plant diseases, for example (Kadarmideen 2014). Molecular analyses are nowadays commonly performed to examine candidate genomic regions or even the whole genome (in so-called "genome-wide association studies" (GWAS)) for causative genomic variants (Katsanis et al. 2013). The knowledge of these influential regions is of particular interest, since they are likely to harbor important genes involved in the onset of the disease and provide clues to the underlying mechanisms. Although these analyses are progressively becoming widespread, and despite successes in studies targeting for example diabetes (Frau et al. 2017) or Crohn disease (Libioulle et al. 2007), a large part of the genetic landscape of most traits is still unknown and uncharacterized, with in many cases the tested genetic variation only explaining  less than 5%-10% of the risk of the disease (Riancho 2012). (Visscher et al. 2012) (Korte et al. 2013) suggested that this low figure could be due to the presence of a large number of different genetic causes and to potential interactions between genes. Consequently, a deeper understanding of the genotype to phenotype relationships will necessitate much more work in many situations (Yee et al. 2016).

In this thesis, we have tried to elaborate methods aiming at discovering simultaneous factors acting on the onset of the disease as an alternative to methods targeting single regions. Although we have focused on genomic regions, such methods could also encompass situations where genes and environment interact to produce the observed phenotypes. The main reason for that choice is that many signs indicate that interactions of several genes to underlie many traits might be the rule rather than the exception (Stanislas et al. 2017). Firstly, from a purely biological point of view, most genes are involved in complex networks where they interact with other genes; changes (mutations) in one gene might have or not an impact on the behaviour of the network, and several simultaneous mutations might be necessary to change the products of the network. Consequently, many scenarios are possible, some of which suggesting epistatic interactions between genes (Zou et al. 2017). Second, from a more pragmatic point of view, the mapping effort using single regions, although sometimes successful, have often failed to demonstrate genotype to phenotype relationship. This might be due to a lack of statistical power, as has been suggested, or to a poorly specified model (or to both): due to interactions, a gene might mask the effect of another gene, preventing to associate clearly this second gene to the studied phenotype (Jung et al. 2016). The next section illustrates this situation.

### 1.1. An example of interaction

Various mechanisms of interactions exist and examples of each of these mechanisms can be found in the genetic literature (Costanzo et al. 2016). We will use the "complementary gene action" to illustrate the principle and to explain the difficulties for gene mapping due to this determinism. A classic example of this type of interactions is the sweet pea flowers colour problem: when crossing two parental white coloured lines, researchers obtained a completely purple F1 line. Next, when generating the F2 line (i.e. crossing the F1 individuals), an unexpected ratio of 9:7 purple-coloured to white-coloured flowers is observed. The explanation is as follows:

- The determinism involves two genes, with two alleles each, noted A and a for the first and B and b for the second.
- The parental lines have (fixed homozygous) genotypes AAbb and aaBB, respectively.
- All F1 individuals are thus AaBb.
- If the genes are on distinct chromosomes (or sufficiently far apart on the same chromosome), four types of gametes are equally likely: AB, Ab, aB and ab.
- These 4 gametes lead to 16 equally likely genotypes in the F2 population, summarized in the following table:

| | AB | Ab | aB | ab |
|---|---|---|---|---|
| AB | AABB | AABb | AbBB | AaBb |
| Ab | AABb | AAbb | AaBb | Aabb |
| aB | AaBB | AaBb | aaBB | aaBb |
| ab | AaBb | Aabb | aaBb | aabb |

Table 1 - an example of interaction. Involved genes show a "complementary gene action": both dominant alleles are needed to obtain one of the phenotypes (purple colour, here)

So, obtaining the purple color necessitates that both A and B alleles be simultaneously present. The underlying genes are said to be "complementary". This type of behavior has strong consequences on mapping experiments: imagine that a set of flowers is collected and that purple and white plants are genotyped in order to identify the genomic regions involved in the color determinism. Mapping single regions would probably fail to identify the individual genes (for example, some AA plants are white, but some other AA plants are purple), while using 2 simultaneous regions would probably identify the 2 genes (all A-B- plants are purple, while all other genotypes lead to white flowers).

## 1.2. Definition of genetic interactions

A gene interaction is an interplay between multiple genes that has an impact on the expression of an organism's phenotype (Costanzo et al. 2016). In this work, we will only consider interactions between genes, although other types of interactions are possible, such as for example the dominance (interaction between the alleles of a single gene) or interactions between proteins. The term gene-gene interaction is also known as epistasis or genetic interaction (Moore et al. 2005). We showed above an example of interaction, but various other types are possible. Three well-known examples are:

- Recessive epistasis: when the recessive allele of one gene masks the effects of either allele of the second gene (Marcelo et al. 2005). An example of this is the coat colour in Labrador retriever (Schmutz et al. 2007): one gene codes for pigment production (B) and the other for diffusion of the pigment into the air shaft (E). Mutations in E (e) leads to no diffusion of the pigment in the coat, no matter whether black (B) or brown (b) pigments were produced: all individuals carrying the recessive ee genotype will end up as golden coat. This is summarized in Table 2.

| | BE | Be | bE | be |
|---|---|---|---|---|
| BE | BBEE | BBEe | BbEE | BbEe |
| Be | BBEe | BBee | BbEe | Bbee |
| bE | BbEE | BbEe | bbEE | bbEe |
| be | BbEe | Bbee | bbEe | bbee |

Table 2 - an example of recessive epistasis. The possible genotypes are displayed and the background colour corresponds to the dogs coat colour.

- Dominant epistasis: when the dominant allele of one gene masks the effects of either allele of the second gene (Marcelo et al. 2005). An example is the summer squash, where the colour of the plant is due to 2 genes. If the dominant allele of the second gene (B) is present, the squash will be white no matter the genotype at the first gene. If the genotype at the second gene is the recessive one (bb), then the colour will depend on the presence of the dominant allele at the first gene (A): homozygous (AA) or heterozygous (Aa) individuals will be yellow, while recessive homozygous (aa) plants will be green. This is summarized in Table 3.

| | AB | Ab | aB | ab |
|---|---|---|---|---|
| AB | AABB | AABb | AaBB | AaBb |
| Ab | AABb | AAbb | AaBb | Aabb |

| | | | | |
|---|---|---|---|---|
| aB | AaBB | AaBb | aaBB | aaBb |
| ab | AaBb | Aabb | aaBb | aabb |

Table 3 - an example of dominant epistasis. The possible genotypes are displayed and the background colour corresponds to the summer squash colour.

- Redundant genes: when a gene with a dominant allele is duplicated (and this is also true when genes are replicated several times), only double (multiple) recessive individuals will display the recessive phenotype (Nowak et al. 1997). An example is the snapdragon flower colour, which is red when a dominant allele is present, and white if not. This is shown in Table 4.

| | AB | Ab | aB | ab |
|---|---|---|---|---|
| AB | AABB | AABb | AaBB | AaBb |
| Ab | AABb | AAbb | AaBb | Aabb |
| aB | AaBB | AaBb | aaBB | aaBb |
| ab | AaBb | Aabb | aaBb | aabb |

Table 4 - an example of redundant genes. The possible genotypes are displayed and the background colour corresponds to the plant flowers colour.

## 1.3. Is epistasis important?

There is a debate between those claiming that interactions contribute an important share to the genetic variation, and those who consider that the phenomenon is of minor importance to explain that variation. A first remark is that a distinction should be made between additive and total genetic variations (i.e. including non-additive effects, such as epistatic effects): the first leads to the so-called narrow-sense heritability — where is the additive genetic variance and is the phenotypic variance, and ——— where is the genetic variance due to non-additive effects (dominance, epistasis) (Mackay and Moore, 2014). Therefore, the relative importance of the non-additive variance in the genetic determinism of the traits is debated. In (Hill et al., 2008), it is argued that most of the genetic variation is additive. Since additive effects are transmitted from each of the parents to the descendants, while non-additive affects are not (they are rebuilt from the new combinations arising from the new combination of gametes), this is of course of special importance for breeders, who will mostly select on additive values. Nevertheless, other authors show that considering non-additive effects could improve prediction accuracy in situations where the underlying determinism is largely or partially due to epistatic interactions (Morgante et al., 2018), (Carlborg and Haley, 2004). For most complex traits, the determinism is largely unknown and the presence of epistatic interactions cannot be a priori discarded. Consequently, our view is that unravelling such interactions might contribute, in variable proportions, to a better knowledge of complex traits. This view has been illustrated in the previous section.

## 1.4. Position of the problem.

As mentioned above, the massive amount of available molecular information did not allow, in many applications, to unravel the exact relationship between the genomic configuration, including the interactions between the involved genes, and the phenotypic expression (Fuxman Bass et al. 2016). The failure of “simple” association models led to try to associate observed variations at the macroscopic level (phenotype) to identified variations and their interactions at the molecular level (Hu et al. 2011).

This approach introduces at least two challenges:

1. the genetics underlying most traits of interest is complex and probably involves most of the time many genes and many interactions between these genes, leading to a complex relationship between genomic variants and phenotypes. Properly modelling such intricate network of genes and interactions is a potentially very challenging task. Consequently, identification of every (or even of any) interaction is a potentially very difficult aim.
2. from a more statistical point of view, fully modeling the underlying genetic complexity leads to models with large dimensionality, causing the well-known ‘curse of dimensionality’ problem: higher complexity corresponds to larger sets of parameters to estimate, to larger search spaces and to the need for huge collection of observations to efficiently scan these search spaces and accurately estimate the parameters with sufficient power.

In our work, we have investigated the use of non-parametric modelling as an alternative to parametric methods to solve these problems. One of the reasons under this choice is that many methods, described below, have been developed with some success using that approach. Another one is that the problems linked to the estimation of the parameters in parametric approaches could make these estimations less affordable in models involving interactions, and therefore render the use of such models more questionable (Ma et al. 2011).

Note nevertheless that increasing the number of parameters to be identified, although potentially making the power issues developed above even more critical, might also lead to more accurate models of the underlying genetics by introducing interactions (Maity et al. 2011). Better models of the underlying genetics might in turn improve the detection power of the effects of interest. Consequently, it is not necessarily obvious that interaction models will present poor power when compared to non-interaction ones, which should motivate more research on this subject.

## 1.5. Interest of the work

The knowledge of the relationship between variations at the molecular or cell level and phenotypic variation is of major importance from various points of view. It is of fundamental interest to understand how subtle molecular variations lead to various phenotypes and to be in a position to dissect complex mechanisms into small manageable pieces, allowing to cope with the inherent complexity underlying a trait of interest. Applied aspects are most of the time at least as important as fundamental ones: a better understanding of the genetic components and of the mechanisms leading to some diseases might give some keys to potential therapies, and the discoveries of molecular mechanisms at the root of quantitative traits of interest, such as pathogen resistance or animal production, might assist the breeders in the production of more robust and sustainable animals or plants.

## 1.6. Biological background

Unlike Mendelian diseases, in which disease phenotypes are largely driven by mutations in one or two gene loci, complex diseases such as rheumatoid arthritis and many cancers are influenced by a complex interplay of genetic and environmental factors (Quintana-Murci 2016). The examples provided above should have explained what makes the interacting factors hard to discern, and this gets of course even truer when the interacting genes are unknown.

Genome-wide association studies (GWAS), in which several hundred thousands to more than a million single nucleotide polymorphisms (SNPs) are assayed in thousands of individuals, represent a powerful tool to study the genetic architecture of complex diseases (Visscher et al. 2012). During the past few years, these studies have identified hundreds of genetic variants associated with complex diseases and have provided valuable insights into the complexities of their genetic architecture (Manolio et al. 2009). Nevertheless, most variants identified so far have been found to confer relatively small information about the relationship between the genomic variants and the phenotypes because of a lack of reproducibility of the findings, or because these variants most of the time explained only a small proportion of the underlying genetic variation (Fang et al. 2012). This observation has been quoted as the ‘missing heritability’ problem (Manolio, et al. 2009). Moreover, hundreds of studies have searched for gene-gene and gene-environment interaction effects in GWAS data with the underlying motivation of identifying or at least accounting for potential biological interactions. So far, this quest has been mostly unsuccessful (Aschard 2015). We therefore developed statistical methods to contribute to address this problem.

### 1.6.1. Molecular markers

Genetic mapping models rest on molecular markers to serve as proxies of neighbouring genes: detected significant interactions between markers will be interpreted as potential interactions between genes close to these markers (Moore et al. 2005). In this context, “close to a marker” means “in linkage disequilibrium (LD) with the marker” (although other reasons, such as genetic drift or selection, might also lead to LD without the need for the gene and the associated markers to be physically close). Any DNA polymorphism is eligible as a molecular marker. This includes microsatellites, copy number variations, insertions/deletions, single nucleotide polymorphisms, among others. We next describe a few of these polymorphisms.

### 1.6.1.1. Single nucleotide polymorphisms

A single nucleotide polymorphism, or SNP (pronounced "snip"), is a variation at a single nucleotide position in a DNA sequence, as exemplified in Figure 1: the DNA sequences in the 2 pieces of DNA are identical except for a nucleotide, where they differ. Most SNP are biallelic, with a vast majority exhibiting either C/T alleles or A/G alleles. These distinct alleles can be present in a single individual (making this individual heterozygous for the SNP) and/or throughout the population. The frequency of the minor allele (the less frequent one) varies from SNP to SNP, with values ranging from close to 0 % up to values close to 50 %. On average, human DNA consists of a SNP for every 300 bases, meaning that, for the whole genome (3 billion bases), there would be roughly 10 million SNPs (Liu et al. 2015).

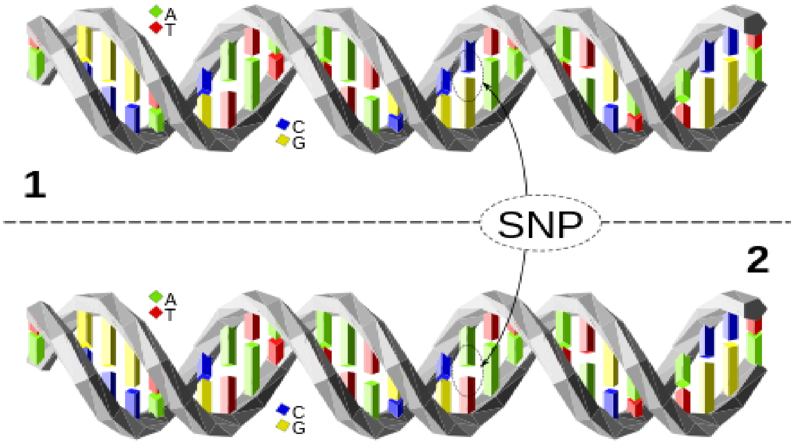

Figure 1 - a SNP. The two molecules of DNA only differ at one nucleotide

### 1.6.1.2. Microsatellites

A microsatellite, or Single Sequence Repeats (SSRs), is a group of repetitive DNA in which certain DNA motifs (1 to 10 nucleotides) are repeated, typically 5–50 times. They tend to occur at thousands of locations within an organism's genome. They have also a higher mutation rate and a higher genetic diversity than other areas of DNA, which makes them a good candidate as a genetic marker (Vieira et al. 2016).

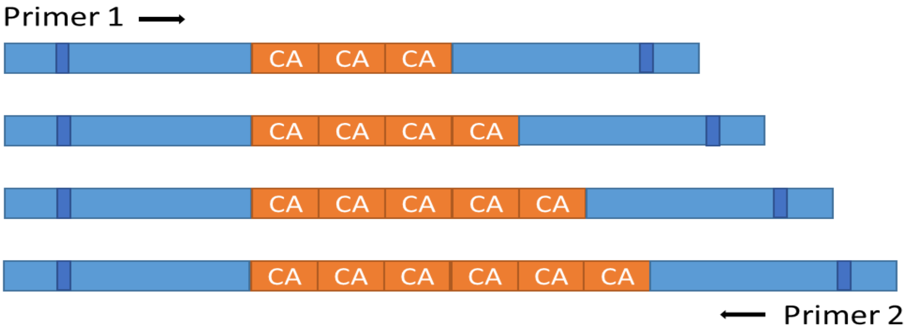

Figure 2 - a microsatellite marker. The number of copies of the CA nucleotides tandem varies from copy to copy.

### 1.6.1.2. CNV (Copy Number Variations)

A copy number variation is a phenomenon in which large regions of the genomes are either duplicated or deleted, creating structural variant regions. The supplementary copies can involve fairly large stretches of DNA (sometimes thousands of nucleotides). The regions with such variations cover a relatively large portion of the genome (up to 10% of the human genome, for example) (Thapar et al. 2013). These structural variants provide a support for the evolution of genes to new functions, but can also be causative of disease.

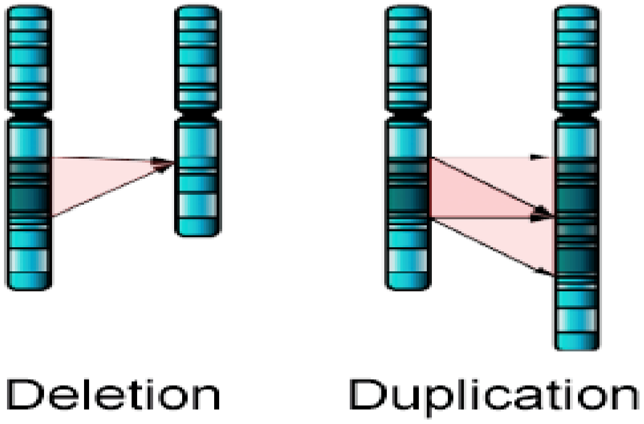

Figure 3 - a Copy Number Variation.

### 1.4.1.2. Insertions and deletions (INDELs)

Small insertions or deletions - commonly called INDELs - are another important source of genetic polymorphisms. In terms of base pairs of variations, INDELs cause similar levels of variation as SNPs (Mullaney et al. 2010).

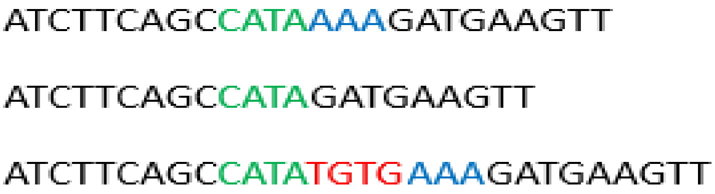

Figure 4 - INDELs. Using the first sequence as a reference ("wild type") sequence, 3 base pairs (AAA, in blue) have been deleted in the second sequence (deletion), and 4 base pairs (TGTG, in red) have been inserted in the third sequence.

### 1.7. Statistical background

A statistical model is an attempt to provide a mathematical abstraction of the mechanisms that produced the observations. The model makes some assumptions and some simplifications of the reality in order to make things still manageable in terms of the involved mathematics and of the computing burden while still providing a hopefully useful view of the studied phenomenon (Calzone et al. 2015). The goal is to be able to properly describe the state of nature and to make accurate predictions. In our mapping context, the models should be able to identify pieces of the genome that are involved in the studied trait. Although several taxonomies exist for the models, we will concentrate in the following paragraph on the distinction between parametric and non-parametric models, which is important to describe our work.

#### 1.7.1. Parametric models

A parametric model is a family of distributions such that each member of the family can be described using a finite set θ of parameters and all the parameters are in finite-dimensional parameter spaces. An example of such families is the normal family, indexed by the 2 parameters μ (the mean, which is also the expected value of the modeled variable) and σ (the standard deviation, the squared root of the expected squared distance to the mean). The classical normal distribution formula:

where μ can take any real value and σ can take any positive value, thus defines a family of distributions, used to model a variable (and, more generally, a set of variables) assumed to originate from the well-known bell-shaped distribution. Such models are widely used, including in the field of genetic mapping. Most common applications targeting genes and quantitative trait loci mapping strategies include linear regression, logistic regression, classical and generalized mixed models, among others (Howard et al. 2014).

Common features of parametric models are:

- The data is summarized by a fixed set of parameters (Jackson 2016).

- They perform correctly with relatively small dataset and can avoid overfitting due to the a priori imposed structure (Sun et al. 2017).
- They are optimal ("best") when correct parameters are chosen (Elster et al. 2005).
- Although they make stronger assumptions about the data, they work well if the assumptions are correct (Goodrich 2012).

These models generate several statistical challenges. For example, in the "genomic selection" procedures, the phenotypes of interest are modeled as a sum of marker effects added to a sum of other (non-genetic) effects. Since many markers (p) are available on a restricted set of (n) individuals, leading to much more unknown than data points, techniques such as the LASSO (Least Absolute Shrinkage and Selection Operator) or Bayesian techniques must be used to solve such problem, leading to more difficult interpretation of the effects of the variables (Howard et al. 2014). Another challenge is that traditional parametric methods need strong model assumptions to model interactions, such as assuming linear G×G interaction. This assumption, however, could be easily violated due to the underlying nonlinear machinery between the genetic factors. As mentioned above, misspecification in parametric models could lead to large bias (Ma et al. 2011, Maity et al. 2011).

### 1.7.2. Non-parametric models

Unlike parametric models, non-parametric models are not based on parameterized families of probability distributions. Nonparametric statistics make no assumptions about the probability distributions of the variables being assessed, and they can take big dimensional parameter spaces (Li et al. 2013).

In short, in non-parametric models:

- The data is summarized using an unknown set of parameters (Hamilton et al. 2017).
- Some of the original data must be kept to make predictions or to update the model (Ghahramani 2012).
- No assumption is made about probability distributions (Ghahramani 2012).
- Models are generally slower, but potentially more accurate, especially when the assumptions made in parametric models are questionable (Ho et al. 2017).

In the interaction mapping field, many non-parametric methods have been devised (Support Vector Machine, Neural Networks, Random Forest, k-nearest neighbors algorithms, ...). These nonparametric models have the potential to bring new solutions for the challenges in the domain (Gianola et al. 2006) because of their ability to handling multiple genetic variants with the consideration of possible high-order G-G interactions, and because they do not make any assumption on the disease models (Li et al. 2013, Howard et al. 2014, Li et al. 2014).

### 1.7.3. Modelling interactions

The regulatory interactions in genetic networks form a complicated system and an important objective of systems biology is to model and infer these interactions. Proper modeling and inference of these genetic interactions requires understanding of the distinction between biological and statistical interactions (Forsberg et al. 2017).

#### 1.7.3.1. Statistical interactions

The most common statistical definition of interaction relies on the concept of a linear model describing the relationship between some outcome variable and some predictor variable(s). The case of statistical interaction potential arises when there are two or more independent variables. The simplest case is when the effect of each independent variable is completely separate from the other independent variables. In this “no interaction” setting, the effects of the different independent variables act just additively. A more complicated situation arises when the effect of one independent variable depends on other independent variable(s). This is referred to as an "interaction" situation (Moore et al. 2005, Cordell 2009). Since, in this context, the effect of a variable cannot be obtained without considering other potentially interacting variables, it turns out that this situation can significantly complicate certain types of multivariate analyses with respect to situations where variables are assumed to be independent of each other (Yi 2010).

#### 1.7.3.2. Biological interactions

Examples of biological interactions have been presented in a previous section. In our genetic mapping context, biological interactions mean in the widest sense the effect of a particular genotype on the phenotype depends on the genetic background. It can be defined in a simple, as the phenotypic effect of one locus depends on the genotype at the second locus (Carlborg and Haley, 2004). Also, it can be defined in a general, as the effect of a gene on a phenotype is dependent on the configuration of one or more other genes (Moore et al. 2005). In this context, many epistasis were detected as responsible for differences in phenotype such as comb type in chickens and coat color in various animals (Carlborg and Haley, 2004).

## 2. Use of non-parametric statistical methods in genetic interaction mapping applications

Various non-parametric methods have been used in genetic problems. In this section, we will review some of the most important methods in this field. Although notations in the following descriptions are mostly borrowed from the papers at the roots of the described methods, the data structures used by the methods are similar:

- A set (ranging from a few hundreds to a few thousands, in most situations) of observations, each observation corresponding to a subject (human, animal or plant) used in the experiment.
- For each subject (i.e. observation), we have at least:
  - A phenotype, to be seen as the dependent variable of interest. This phenotype, measured on the subject, can either be discrete (as, for example, in case-controls experiments, where the phenotype is either 0 (control) or 1 (case)) or continuous (blood pressure, or annual milk yield, for example).
  - Genotypes, to be considered as the putative explaining variables. These genotypes have been obtained from the laboratory based on DNA samples originating from the subjects of the experiment. These genotypes are coded as discrete values representing the various possible genotypic configurations. For example, a SNP genotype could be coded as 0, 1 or 2 to represent the various possible genotypes (AA, AB or BB) for that SNP. The number of genotypes, typically corresponding to the number of markers used in the experiment, is generally large (from a few dozens to a few millions) and, in most situations, larger than the number of subjects.

One of the goals of the experiment is then to try to identify the set of genotypes with a significant impact on the studied .phenotype.

### 2.1. Support Vector Machines

Support vector machines (SVMs) are supervised non-parametric statistical learning techniques that analyze data and recognize patterns. They are used for classification and regression analyses. There is no assumption made on the underlying data distribution, which is utilizing hyperplanes in high dimensional spaces (Mountrakis et al. 2011).

#### 2.1. 1. Methodology of Support Vector Machine

Let us assume an input (features) - output (classes) process producing paired data where is a p-dimensional vector of features measured on a sample, that has been classified into a class (coded -1 or 1) represented by , . The training process of a (linear) support vector machine aims to find a linear separating hyperplane with the maximal margin (2 , the distance between and ) under the classification conditions:

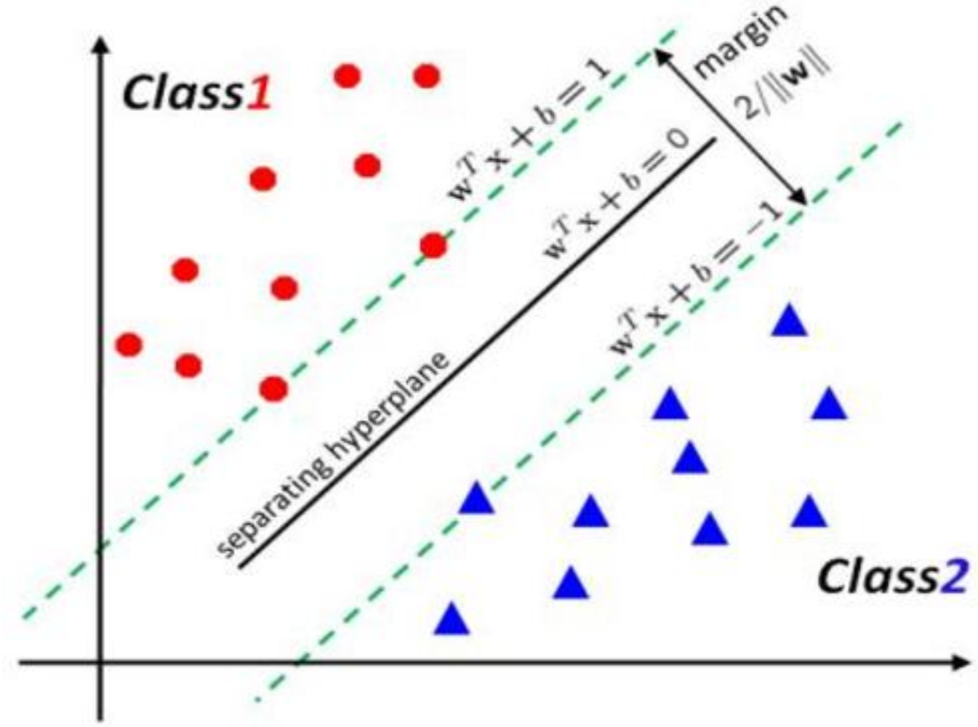

Figure 5 - Linear SVM with maximum-margin hyperplane (Agrawal, et al., 2012).

In Figure 5, a hyperplane separates the two classes of input, where is the normal vector, is the bias, and “ ” is the dot product (Chen et al. 2008) (Koo et al. 2013).

Two additional hyperplanes separate the data from the previous hyperplane (defined by ) with no data between them. The additional hyperplanes are located at a maximum distance (known as margin) from the separating hyperplane. These hyperplanes have equations and , and are such that all points for which are from the first class and those for which are from the second class.

Furthermore, in Figure 5, the distance between two hyperplanes is equal to , and the offset of the separating hyperplane from the origin along the normal vector is determined by . Overall equation for the additional hyperplane can be written as

When the data points clouds overlap, a solution is to map the input vector data into higher dimensional space, known as the feature space, so that the linear separation can be achieved within that space (Koo et al. 2013).

SVM can also be extended to non-linear separating hyper surfaces using “the kernel trick”: the original input space is mapped into a high dimension space (the feature space) using a kernel function that is defined as where is kernel function that map input space into feature space (Sheng et al. 2014) and SVM is applied to these transformed couples.

Several kernel functions have been proposed in SVM to obtain the optimal solution; the most frequently used such kernel functions are (linear kernel is the simplest kernel function, given by the inner product (x,y)), (polynomial kernel is a nonstationary kernel that it is well suited for problems where all the training data is normalized and where (the polynomial degree) and are kernel parameters (Chen et al. 2008) and (Sheng et al. 2014)), and (radial basis function)).

### 2.1.2. Application of Support Vector Machines to the detection of gene-gene interactions

SVM have been used to predict genetic interactions, which can be learned from the features of known genetically interacting pairs in order to predict which other pairs genetically interact.

In order to achieve this, the training data consists of two sets of features vectors, each set labelled as either positive (corresponding to the presence of genetic interaction) or negative (corresponding to the lack of genetic interaction). Each features vector characterizes a pair of genes rather than a single gene. When the features are mapped into a high-dimensional space, the SVM constructs a separating hyperplane that maximizes the margin between the features of genetically or not genetically interacting pairs. For this mapping, (Koo et al. 2013) used kernel function such as polynomial or radial basis.

In (Fang and Chiu, 2012), the authors have proposed an extended SVM method and a SVM based pedigree-based generalized multifactor dimensionality (PGMDR) for detecting gene-gene interactions in the absence or presence of the main effects of genes with an adjustment for covariates and on a limited sample of families. The results show that the proposed approaches of SVM and SVM-based PGMDR have higher power than other methods (PGMDR and FAM-MDR (family-based multifactor dimensionality reduction)) used for comparisons. In addition, although more computationally expensive than the other methods, these methods show higher prediction accuracy and power, making them valuable for the interactions detection problem.

In (Fang et al. 2013), the authors have also developed a novel approach named "backward support vector machine (BSVM)-based variant selection procedure" to identify informative disease-associated rare variants. The idea of this approach is that the rare variants are weighted and selected according to their positive or negative associations with the disease. The results on both simulated and real data show that the proposed BSVM approach is more powerful than the other approaches used in this study (such as set Kernel Association Test (SKAT)) .

In (Chen et al. 2008), the authors have also proposed SVM methods in various situations to detect gene-gene interactions and compared this approach to MDR (MDR is described below). The results show that SVM methods are a useful tool for the identification and characterization of high order gene-gene and gene-environment interactions but is computationally more costly than MDR.

### 2.1. 3. Strengths of Support Vector Machine

(i) SVM can deal with high dimension data set (Upstill-Goddard et al. 2013).

(ii) SVM can be utilized to classify complex biological gene expression data (Ban et al. 2010).

(iii) SVM is robust to noise and not prone to overfitting (Ozgur et al. 2008).

### 2.1. 4. Weaknesses of Support Vector Machine

(i) SVM is restricted to pairwise classification (Chen et al. 2008).

(ii) SVM cannot be directly used for features selection (Mountrakis et al. 2011).

(iii) The power of SVM might be reduced in the presence of genetic heterogeneity (Chen et al. 2008).

(iv) Computationally intensive (Wang et al. 2008).

## 2.2. Neural Networks

Another computational approach proposed for the study of disease susceptibility genes is neural networks (NN). Neural Networks are a class of pattern recognition methods developed in the 1940's to model the neuron, the basic functional unit of the brain. Neural networks process information in a way similar to the human brain. It consists of a large number of highly interconnected processing units (neurons) working in parallel to solve a specific problem (Motsinger-Reif et al. 2008).

### 2.2. 1. Methodology of Neural Networks

Single neuron model (also known as perceptron) is the basic neural model in neural networks. In this basic model, a neuron consists of a set of weighted inputs producing a single output. In Figure 6, the model is represented through multiple inputs, sent through connections providing the weighting ……, ) (Koo et al. 2013).

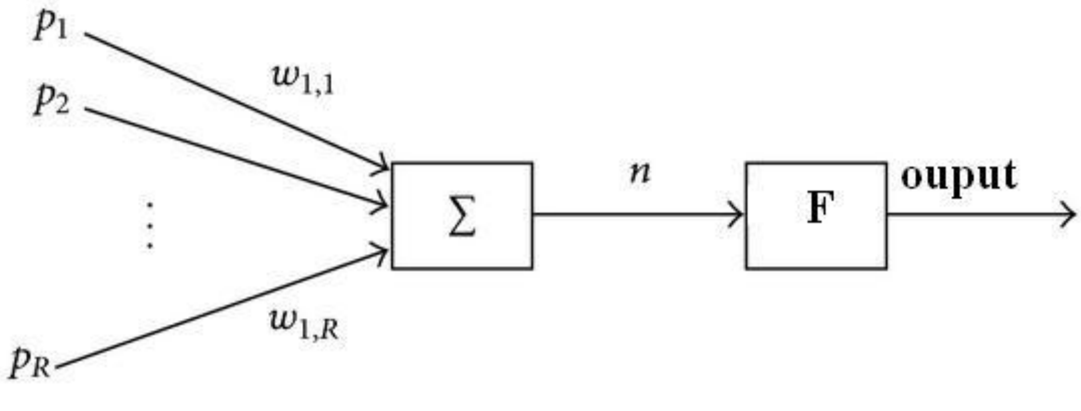

Figure 6 - Basic neural model (Koo et al. 2013).

The generated output is computed in two steps. A weighted sum of the inputs is first calculated using:

(1)

Next, the weighted sum is compared to a threshold. For example, a F(x) Heaviside activation function is used when the weighted sum of inputs is compared to a null threshold, where F(x) is defined as

(2)

The perceptron, classifying individuals (represented by an input vector) into an output class (0 or 1 in the example given above), can serve as the basic building block for an artificial neural network (ANN), which is a more general classifier

As a simple example of ANN, a feed-forward network was the first type of used artificial neural network. It contains multiple neurons (perceptrons) arranged in layers. Perceptrons from adjacent layers have connections or edges between them Figure 7. All these connections have weights associated with them (Konomi et al. 2017).

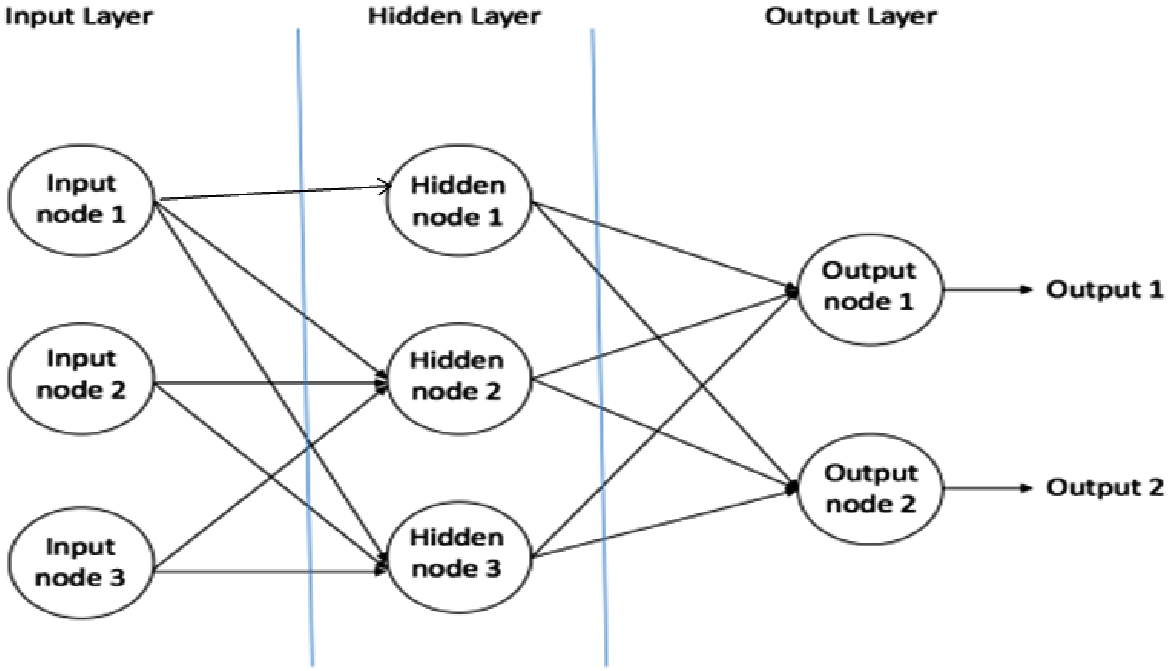

Figure 7 - An example of feedforward neural network.

From Figure 7, the outputs from the input layer are used as inputs for each node in the hidden layer. Similarly, the outputs of the nodes in the intermediate layer(s) serve as inputs for the next layer, propagating the signal down to the output layer. Classically, multilayer ANN consist of three layers or more, including an input layer, an output layer, and one or more hidden layers. Each node in one layer connects with varying weights to every node in the following layer, and the transfer function F(x) is very commonly a sigmoid function. Note that in the example presented here, the information moves in only one direction (feed forward). No cycle sending information from outputs to previous layers of the neural network is included.

### 2.2. 2. Application of Neural Networks in the detection of gene-gene interactions

Neural network methods are used to identify disease susceptibility genes in both linkage and association analyses. Although both types of analyses have the same objective - i.e. identifying markers significantly associated to loci involved in the trait of interest -, the approaches differ: in linkage analysis, information from the pedigree and from the genotypes is used to follow the segregation of the trait in the pedigree and detect associations between the genotypes and the trait. In association analyses, this link is sought using individuals randomly sampled from the investigated population, and no pedigree is used. In disease mapping experiments, the sampling is stratified and samples are collected in cases and controls sub-populations (Curtis 2007). A consequence is that regions detected using association analyses are generally smaller than those detected using linkage analyses, but require denser markers maps. In general, in the genetic mapping context, the genotypes serve as input and the phenotype is the output of the neural network (Koo et al. 2013).

Various coding schemes are possible for the inputs and output of ANN. For example, the inputs can be the presence or absence of a specific marker allele (a value of 1 would represent the presence of the allele, and a value of 0 an absence of the allele). Another common encoding strategy for the inputs of a neural network is to use identity-by-descent (IBD) status of the genotypes: variable is set to 1 when the alleles in a genotype are supposed to be IBD, to $-1$ when not and to 0 when the genotype is uninformative. On the other side, several coding are also possible for the outputs of neural networks. For example, the output could be the disease status, in which a value of 1 would represent a case whereas a value of 0 would indicate a control (Motsinger-Reif et al. 2008).

In (Tomita et al. 2011), the authors have proposed artificial neural networks (ANN) for the detection of gene-gene interactions. The idea of this study is based on the use of artificial neural networks with the parameters decreasing method (PDM). The procedure of PDM begins by excluding one SNP from the total number of SNPs and constructs a model containing the remaining SNPs. In turn, each SNP is deleted from the total number of SNPs and with the remaining SNPs a model is constructed. The results demonstrate that the artificial neural network approach had more power than logistic regression (LR) to characterize the development of complex diseases such as an allergic disease.

In (Gunther et al. 2009), the authors have also proposed neural networks for the detection and the modelling of various types of gene-gene interactions. In their study, the authors used feed-forward multilayer perceptron (MLP) as a neural network, given that this method is able to approximate arbitrary functional relationships between covariates and response variables. The results on simulation data demonstrate that neural networks have more ability to detect and model different types of biological gene-gene interactions than others methods (logistic regression and MDR) which were used for comparison in the study.

### 2.2. 3. Strengths of Neural Networks

(i) NN are able to model the relationship between disease and single nucleotide polymorphism (SNP) (Tomita et al. 2011).

(ii) NN can make prediction on data where the disease outcome is unknown by learning the outcome given on a dataset (Basheer et al. 2000).

(iii) NN can deal with large volumes of data (Ritchie et al. 2003).

(iv) NN are still efficient in the presence of genetic heterogeneity, high phenocopy rates, polygenic inheritance, and incomplete penetrance (Motsinger-Reif et al. 2008).

### 2.2. 4. Weaknesses of Neural Networks

(i) NN work as a black box (Motsinger-Reif et al. 2008).

(ii) Difficult to list out all possible NN architectures, which causes the difficulty to find the optimal architecture (Basheer et al. 2000).

(iii) Result of NN are hard to interpret due to the dimensionality problem (Ritchie et al. 2003) (Curtis 2007).

## 2.3. Multifactor Dimensionality Reduction

MDR has enjoyed great popularity in the field of interaction mapping and a vast amount of extensions and modifications of the original method (Ritchie et al. 2001) have been suggested and applied, building on the general idea (Gola et al. 2015). It is a data mining approach for detecting and characterizing combinations of attributes or independent variables that interact to influence a dependent or class variable (complex gene–gene and gene–environment interactions) (Martin et al. 2006). The MDR method is nonparametric (i.e., no hypothesis about the distribution of statistical parameters is made), is model-free (i.e., it assumes no particular inheritance model), and is directly applicable to case-control and discordant-sib-pair studies (Manuguerra et al. 2007).

### 2.3. 1. Methodology of Multifactor Dimensionality Reduction

Figure 8 demonstrates the process for the MDR algorithm. Before the MDR analysis begins, the data set is divided into multiple partitions for cross-validation. Cross-validation is an important part of the MDR method, as it aims to find a model that not only fits the given data, but can also predict on future, unseen data (Ritchie et al. 2006).

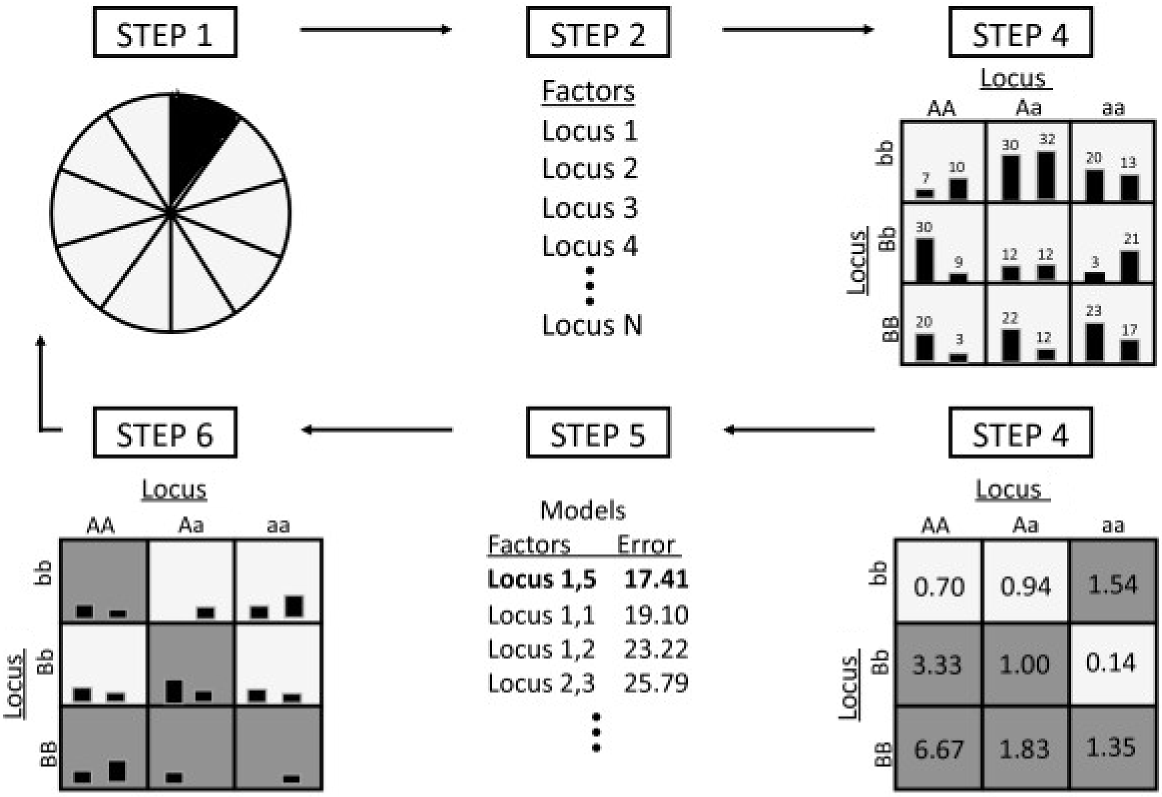

Figure 8 **-** Steps of the MDR (Motsinger-Reif, 2008)

A summary of the general steps needed to implement the MDR method detailed in (Ritchie et al. 2006) are as follows:

1. In step one, the data is divided into k (typically, 10) random subsets. (k-1) of the subsets make up the “training set” while the last subset becomes the “testing set” (see also step 6, “cross-validation”).
2. In step two, a set of f factors is then selected from the pool of all factors. These factors can include both genetic and environmental data. There is no predefined limit on the number of independent variables that can be examined. However, limits due to computation time may arise, especially when the number of potential factors is high. For example, in genetic interactions mapping problems, if m markers are used, the number of possible configurations that could be tested is of the order of $m^f$. Since m could easily be from several thousands to several millions in today applications, the search space can become intractable for values of f larger than 2 or 3. In Figure 8, f is equal to 2.
3. In step three, the f factors and their possible multifactor cells are represented in f-dimensional space, with all possible multifactorial combinations represented as cells in the table. The number of cases and controls for each locus combination are counted.
4. In step four, each multifactor cell in the n-dimensional space is labelled as high risk if the ratio of affected individuals to unaffected individuals exceeds a threshold of one (dark grey background cells), and low risk if the threshold is not exceeded (light grey background cells).

5. In steps five and six, the classification performances are estimated using the “testing set” data for each of the tested set of f factors.
6. The six steps are repeated using the k possible partitions of the original dataset into “training” and “test” sets.
7. The model with the best average performances is selected and the prediction error of the model is estimated using the independent test data.

Commonly, the classification performances are assessed using a ”balanced accuracy” criterion where the balanced accuracy is computed as a simple average of the sensibility and the sensitivity of the classifier. Repeating this procedure over all possible markers sets allows obtaining the best model, which is defined as the set of markers providing the best allocation performances. Significance for the optimal model can be obtained through a permutations test, in which the potential links between the individuals’ genotypes and the phenotypes are disrupted by randomly shuffling the phenotypes. The p-values obtained using this test have then to be corrected for multiple testing, where multiple tests are due to the number of models that are successively tested.

### 2.3. 2. Application of Multifactor Dimensionality Reduction to the detection of gene-gene interactions

A lot of applications use the principles of MDR, only a few of them will be mentioned below.

In (Calle et al. 2008), the authors have proposed a novel approach of MDR named “Model-Based Multifactor Dimensionality Reduction (MB-MDR)”. MB-MDR aims at identifying specific multi-locus genotypes associated with a disease susceptibility while allowing to adjust for marginal effects and confounders. Another difference between MB-MDR and MDR is that just those cells exhibiting significant evidence of (high or low) risk will be merged. The other cells which either show no evidence of association or have no sufficient sample size are included in an additional category, that of no evidence of risk. The results show that MB-MDR has improved power over MDR in the presence of genetic heterogeneity.

In (Cattaert et al. 2010), the authors have proposed an approach named “FAMily Multifactor Dimensionality Reduction (FAM-MDR)” for detecting gene-gene interactions. This method combines features from the genome-wide rapid association using mixed model and regression approach (GRAMMAR) (Aulchenko et al. 2007) with the approach (MB-MDR). The applications of this approach are on continuous traits, however it can be used for any type of binary traits. The result shows that FAM-MDR has improved power over the approach Pedigree-based Generalized MDR (PGMDR) in most of the simulations using continuous traits.

In (Yang et al. 2013), the authors have also proposed balancing functions for adjusting the ratio in risk classes and classification errors for Imbalanced cases and controls using multifactor dimensionality reduction (MDR-ER) as a novel method to improve MDR. The difference between MDR-ER and

MDR is that the former detects the high or low risk cells through a comparison of the percentages of cases in case data and of controls in control data rather than using the raw numbers. The authors concluded that MDR-ER can be useful for detecting gene-gene interactions in imbalanced data sets.

In (Gui et al. 2011), the authors have developed still another approach called "Robust Multifactor Dimensionality Reduction (RMDR)" for detecting gene-gene interactions. The main difference between RMDR and MDR is the use of Fisher's Exact Test by RMDR to identify whether sets of genotype combinations must be listed in MDR model. The authors concluded that RMDR is more robust than MDR in both simulated and real data.

### 2.3. 3. Strengths of Multifactor Dimensionality Reduction

(i) The main strength of the MDR is that it facilitates the detection and characterization of multiple genetic loci associated with a trait by reducing the dimensionality of the multilocus data (Ritchie et al. 2001) (Fang et al. 2012).

(ii) MDR is also non-parametric, since no parameters are estimated, which eliminates the uncertainty introduced by the parameter estimates of parametric methods, such as logistic regression (Ritchie et al. 2006).

(iii) Theoretically allow to highlight gene–gene interactions of any order (Mahachie John et al. 2011).

### 2.3. 4. Weaknesses of Multifactor Dimensionality Reduction

(i) MDR can be computationally intensive, especially when more than 10 polymorphisms need to be evaluated (Ritchie et al. 2001).

(ii) The MDR method can fail in finding the correct models, because it assumes that there is no genetic heterogeneity, as in situations where a group of cases are explained by a combination of loci different from the one that explains another group of cases (Manuguerra et al. 2007).

(iii) Some important interactions could be missed (Cattaert et al. 2011).

(iv) Lack of adjustment for main effects (Calle et al. 2008).

(v) Lack of adjustment for confounding factors (He et al. 2009).

(vi) Low power under genotyping error, missing data, phenocopy and genetic heterogeneity (Calle et al. 2008).

### 2.4. Boosting

Boosting is a machine learning ensemble meta-algorithm for reducing bias in supervised learning. It has been introduced by (Schapire 1990). Boosting is a technique for combining multiple base classifiers whose combined performance is significantly better than that of any of the base classifiers by combining many weak classifiers - i. e. classifiers that perform poorly, barely better than by chance only - to produce a powerful committee (Timo et al. 2011). It is mainly utilized in high-dimensional data (Binder et al. 2008).

### 2.4. 1. Methodology of Boosting

Theoretically, any method performing an aggregation of weak classifiers to output a strong classifier could become a boosting algorithm. Many proposals have been made in this spirit. In this text, we will use one of the most famous developed algorithm, named AdaBoost, to explain the methodology. AdaBoost is the first algorithm that could adapt to the weak learners. It iteratively generates a robust final hypothesis by giving increased weight to misclassified training samples from previous learning rounds (Liu et al. 2003). The way this is done is briefly provided, using a derivation from (Rojas 2009):

1. Assume we have a dataset with N (sets of) features and the corresponding categories ( . We have also a set of L (weak) classifiers such that .
2. We aim at building a composite (strong) classifier based on a linear combination of the weak classifiers. After (m-1) updates, the current composite classifier C is therefore: , where the constants , , . . . , are the weights we assign to each classifier
3. In the next iteration, we want to progress from to , with: . We thus need to choose the next classifier and the corresponding weight.
4. It is easy to show that the classifier with the lowest weighted errors ( , where the sum extends over the misclassified points and where ) is optimal (i.e. makes the exponential loss function minimal). So this classifier becomes $k_m$.
5. It is also shown that the corresponding optimal coefficient can be computed as:

   where is the ratio of the weights of the misclassified points to the weights of all the points:

### 2.4. 2. Application of Boosting to the detection of gene-gene interactions

In (Wan et al. 2009), the authors have proposed an alternative learning approach (MegaSNPHunter) to detect disease predisposition SNPs and high level interactions in genome wide association studies. MegaSNPHunter uses a hierarchical learning approach to discover multi-SNP interactions. It takes case-control (potentially genome-wide) genotype data as input and produces a ranked list of multi-SNP interactions as output. More precisely, the whole genome is first partitioned into multiple short subgenomes where each subgenome covers a genomic area representing possible haplotype effects in practice. For each subgenome, MegaSNPHunter builds a boosting tree classifier where nodes represent markers and paths represent interactions, and measures the importance of SNPs on the basis of their contributions to the classifier. The method keeps the SNPs from all subgenomes that reach a predefined threshold and lets them compete with each other in the same way at the next level. The competition terminates when the number of selected SNPs is less than the chosen size for subgenomes. At the last step, MegaSNPHunter extracts and reports the valuable multi-SNP interactions. To handle the multiple test issue, an extra permutation-based test at the chromosome level on both single SNP and SNP interactions is used to correct P values. The results in the paper show that MegaSNPHunter is useful in handling large-scale SNP data and performs better on both simulated data and real data than another approach (BEAM) used for comparison in this study.

In (Li et al. 2011), the authors have also used an ensemble method based on boosting for detecting epistasis. The idea in this study based on the AdaBoost algorithm was to combine an intuitive importance score with Gini impurity (see the definition in the paragraph on decision trees) to select genotype data. The results on both simulated data and real data show that the proposed approach is valid and more powerful than other approaches (RFs and BEAM) used for comparison in the study.

In (Pashova et al. 2013), the authors have proposed a novel approach of boosting based on another loss function. This method, referred to as L2 boosting has advantages in high dimensional problems and can potentially detect small effects due to combinations of environmental variables with a genotype on a phenotype. L2 boosting is based on a functional gradient descent algorithm with the L2 (squared error) loss function. The study concluded that L2 boosting is particularly useful to pick out ensembles of weaker effects of SNP that interact with another phenotype than other methods such as AIC and BIC on both simulated data and real data.

### 2.4. 3. Strengths of Boosting

(i) Very simple to implement (Wang et al. 2010).

(ii) Feature selection on very large sets of features (Lubke et al. 2013).

(iii) It has a learning strategy allowing to extract both local SNP interactions and global gene interactions in an efficient manner without exhaustive enumeration (Wan et al. 2009).

### 2.4. 4. Weaknesses of Boosting

(i) Suboptimal solution (Ganatra et al. 2010).

(ii) Can overfit in presence of noise and outliers (Sariyar et al. 2014).

(iii) In applications such as MegaSNPHunter, boosting requires that the marginal effects of SNP be above the median of the marginal effects of the subgenomes they reside on (Wan et al. 2009).

### 2.5. Decision Tree

A Decision Tree (DT) is one of the most often used non-parametric method to solve a classification and regression problem. The decision tree can be used in high-dimensional data to facilitate the decision by choosing the most appropriate one to reach a goal (Motsinger-Reif et al. 2010). Various algorithms targeting an optimal tree (defined as a tree that accounts for most of the data, while minimizing the number of levels) have been devised, including the classical ID3 (Quinlan 1986) and CART (Breiman et al. 1984).

### 2.5. 1. Methodology of decision trees

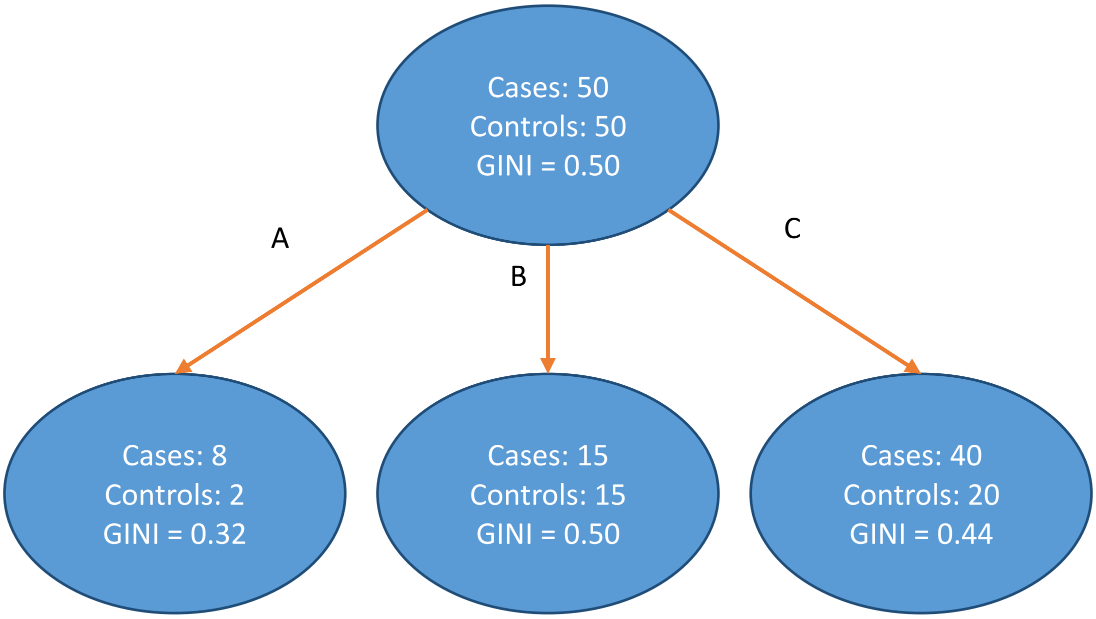

Figure 9 **-** A first step in the construction of a decision tree on a case-control problem. The data in the top node are split into child nodes according to each of the possible values A, B and C of a given attribute. An index (GINI here, see text) is computed to evaluate the “purity” of each node (lower values of GINI means “higher purity”).

The dataset is made of a list of N items, represented individually by a vector of attributes $\mathbf{x_i}$ and a class $y_i$ (in Figure 9, $y_i$ is either “Case” or “Control”). This dataset is firstly divided into a “training set”, used to build the tree, and a “test set”, used to prune the tree to avoid overfitting. Decision trees are built in a top-down manner (Li et al. 2011). The top node gathers the whole training set data. A scan of the various attributes of the node items allows obtaining the attribute leading to the “purest” children nodes when the data is split according to the attribute values. For example, in Figure 9, the tested attribute takes 3 values A, B or C, leading to the 3 child nodes gathering 10, 30 and 60 items from the 100 items present in the parent node. In this context, “purest” means “showing the lowest variability for the classification variable”. Various measures can be used to measure the (im) purity, including entropy measures and Gini coefficient. Figure 9 shows the value of the Gini coefficients computed in each node as:

—— —— (1)

where is the proportion of cases in the node, is the proportion of controls, is the number of cases in the node, is the number of controls in the node and is the total number of individuals in the node. This coefficient is a measure of impurity since pure nodes (only one class present) would lead to GI = 0, while a completely mixed node, with half of the individuals being cases and the other half being controls would lead to the maximal value GI=0.5. Consequently, a possible criterion to choose the “best” attribute (i. e. the one leading to the “purest” child nodes) could be to select the attribute leading to the largest reduction in GI, computed as:

—— (2)

Where, n(d) is the number of individuals at a child node d, N is the number of individuals at the parent node p, GI(d) is the Gini impurity of node d, and x is the tested attribute.

In the example in Figure 9, the reduction in GI can be calculated:

$$\text{Gain} = (0.5 - 0.1 * 0.32 - 0.3 * 0.5 - 0.6 * 0.44) = 0.054.$$

The decision tree can be grown using recursively the procedure described above, until a stop criterion is met. Alternatively, the tree can be fully grown to obtain exclusively pure nodes, and subsequently pruned to avoid overfitting. Various techniques and algorithms exist to perform the pruning.

### 2.5. 2. Application of decision trees for the detection of gene-gene interactions

In (Estrada-Gil et al. 2007), the authors have proposed a novel approach named “Genetic Programming based on a Decision Tree (GPDTI)” for the detection of gene-gene interactions. The strategy used in this method to complete the tree is to use a cross-validation strategy and classification error to avoid overfitting and take the best estimates. The authors concluded that GPDTI can be useful for detecting gene-gene interactions in big datasets and leads to an easy interpretation of the results.

In (Motsinger-Reif et al. 2010), the authors have developed a novel approach using decision trees, named “Grammatical Evolution Decision Trees (GEDT)”, for the detection of genetic interactions. A difference between (GEDT) and traditional decision trees is the genome representation, which is split by (GEDT) into logic expressions. The results on simulated data show that the proposed novel approach (GEDT) is useful and more powerful than a traditional decision tree approach in detecting gene-gene interactions with and without main effects.

### 2.5. 3. Strengths of Decision Tree

(i) Decision trees can handle large quantities of data in reasonable computation time (Yoo et al. 2012).

(ii) Decision trees inherently include an interaction between feature subset search and a classification model (Huang et al. 2009).

(iii) Decision trees are simple and the resulting tree can be interpreted as a series of IF-THEN rules that are easy to understand (Moore et al. 2010).

### 2.5.4. Weaknesses of Decision Tree

(i) Decision trees may not be able to discover particular important interactions because of limitations imposed by the stopping rules, the competitive importance of the variables and/or the pruning procedure (Barnholtz-Sloan et al. 2011).

(ii) Possibility of duplication with the same sub-tree on different paths (Guy et al. 2012).

(iii) Decision trees rely on modest marginal effects to construct data learning (Guy et al. 2012).

## 2. 6. Random Forest

Random Forests (RF) or random decision forests are a non-parametric method, based on Decision Trees (DT). Random Forests are also an ensemble learning method grouping Decision Trees (Winham et al. 2012). Since fully-grown DT tend to over-fit the data and provide variable results in the deep nodes, an idea would be to obtain a reduction of this variance by averaging several trees created on independent training sets. This idea is at the root of RF.

### 2. 6. 1. Methodology of Random Forest

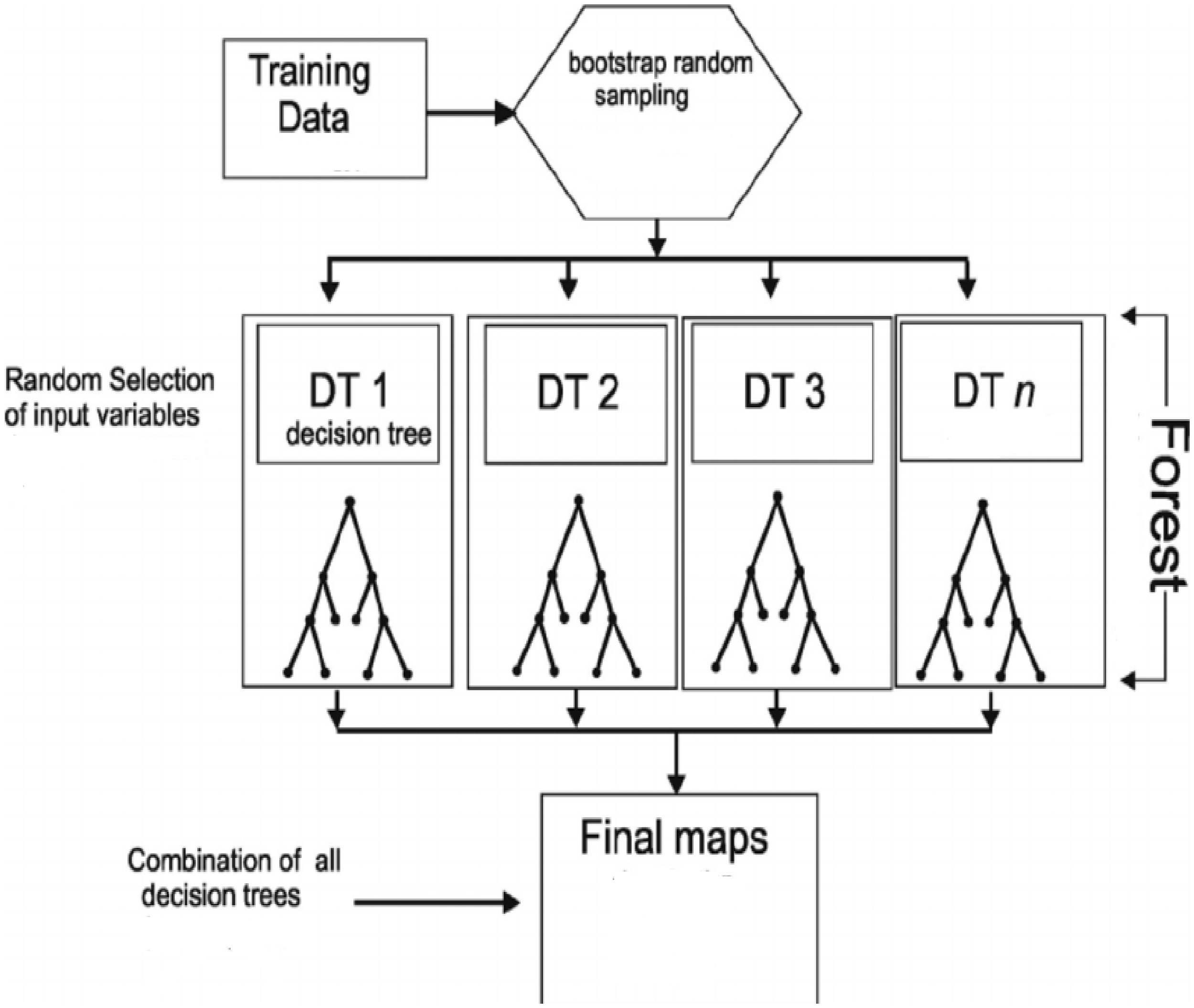
Training Data
bootstrap random sampling
Random Selection of input variables
DT 1 decision tree
DT 2
DT 3
DT *n*
Forest
Combination of all decision trees
Final maps

Figure 10 - Random Forest

A RF is a collection of individual decision tree classifiers, where each tree in the forest has been trained using a bootstrap sample of instances (i.e. subjects) from the data, and each attribute in the tree is chosen from among a random subset of attributes (Breiman 2001). Classification of instances is based upon aggregate voting over all trees in the forest (Moore et al. 2010). The methodology of Random Forest can be described as follows (Koo et al. 2013):

1- Grow many trees using bootstrap samples from the training data.

2- The data from the training data and not present in the bootstrap sample is named OOB (“out-of-bag”) data and can used to estimate the prediction error.

3- Each node in the tree is split by taking the best among a randomly chosen subset of predictors at the node.

4- Use OOB error rate from each tree to measure the prediction error and to get an unbiased measure of the accuracy of the model over test data.

### 2. 6. 2. Application of Random Forest for detecting gene-gene interactions

In (Lunetta et al. 2004), the authors have proposed an approach of Random Forest to detect gene-gene interactions associated to a complex disease using SNPs genotypes. The method uses variable importance to determine the sets of SNPs that are important in the prediction. The results show that this approach is useful in selecting a set of SNPs associated with an increased risk from a large number of SNPs. The authors also show that their method has advantages over other methods (logistic regression, Fisher Exact) used for comparison in the study.

In (Jiang et al. 2009), the authors have developed a Random Forest method to reduce the number of combinations of genetic variants into a small number, which can be controlled in the search for the interactions. The idea behind this method is to use a sliding window sequential forward feature selection (SWSFS) algorithm to take the sets of SNPs able to minimize the classification error. The results in the paper show that this approach is useful in handling large-scale SNP data and identify the epistatic interactions on both simulated data and real data, and performs better than other approaches (BEAM, logistic regression, Chi-squared test) used in this study.

### 2. 6. 3. Strengths of Random Forest

(i) RF detect gene-gene and gene-environment interactions without a strong marginal effect (Winham et al. 2012, Botta et al. 2014).

(ii) RF reduces the overfitting data, and are therefore more accurate (Jiang et al. 2009).

### 2. 6. 4. Weaknesses of Random Forest

(i) Bootstrapping and random feature selection make RF look more like a black box (Winham et al. 2012).

(ii) RF difficult to scale up to GWAS data (Upstill-Goddard et al. 2012) .

## 3. Bayesian methods

Although inherently parametric, Bayesian methods are important in the field of genetic interaction mapping and have been used for comparison and as candidates in some of our procedures. Consequently, we will say a few words on these methods.

Bayesian statistics is at the root of many machine learning methods; based on the famous Bayes theorem due to Thomas Bayes (1701-1761), the Bayesian methods combine prior information to sample information from the data to obtain a so-called posterior distribution from which inference and conclusions can be drawn. Bayesian methods provide a big flexibility to treat and analyze any type of problem, partly because they can be used cumulatively across progressive experiments (Sebastiani et al. 2003) and also because they lend themselves to the use of nowadays efficient computational tools and algorithms.

## 3. 1. Methodology of Bayesian methods

The general idea in Bayesian methods is that the data (noted **x**) we collect transforms the prior information we have on a set of parameters (noted θ) into a posterior information integrating both (prior and collected) types of information. This transformation takes a rough and nice form, usually referred to as Bayes theorem:

In this expression, is the 'prior distribution' of the parameters vector, representing the information we have (or we assume) before starting the experiment, is the conditional distribution of the observations, given a set of parameters, most often called the 'likelihood', is the 'posterior distribution' of the parameters vector, from which we would like to draw inferences on the parameters values, and means that both sides are equal up to a constant.

The posterior distributions will include all the current information about the parameters and can be updated as new information accumulates. The main objective of Bayesian inference is to explore the full posterior distributions of all the parameters. Even in the situations where the posterior distributions cannot be identified as a known distribution, inference using that distribution can be carried out using computational algorithms such Markov chain Monte Carlo (MCMC) algorithms (Liu 2013).

In short, typical Bayesian methods can be described in these steps (Glickman et al. 2007):

1- Select a suitable probability model for the data.
2- Select a prior distribution, which can be selected from past information, or chosen to present a balance among outcomes when no information is available.
3- Create the likelihood function, which is depended on the data and the probability model.
4- Construct the posterior distribution, which is combining the likelihood with the prior distribution.
5- Extract statistical results based on the posterior distribution.

## 3. 2. Application of Bayesian methods in the detection of gene-gene interactions

Various methods have been devised using Bayesian approaches. We will restrict ourselves to a more detailed description of only one recent application to show an application using the Bayesian framework. One reason for this choice is that the chosen method is representative of the proposed approach. The other reason is that we have been using that method in our work (for comparison or as part of an aggregation method).

In (Wang et al. 2015), the authors have proposed a novel approach named “Bayesian High-order Interaction Toolkit (BHIT)” for detecting high-order gene-gene interactions on both discrete and continuous phenotypes.

In this study, the data is made of a set of phenotypes **Y** measured on G individuals and, for each of these individuals, a set of R genotypes **X**. The goal of the algorithm is to associate subsets of **X** to distinctive patterns of the phenotypes. To that end, they define a vector **I**, of length R, where $I_j$ should provide the subset number for marker j, which is number between 1 and H. Consequently, solving the problem corresponds to finding **I** and H, which are therefore the parameters in this Bayesian problem: we want to infer .

Using the Bayes theorem, this distribution can be inferred as:

Now, , the likelihood can be further detailed into:

The first factor is the probability of the phenotypes given a partition of the genotypes; it can be modelled using (for example) multivariate Gaussian distributions, introducing the need for distinct means, standard deviations and correlations as new parameters for these normal distributions. The second factor is the probability of the genotypic configuration given a partition of the genotypes. This can be modelled as a multinomial distribution, with unknown probabilities. Therefore, these probabilities used in the multinomial distributions act as new parameters, which are supposed to arise from a Dirichlet distribution (see the paper for details).

Sampling from the joint posterior distribution can be done using a Metropolis-Hastings algorithm: after burn-in iterations, new configurations (**I**) are sampled, and the joint posterior probability of this new configuration $P_N$ is compared to the probability of the previous configuration $P_P$: the new configuration is accepted if $P_N \geq P_P$. If $P_N < P_P$, the new configuration is accepted with a probability equal to $P_N / P_P$.

The results on both simulated data and real data demonstrate that the performances are better than the other approaches (BEAM and BEAM2, also using a Markov-Chain Monte-Carlo partitioning strategy) used for comparison in this study.

Researchers have constructed other Bayesian strategies. For example, (Yi et al. 2011) have used Bayesian models to detect main effects of locus, gene-gene and gene-environment interactions in case-control studies. Their approach uses Bayesian generalized linear models with Student-t prior distributions in different situations. The results obtained on simulated data and real data show that the approach is potentially useful to detecting gene networks underlying complex diseases.

Another example is (Liu, et al., 2015). The authors have used a Bayesian hierarchical mixture model for detecting gene-gene and gene-environment interactions in the same model. The results on simulation data and real data demonstrate that this method is useful to detect gene-gene and gene-environment interactions in various situations and has more power than the logistic regression used for comparison in this study.

## 3. 3. Strengths of the Bayesian methods

(i) They can efficiently explore high-order interactions (Wang et al. 2015).

(ii) They can accurately deal with missing data and genotyping errors (York et al. 2005).

(iii) They can be used as variable selection methods (Isci et al. 2014).

## 3. 4. Weaknesses of the Bayesian methods

(i) Uninformative prior probabilities may be difficult to specify (Yang et al. 2012).

(ii) They require high computational resources for big dataset (Wang et al. 2015).

(iii) They use stochastic way and are therefore not guaranteed to find the optimal solution (Chris et al. 2015). Actually, they can easily get stuck in a reduced portion of the parameters space.

# Objectives

The main objective of this thesis was to develop statistical non-parametric methods allowing to detect genetic interactions. This is an important part of the job aiming at a better understanding of the relationship between genomic configuration and phenotypic expression.

As parts of this main objective, we have several more specific objectives:

- Understanding the mechanisms of nonparametric statistical methods and working on the development of these methods.
- Entering the world of ‘big data’ and discovering tools helping to handle, manage and analyze large data sets such as the ones met in genome-wide association studies.
- Understanding mechanisms of genetic interactions and finding methods and algorithms to discover them.

# Experimental Section

# Study 1: KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies

# Preamble

In this study, we propose the main contribution of this PhD thesis: a novel approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) methods for detecting gene-gene interactions as a possible alternative to existing algorithms, especially in situations where the number of involved determinants is high. This method illustrates how taking into account the physical nature of the problem - the markers present on today dense maps are physically linked on a chromosome, introducing a disequilibrium between close markers due to linkage - allows introducing more information in existing methods, and how this can be used to improve these methods. In our case, we have demonstrated that KNN-MDR is more computationally efficient than other exhaustive strategies, using windows of linked markers instead of single markers, which is facilitating the analysis of large-scale data sets with potentially genome-wide SNPs. The improvements on the efficiency of the method make it eligible for the detection of higher-order interactions, although this would admittedly remain a notably challenging task. Another reason making KNN-MDR useful is its ability to detect interactions in the absence of marginal effects. Several methods use marginal effects to pre-filter the data, assuming that only markers showing some effect individually are likely to be involved in interactions. We have considered this as an excessive assumption, and consequently developed a strategy where this assumption is not necessary. Relaxing this assumption in our simulations has proved that KNN-MDR performed generally better than concurrent methods in this context. Although we have demonstrated some of the advantages of the method, we are aware that improvements are possible, and some ideas in that direction are proposed in the perspectives. These perspectives could render the method more useable for external users and increase its use.

# Study 1: KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies

*BMC Bioinformatics* 18:184 (2017)

Sinan Abo Alchamlat and Frédéric Farnir

# *Abstract*

## Background

Finding epistatic interactions in large association studies like genome-wide association studies (GWAS) with the nowadays-available large volume of genomic data is a challenging and largely unsolved issue. Few previous studies could handle genome-wide data due to the intractable difficulties met in searching a combinatorial explosive search space and statistically evaluating epistatic interactions given a limited number of samples. Our work is a contribution to this field. We propose a novel approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) methods for detecting gene-gene interactions as a possible alternative to existing algorithms, especially in situations where the number of involved determinants is high. After describing the approach, a comparison of our method (KNN-MDR) to a set of the other most performing methods (i.e. MDR, BOOST, BHIT, MegaSNPHunter and AntEpiSeeker) is carried on to detect interactions using simulated data as well as real genome-wide data.

## Results

Experimental results on both simulated data and real genome-wide data show that KNN-MDR has interesting properties in terms of accuracy and power, and that, in many cases, it significantly outperforms its recent competitors.

## Conclusions

The presented methodology (KNN-MDR) is valuable in the context of loci and interactions mapping and can be seen as an interesting addition to the arsenal used in complex traits analyses.

## Keywords

Gene-gene interaction - Epistasis - Single Nucleotide Polymorphism - Genome-wide association study - Multifactor Dimensionality Reduction - K-nearest neighbors.

## Background

These last years have seen the emergence of a wealth of biological information. Technical improvements in genotyping and sequencing technologies have facilitated the access to the genome sequence and to massive data on genes expression and on proteins. This large availability of molecular information has revolutionized the research in many fields of biology. In parallel to these technical developments, methodological advances are needed to address the various questions of scientific interest that have been targeted when developing these new molecular tools. For example, the identification of up to several millions genomic variations in many species and the development of chips allowing for an effective genotyping of SNPs panels in large cohorts have triggered the need for statistical models able to associate genotypes from individuals and interacting SNPs to phenotypic traits such as diseases, physiological and productions traits (Wu et al. 2010). Our paper is a contribution to this association problem.

The systematic exploration of the universe of variants spanning the entire genome through genome-wide association studies (GWAS) has already allowed the identification of hundreds of genetic variants associated to complex diseases and traits, and provided valuable information into their genetic architecture (Wu et al. 2010) while allowing to improve prediction of phenotypic outcomes (Wei et al. 2014). Nevertheless, most variants identified so far have been found to confer relatively small information about the relationship between changes at the genomic locations and phenotypes because of the lack of reproducibility of many of these findings, or because the identified variants most of the time explain only a small proportion of the underlying genetic variation (Fang et al. 2012). This observation, quoted as the ‘missing heritability’ problem (Manolio et al. 2009) of course raises the following question: where does the unexplained genetic variation come from? Several authors have postulated that many genes and mutations could be involved, with individual small effects, resulting into a low detection power in most of the performed studies, but with large collective effects (Visscher et al. 2012). Another tentative explanation is that genes do not work in isolation, leading to the idea that sets of genes (“gene networks”) could have a major effect on the tested traits while almost no marginal effect is detectable at individual locus level. Note also that this gene network hypothesis is a potentially credible explanation to the lack of reproducibility of obtained positive results (Boos et al. 2011), due to situations where different mutations or mutations combinations within the network (within the same genes or on different genes in the networks) could lead to similar phenotypic effects (Manceau et al. 2010).

Consequently, an important question still remains about the exact relationship between the genomic configuration, including the interactions between the involved genes, and the phenotypic expression. The major idea in this respect is to try to associate observed variations at the macroscopic level (phenotype) to identified variations and their interactions at the molecular level.

This view introduces at least two challenges. First, the genetic mechanisms underlying most traits of interest are complex and probably involve most of the time many genes and many interactions between these genes, leading to a complex relationship between genomic variants and phenotypes. So, modeling and identification of every, and even of any, interaction is a potentially very challenging task (Marchini et al. 2005). Second, from a more statistical point of view, fully modeling the interactions leads to models with large number of parameters to be estimated and large search space, leading to the well-known ‘curse of dimensionality’ problem (De los Campos et al. 2010). Furthermore, increasing the number of parameters to be estimated potentially makes the power issues mentioned above even more critical. Nevertheless, introducing interactions into the model might lead to a more accurate model of the underlying genetics, which in turn might improve the detection power of effects of interest. So it is not obvious that interaction models will present poor power when compared to non-interaction ones, which should motivate more research on the subject.

In the literature, various statistical methods have been used to detect gene-gene or gene-environment interactions (Koo et al. 2013), (Millstein 2013). Many of these statistical methods are parametric and rely on large samples properties (Park et al. 2008), (Usai et al. 2012). On the other hand, nonparametric methods have generated intense interest because of their capacity to handle high-dimensional data (Musani et al. 2007). In order to limit the size of the search space, many of the proposed approaches may have missed potential interactions by only considering variants that have a significant genetic marginal effect as, for example, in the logistic regression method proposed by (Fang et al. 2012), where the model relates one or more independent variables (i.e. main effects for genes) and their corresponding interaction terms (i.e. gene-gene interaction effects) to a discrete dependent variable (e.g. disease status). Because of issues linked to the dimensionality, models such as the logistic regression are limited in their ability to deal with interactions involving many factors (Ritchie et al. 2003). In response to these limitations, novel methods for detecting interacting variants have been designed, such as neural networks (Gunther et al. 2009), random jungles (Schwarz et al. 2011), random forests (Winham et al. 2012), BOOST ‘‘BOolean Operation-based Screening and Testing’’ (Wan et al. 2010), support vector machine (Ban et al. 2010), MegaSNPHunter (Wan et al. 2009), AntEpiSeeker (Wang et al. 2010) or odds ratio (Wu et al. 2010).

One of the most successfully used family of methods in the gene-interactions problems is multifactor dimensionality reduction (MDR) (Ritchie et al. 2003). The MDR method is nonparametric (i.e., makes no hypothesis about the distribution of the statistical parameters), model-free (i.e., it assumes no particular inheritance model), and directly applicable to case-control and discordant-sib-pair studies (Ritchie et al. 2001). The main idea in MDR is to reduce the dimensionality of multi-locus data to improve the ability to detect genetic combinations that confer disease risk (Ritchie et al. 2001). MDR has been proposed to identify gene–gene or gene-environment interactions when marker and/or environment information is available (Ritchie et al. 2006). An advantage of the MDR methods is, as pointed out in (Mahachie John et al. 2011), that, due to their nature, they theoretically allow to highlight gene–gene interactions of any order (Manuguerra et al. 2007).

Refinements of the method have been proposed to deal with potential limitations. (Cattaert et al. 2010) has proposed a novel multifactor dimensionality reduction method for epistasis detection in small or extended pedigrees, FAM-MDR. (Cattaert et al. 2011) and (Calle et al. 2008) have also developed Model-Based Multifactor Dimensionality Reduction (MB-MDR), a MDR-based technique that is able to unify the best of both nonparametric and parametric worlds, allowing to include corrections for cofactors, as in parametric models, while using the flexible framework of non-parametric MDR analyses. Another extension is Generalized MDR (GMDR), a version of the MDR method that permits adjustment for discrete and quantitative covariates and is applicable to both dichotomous and continuous phenotypes (Lou et al. 2007).

Although applied to numerous genetic studies (Collins et al. 2013), (Gui et al. 2011), MDR faces important challenges. First, MDR can be computationally intensive, especially when a large number of markers needs to be tested (Ritchie et al. 2001). Second, the interpretation of MDR results is difficult, for example in situations where a strong marginal effect makes the effects of the other polymorphisms in the interaction questionable (Calle et al. 2008). Third, the MDR method can fail in finding the correct models, because it assumes that there is no genetic heterogeneity, as in situations where a group of cases are explained by a combination of loci different from the one that explains another group of cases (Cattaert et al. 2011). Lastly, the number of possible combinations explodes exponentially with the number of interacting factors, which makes the approach impractical in terms of needed cohorts sizes and computing time in situations where large numbers of genetic and/or environmental determinants are involved, another instance of the ‘curse of dimensionality’ problem.

In this paper, we propose a novel MDR approach using K-Nearest Neighbors (KNN) methodology (KNN-MDR) for detecting gene-gene interaction as a possible alternative to current MDR methods in situations where the number of involved determinants is potentially high and the number of tested markers is large. After explaining the rationale of our method, we will provide results on the comparison of KNN-MDR to a set of competitor methods on both simulated and real datasets.

## Methods

### KNN method

KNN stands for “K Nearest Neighbors” and is one of the most popular algorithms for pattern recognition and classification. Roughly, classification of an observation can be made using a majority vote within the K nearest neighbors of the observation (Aci et al. 2010), where the neighborhood is based on a defined distance between observations. Although simple, many researchers have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets (Suguna et al. 2010). Also, KNN is a multivariate method that retains the variable relationships seen in the data because the logical relationships among response variables will be maintained (Ver Hoef et al. 2013), a feature of importance in our genetic context. The flexibility of KNN is also a great advantage and this technique helps to alleviate the curse of dimensionality by shrinking the unimportant dimensions of the feature space, bringing more relevant neighbors close to the target point (Aci et al. 2010).

## MDR method

The method will be described for dichotomous traits for the sake of simplicity, but could be extended to other situations using the approach described for GMDR (Lou et al. 2007). The Multi-Dimensional Reduction (MDR) method is designed to replace large dimension problems with reduced dimension ones, allowing to make inferences based on a smaller set of variables. In the context of genomic studies, the idea in (Ritchie et al. 2001) is to replace the high dimensional problem arising from considering several markers simultaneously, with one unique variable (for example, a status) that can take only 2 values (for example, '*high risk*' or '*low risk*'). To illustrate, if a set of N SNP markers is used in a case-control study to define the multi-locus genotype, $3^N$ genotypes are possible. Each of these genotypes can be mapped to a status with only 2 values (case or control) using a majority vote on the statuses of the training set individuals falling into that genotype. The classification performances of any set of markers used to define the genotypes can then be assessed, typically using a cross-validation procedure, where the performance is estimated on a test set for each partition trough a measure involving sensitivity and/or sensibility of the classifier, and averaged over all partitions. For all computations reported in this paper, we have used a 10-fold cross-validation procedure and assessed the performances using 'balanced accuracy', which is a simple average of the sensibility and the sensitivity of the classifier. Repeating this procedure over all possible markers sets allows obtaining the best model, which is defined as the set of markers providing the best allocation performances. In practical situations, the potential number of tested markers sets might be huge: if an exhaustive search is to be performed on all P-markers interactions in a GWAS with M markers, about $M!/[P!*(M-P)!] \sim M^P/P!$ combinations would need to be checked, a huge number with nowadays available markers panels. Significance for the optimal model can be obtained through a permutations test, in which the potential links between the individuals' genotypes and the phenotypes are disrupted by randomly shuffling the phenotypes. The p-values obtained using this test have then to be corrected for multiple testing, where multiple tests are due to the number of models that are successively tested.

## KNN-MDR method

Although a widely used and well-established technique, MDR faces several problems, as detailed above. The computational load described in the previous section remains a major issue. Although recent publications (Lishout et al. 2015) have provided some tools to achieve low order interactions screening in a GWAS, the task will remain very challenging for larger order interactions and for larger markers sets, such as sequencing data, and alternative approaches reducing the computer burden remain desirable. Another problem linked to the MDR methodology arises when a test set individual's multi-locus genotype has not been observed in the training set, making it impossible to classify the newcomer. Furthermore, in situations where very few training individuals share the same multi-locus genotype as the tested one, the accuracy of the assignment can also be questioned. Since the number of multi-locus genotypes explodes exponentially when the number of markers in the markers sets increases, this problem becomes rapidly critical, and could finally render the approach inaccurate (few individuals are used to classify) or even unusable (no individual useable to classify) in situations where more than 3-4 markers are to be used simultaneously and with classical cohorts' sizes. Another consequence of the limited number of markers that can be considered simultaneously in MDR is that the genomic regions involved in interactions will most of the time be represented through a single marker, although, due to linkage disequilibrium, considering several linked markers might increase the association signal intensity, and consequently improve the detection power.

Our proposal is therefore to slightly modify MDR to allow facing some of the shortcomings of the method. The only modification is in the status allocation procedure: while MDR uses a majority vote among the (potentially scarce or empty) set of individuals sharing the same multilocus genotype as the tested individual, we propose to use a majority vote within a set of the K nearest neighbors of the tested individual. This procedure has the obvious advantage to eliminate the problem of potentially scarce or empty genotypic configurations mentioned above. On the other hand, this strategy introduces the need to define the neighborhood: a "distance" between individuals based on the genotypic configurations at the selected markers will be needed, and the size K of the neighborhood will have to be provided. These parameters of the method - the chosen distance, K - are further discussed in the discussion section. A second advantage of our approach is that more markers can be considered at once than in the classical MDR strategy. The idea, also detailed in the discussion section, is thus to replace the sets of single markers used in MDR by sets of windows spanning several markers: the M markers are split into W windows of contiguous markers, where the choice of the windows sizes and positions could use genetic criteria explained in the discussion section, and the distances used in KNN-MDR are based on these windows. All the other steps are similar to the classical MDR steps (partitioning for the cross-validation, performance and significance assessments, best model selection). Note that the number of windows W might be much smaller than the number of markers M, as explained below. Consequently, the proposed approach might greatly reduce the needed amount of computations, and consequently make higher-order interactions more affordable. Although alternatives are possible, we have used Mahalanobis distances in our analyses because of its numerous advantages in our setting (see the discussion).

Note that, in KNN-MDR, the computer burden scales quadratically with the number of individuals since the distances between pairs of individuals are needed, but is less sensitive to the number of markers since markers are pooled into windows. So, the important parameter from a computing point of view is the number of windows W, which does not necessarily increase when the number of markers increases.

## Competitor methods

After designing our method, we needed to compare the performances of our approach to some of the other proposed algorithms. Since many methods are available (Wei et al. 2014), we decided to consider four of the most popular ones to be used in the comparison, namely: MDR, BOOST, MegaSNPHunter and AntEpiSeeker. The rationale for choosing this set of methods is the following:

- AntEpiSeeker (Wang et al. 2010) and BOOST (Wan et al. 2010) have been recommended as efficient and effective methods in the comparative analysis of (Shang et al. 2011),
- MDR (Ritchie et al. 2001) is one of the most famous methodologies for detecting interactions (Wei et al. 2014),
- MegaSNPHunter (Wan et al. 2009) is targeting high level interactions, one of the potential advantage of KNN-MDR. Also, a method for exploiting large genotypes sets is provided, which is another objective of our algorithm,
- All these methods have been applied successfully to real datasets,
- These methods have different search strategies: exhaustive search (MDR, BOOST), stochastic search (MegaSNPHunter) and heuristic search (AntEpiSeeker),
- Software implementing the methods is available.

## Simulation

In order to assess the performances of the proposed method, we have simulated various situations and ran MDR, BOOST, MegaSNPHunter, AntEpiSeeker and KNN-MDR on the same datasets to compare the performances in terms of detection power and accuracy. The generation of the simulation datasets will be described in the following lines.

One of the aims of our study was to assess the performance of the methods to unravel gene-gene or gene-environment interactions in the absence of large marginal effects. The reason for that choice was that many methods are able to detect such large marginal effects and to infer interactions within a limited set of loci selected on that basis. Accordingly, we wanted to devise an approach that is able to detect interactions even in the absence of marginal effects. For that reason, efforts have been devoted to generate datasets with interacting genes in the absence of significant marginal effects. Furthermore, heterogeneity between samples has been shown to be a major source for the non-reproducibility of significant signals (Yang et al. 2009). We have modeled heterogeneity by associating penetrances to the multi-locus genotypes underlying the simulated binary trait. The data generation algorithm proceeds along the following lines:

(1) To obtain a linkage disequilibrium (LD) pattern similar to patterns that can be observed in humans, SNPs spanning the human chromosome 9 (HSA9) have been obtained from a study on Crohn disease in Caucasians (Gori et al. , Lou et al. 2007) for 197 individuals. 2000 markers with minor allele frequencies (MAF) above 0.3, and no missing genotype have been selected. Hardy-Weinberg equilibrium tests have been performed on the genotypes for these markers, and the high MAF threshold has been chosen to select informative markers among the complete list of markers, to compensate for the information loss resulting from discarding the other available markers to decrease the computational load. Nevertheless, since experimental data has been used, genotyping errors might be present. Presence of LD in the data was checked using simple association tests between consecutive markers (data not shown).

(2) Since many different individuals are needed in the simulations, we used a trick similar to (Chen et al. 2011) to generate new individuals based on the few available genotypes: each individual genotype was chopped into 10 SNP windows, leading to 200 windows with (maximum) 197 different 10 loci genotypes. Each simulated individual genotype was then built by randomly sampling a genotype for each window and concatenating the 200 genotypes into a new complete genotype with 2000 markers. This technique allows for $197^{200}$ potentially different individuals while conserving some LD.

(3) G SNP were then randomly chosen as having an effect on the simulated phenotype, where G = 2, 3, 4 or 5. Since SNP selection is random, SNP might be linked or not.

(4) Selected SNP genotypes were then used to generate the binary phenotypes. More details of the algorithm are given in an appendix (see Appendix 3 (Additional file 8)), but roughly:

   a. A penetrance is computed for each multi-locus (G SNP) genotype in such a way that each of the G SNP shows no marginal effect: $P( A \mid G_i = 0) = P( A \mid G_i = 1) = P( A \mid G_i = 2) = P$ where $G_i$ denotes the genotype for locus i (i = 1, 2, …, G), 0, 1, 2 are the number of instances of the minor allele in the SNP genotype, A means Affected, $P( A \mid G_i )$ is the penetrance for genotype $G_i$, and P is the prevalence of the disease in the sample (since we used a more or less balanced case-control design, we used a prevalence of P = 0.5).

   b. The multi-locus penetrances $MP = P( A \mid G_1 = k, G_2 = m, \ldots)$ where k, m, … = 0, 1 or 2 are obtained to meet the requirement of no marginal effect (see previous step). An algorithm to compute these penetrances is provided in the Appendix 3 (Additional file 8).

   c. The phenotypes (i.e. affected or non-affected status) are then obtained by randomly sampling a uniform distribution between 0 and 1 and comparing the obtained deviate d to the multi-locus penetrance MP: if d < (>) MP, the individual is (not) affected.

(5) One SNP out of 2 consecutive SNPs was then randomly discarded, leaving 1000 markers for the analyses. The rationale of this selection is that causative mutations might nowadays be present or not in the genotyped variants. This will also be the case in our simulations.

(6) Genotypes and corresponding phenotypes were generated for each simulation, and the obtained datasets were studied using all four methods. KNN-MDR windows size was set to 10 markers, leading to 100 non-overlapping windows, and K value was set to 10. The parameters for the other methods were chosen so that resolution was almost similar for all methods.

(7) Finally, 100 permutations of the phenotypes were performed for each simulation (unless otherwise stated) and the resulting datasets were analyzed using the four methods in order to assess significance. Although this number of permutations is too low for routine work, it was used to reduce the computing burden and help us to discriminate between results clearly non-significant (i.e. $p > 0.05$) and those potentially significant (i.e. $p < 0.05$). When a higher precision was needed for the p-values (see below), an adaptative permutations scheme was used, in which windows not reaching a pre-determined p-value threshold are progressively abandoned in the permutations scheme since these windows are very unlikely to finally reach a significant result (Purcell et al. 2007).

## Real data

Analyses using real data have also been performed. Rheumatoid arthritis (RA) genotype data on 1999 cases and 1504 controls have been obtained from WTCCC (Wellcome Trust Case Control 2007, Wang et al. 2015). Genotypes from the Affymetrix GeneChip 500K Mapping Array Set have been filtered using the usual quality controls tests on DNA quality (percentage of genotyped marker for any given individual above 90 %), markers quality (percentage of genotyped individuals for any given marker above 90 %), genotypes frequencies (markers with a p-value below a Bonferroni adjusted 5% threshold under the hypothesis of Hardy-Weinberg equilibrium in the controls cohort have been discarded). Missing genotypes for the GeneChip markers have been imputed using impute2 software (Howie et al. 2009, Lishout et al. 2015). This procedure led to 312583 SNP to be analyzed for the 2 cohorts. (Zhang et al. 2012) and (Shchetynsky et al. 2015) also used this dataset to infer potential interactions. These studies will therefore serve as a comparison for the results obtained with our approach.

## Working on large datasets

When working on large sets of markers, such as for example those commonly met in GWAS analyses, splitting the complete set into a reasonable set of windows could necessitate including large numbers of markers in each window, which would eventually swamp the signals of interest, as explained in the discussion section. An alternative is to pre-select a subset of markers (for example, taking one marker every N markers) and to define a first set of windows based on these markers. This strategy would allow windows to cover potentially large regions while preserving some detection power. After a first run of KNN-MDR using this subset, the detected combinations (i.e. those departing significantly from the distribution of the results, assuming that most combinations do not have an effect on the studied trait, and that this distribution accordingly corresponds to the distribution of the used measure under the null hypothesis) would be used for a second round of KNN-MDR runs. In this new round, the markers hidden in the first round could be partially or totally recovered for each of the identified regions, and the same approach as in the first round could be used recursively on these new regions. The sequential detection of progressively denser regions could continue down to single markers. An example of this strategy in a GWAS study is provided in the of “Results on WTCCC data” section.

## Results

### Results on simulated data

Since performing classical MDR analyses on a large number of markers is not an obvious task, especially when the number of putative involved SNPs (noted G) is 3 or more, we restricted our analyses to G = 2 and G = 3 to make comparisons to other methods feasible. We have defined the “power” as the proportion of simulations where an association signal was detected ($p < 0.05$), and the “corrected power” as the proportion of simulations where the association was detected and involved the causal SNP (i.e. a rough measure of accuracy). The comparison of the five tested methods is presented for situations where G = 2 in Table 5 and for G = 3 in Table 6 (data sets used to generate these 2 tables are provided as additional files (1 and 2) and more details on the comparisons of the methods results are provided in Appendix 2 (Additional file 7).

| Method | MDR | AntEpiSeeker | BOOST | MegaSNPHunter | KNN-MDR |
|---|---|---|---|---|---|
| power | 0.68 | 0.88 | 0.76 | 0.84 | 0.81 |
| corrected power | 0.56 | 0.39 | 0.48 | 0.20 | 0.71 |

Table 5 - Simulation results when G = 2 and the number of cases and controls is 500

As can be seen from Tables 5 and 6, KNN-MDR seems to show reasonable power when compared to its competitors. More importantly, corrected power of the method is significantly better than for the other tested methods (after 100 simulations, p = 0.0143 when comparing KNN-MDR to its closest competitor for G = 2 and p = 7.23e-7 for G = 3).

| Method | MDR | AntEpiSeeker | BOOST | MegaSNPHunter | KNN-MDR |
|---|---|---|---|---|---|
| power | N/A | 0.65 | 0.67 | 0.80 | 0.74 |
| corrected power | N/A | 0.15 | 0.28 | 0.12 | 0.63 |

Table 6 - Simulation results when G = 3 and the number of cases and controls is 500

A short literature survey (Prabhu et al. 2012, Upstill-Goddard et al. 2013, Li et al. 2014, Wei et al. 2014) leads to the conclusions that many of the methods seem to be marred by high false positive rates. To test that, we have simulated situations where no SNP was involved in the generation of the phenotypes, so that SNP detection by the algorithms would correspond to false positives. Table 7 shows the results of these simulations.

We ran another set of simulations to assess the respective effects of the sizes of the windows and of the number K of neighbors on the (corrected) detection power. Results of these simulations are reported in Table 8.

## Results on WTCCC data

Since working on such a large dataset (> 300k SNP) is very demanding in terms of computing time, we proceeded as follows:

1- 20k SNP were first extracted from the data. Although several selection procedures could be applied, we simply selected 1 SNP every 15 SNP.

2- We divided the data into 200 windows of 100 SNP each.

3- We then tested each of the 19900 possible pairs of windows (sets of 200 SNP) using KNN-MDR.

4- We extracted the 83 sets for which the p-values were lower than 2.5e-6 (a threshold obtained after Bonferroni correction at level 0.05). To reach that significance level using a permutations procedure, we used the following adaptative scheme: after 100 permutations performed on the 19900 possible pairs of windows, only those reaching the 0.05 level were considered for the next round of permutations, assuming that those not reaching that level of significance were very unlikely to reach the desired significance at the end of the process. This left us with 2319 combinations. In the next round, 900 more permutations were performed, and only the combinations reaching the 0.005 level were kept (i.e. 1207 combinations). Repeating this procedure for 1.0e4, 1.0e5, 1.0e6 and 2.0e6 permutations, and respective thresholds equal to 5.0e-4, 5.0e-5, 5.0e-6 and 2.5e-6, we ended up with the 83 sets cited above.

5- The SNP hidden in step 1 were then recovered, leading to 83 sets of 3000 SNP (i.e. 200*15).

6- KNN-MDR was applied on every set from step 5: the sets were divided into 30 windows of 100 SNP and all 435 combinations of windows pairs in each set were considered by KNN-MDR.

7- We kept the 241 sets of 200 SNP with a p-value < 1.15e-4 (Bonferroni correction at level 0.05).

8- MDR was then used for the sets from the previous step, leading to examine 19900 SNP-SNP interactions for each set.

9- The interactions with a p-value < 2.51e-6 (Bonferroni correction at level 0.05) were then considered as significant.

Results from this analysis are presented in Table 9. The full version of Table 9 is provided in a supplementary file. Figure 11 provides a view of the significant results at the chromosome level for our study as well as for 2 other similar studies on this dataset ((Zhang et al. 2012) and (Shchetynsky et al. 2015)).

## Discussion

This paper has introduced a new MDR approach to find markers interactions in genomic scans. It could also be used for other attributes than markers, such as environmental factors, leading to a gene-environment interaction search method. Due to the proposed strategy relying on the MDR approach, and in parallel with a recent study (Chen et al. 2011) using ("simple") MDR as a reference strategy, we have compared our proposed method's performances to this reference and other reference methods (MegaSNPHunter, AntEpiSeeker, BOOST), and tried to show that our method could have benefits compared to these methods. Of course, other algorithms might have been tested, such as the recent Bayesian High-order Interaction Toolkit (Wang et al. 2015) which is proposing a MCMC approach to scan the very large search space of potential sets of markers (incidentally, this algorithm has also been tested on a smaller set of simulations, and its power has been found significantly lower than KNN-MDR on this dataset). Our point in this respect was not to be exhaustive, but simply to show that the approach we propose can bring some more information than other popular methods, and might be a useful addition to the arsenal developed to tackle genomic interaction problems.

| Method | MDR | AntEpiSeeker | BOOST | MegaSNPHunter | KNN-MDR |
|---|---|---|---|---|---|
| Power (p-value <0.05) | 0.18 | 0.45 | 0.19 | 0.38 | 0.07 |

Table 7 - Simulation results when G = 0 and the number of cases and controls is 500

The results obtained through the simulations demonstrate some of the features that potentially make KNN-MDR helpful. More specifically, the simulations show the feasibility of scans using large number of markers, as opposed to MDR where the computer burden explodes with the number of markers (when it simply increases linearly with KNN-MDR). This might allow to highlight interactions between markers far apart on the genomic map (trans-interactions), while some strategies proposed to restrict the scans to close-by markers (cis-interactions) to reduce the amount of computations.

We now discuss some of the features of the method:

## Number of interacting loci

In this paper, although the algorithm given in the appendix can be used for G larger than 3, only 3 markers have been used to generate the phenotypes. Nevertheless, in practical applications, it is not unlikely that situations involving more than 3 loci might exist. These situations might increase the interest of using methods such as KNN-MDR. Indeed, when more regions are involved in the phenotype, this could decrease the distance measure between individuals sharing some or all of these regions and better cluster individuals sharing the same status. Conversely, in MDR, discovering such complex patterns would likely necessitate to increase the number of loci scanned simultaneously, which would make computations even more difficult. Also, increasing the number of loci increases the number of cells with no (or very few) observations, making status allocation potentially inaccurate or even impossible.

## Parameter settings

We mentioned earlier that parameters setting in KNN-MDR mainly involves defining the sizes, positions and the number of windows, the number K of neighbors and the distance measure. All parameters are problem dependent, making it difficult to devise general rules. Nevertheless, some guidelines might be given.

| | | W= | | | |
|---|---|---|---|---|---|
| | | 5 | 10 | 15 | 20 |
| K= | 5 | 71 | 68 | 62 | 52 |
| | | 65 | 62 | 51 | 38 |
| | 10 | 70 | 66 | 64 | 56 |
| | | 60 | 53 | 51 | 43 |
| | 15 | 71 | 65 | 59 | 58 |
| | | 59 | 49 | 47 | 44 |
| | 20 | 69 | 60 | 56 | 53 |
| | | 67 | 55 | 52 | 45 |

Table 8 - Power (above) and corrected power (below) when the parameters K (number of markers) and W (windows size) are varied in 100 simulations with 500 cases and 500 controls and G = 2.

In all the analyses performed in this study, we have only used Mahalanobis distances, as already mentioned. The reason was that this distance allows to take into account potential correlations between attributes (typically, linkage disequilibrium between close markers) and because it makes it possible to weight the attributes in the sum (for example to take into account that similarity for rare alleles is more informative that on frequent ones). In our studies, only SNPs have been used, for which the distance proposed in the Mahalanobis measure makes sense, with D(AA,AB) = D(AB,BB) = 0.5* D(AA,BB), where AA, AB and BB are the three possible SNP genotypes. This might be different and might need more investigations if other types of genetic variants are used. Note also that, in most computations, to reduce the computational burden, the correlation between neighboring markers has not been estimated but set to 0 (i.e. we used the normalized Euclidean distance), which might potentially affect the power. Although we did not explicitly test this, we expect that including the correlations would lead to better take into account the linkage disequilibrium, which should have a positive effect on the detection power. So, using this information might be favorable in terms of power, but at the cost of an increase in the computation time. Note also that using this kind of distance makes less sense when working with markers with more than 2 alleles, unless it can be postulated that the distance between, for example, alleles 1 and 3 is roughly twice the distance between alleles 1 and 2. An easy to compute and similar distance measure would then be to square the number of differing alleles (0, 1 or 2) between two compared genotypes, to normalize as for the Mahalanobis distance, to sum over all markers in the window and to take the square root of the product. This “binary” distance is implemented in our KNN-MDR software.

| SNP | Position | Testing balanced accuracy | P-value |
|---|---|---|---|
| rs10979420,rs778980 | 9:108634242 , 19:5863725 | 0,894054 | $2.51*10^{-6}$ |
| rs10979420,rs778982 | 9:108634242 , 19:5866574 | 0,894054 | $2.51*10^{-6}$ |
| rs6781338,rs778982 | 3:180060018 , 19:5866574 | 0,88983 | $2.51*10^{-6}$ |
| rs778980,rs17325560 | 19:5863725 , 20:2614933 | 0,88983 | $2.51*10^{-6}$ |
| rs4979291,rs10979420 | 9:107732763, 9:108634242 | 0,88983 | $2.51*10^{-6}$ |
| rs561259,rs10979420 | 2:79014325 , 9:108634242 | 0,88983 | $2.51*10^{-6}$ |
| rs1862333,rs17325560 | 5:181066946, 20:2614933 | 0,888751 | $2.51*10^{-6}$ |
| rs1862333,rs485409 | 5:181066946 ,18:28918712 | 0,888751 | $2.51*10^{-6}$ |
| rs571307,rs578044 | 13:29942173,18:28918696 | 0,887092 | $2.51*10^{-6}$ |
| rs1169565,rs571307 | 2:71196518 , 13:29942173 | 0,880437 | $2.51*10^{-6}$ |

Table 9 - The 10 most significant results of the analysis on the RA dataset from WTCCC

For the windows dimensions, our idea is to use the assumption that individuals sharing mutations responsible for the trait should look more similar in the surroundings of these mutations than those not sharing these mutations. The resemblance should thus extend to neighbouring markers, where the neighbourhood size is a function of the linkage disequilibrium (LD) in the region. In situations where LD increases (due to the studied population and/or the markers density), distance between individuals sharing genomic regions (including the causal regions) should decrease and detection power should increase. Note that this genomic feature is ignored in the other tested methods. Accordingly, the windows sizes W should ideally be defined to capture the local linkage disequilibrium. Since the measurable LD is dependent on the population history and on the markers density, assessment of this measure should first be made in order to have reference dimensions for the various windows to be used in KNN. Note that the extent of LD need not be the same across the whole genome: accordingly, the size of the windows might be varied along the genome to better reflect the underlying structure and better capture the relevant information.

To illustrate that expected behavior, we have performed the simulations leading to Table 8. As visible from that table, the powers decrease when the windows sizes increase. Our interpretation of this result is that, due to the way the simulated data are generated, chunks of five linked (i.e. showing some LD) markers are used, which should restrict the signal caused by LD to five markers. Adding more markers to the windows adds noise, and consequently reduces the resemblance between the composite pieces of chromosomes harboring the causative mutations, and thus the power.

Next, the number K of neighbors should somehow reflect the number of individuals sharing regions harboring causal mutations. This number is of course unknown and difficult to evaluate a priori because it is dependent on various population and trait parameters such as the history of the population or the genetic heterogeneity of the trait. Furthermore, it might vary from region to region, making it difficult to devise general rules allowing to infer relevant values of K. Possible “brute force” approaches would be to rerun the algorithm with varying number of neighbors (grid search) or to use bootstrap methods (Hall et al. 2008). This strategy could allow to capture regions of interest while integrating potential sources of variations, at the cost of supplementary computer burden. Another point of view is that the corrected powers do not significantly (at the 5% level) disagree between the various K values for the tested windows sizes, which indicates that the results might not be very sensitive to this parameter, at least in our simulations. For this reason, we used K = 5 or K = 10 in our computations. Note also that odd K values might facilitate the majority vote.

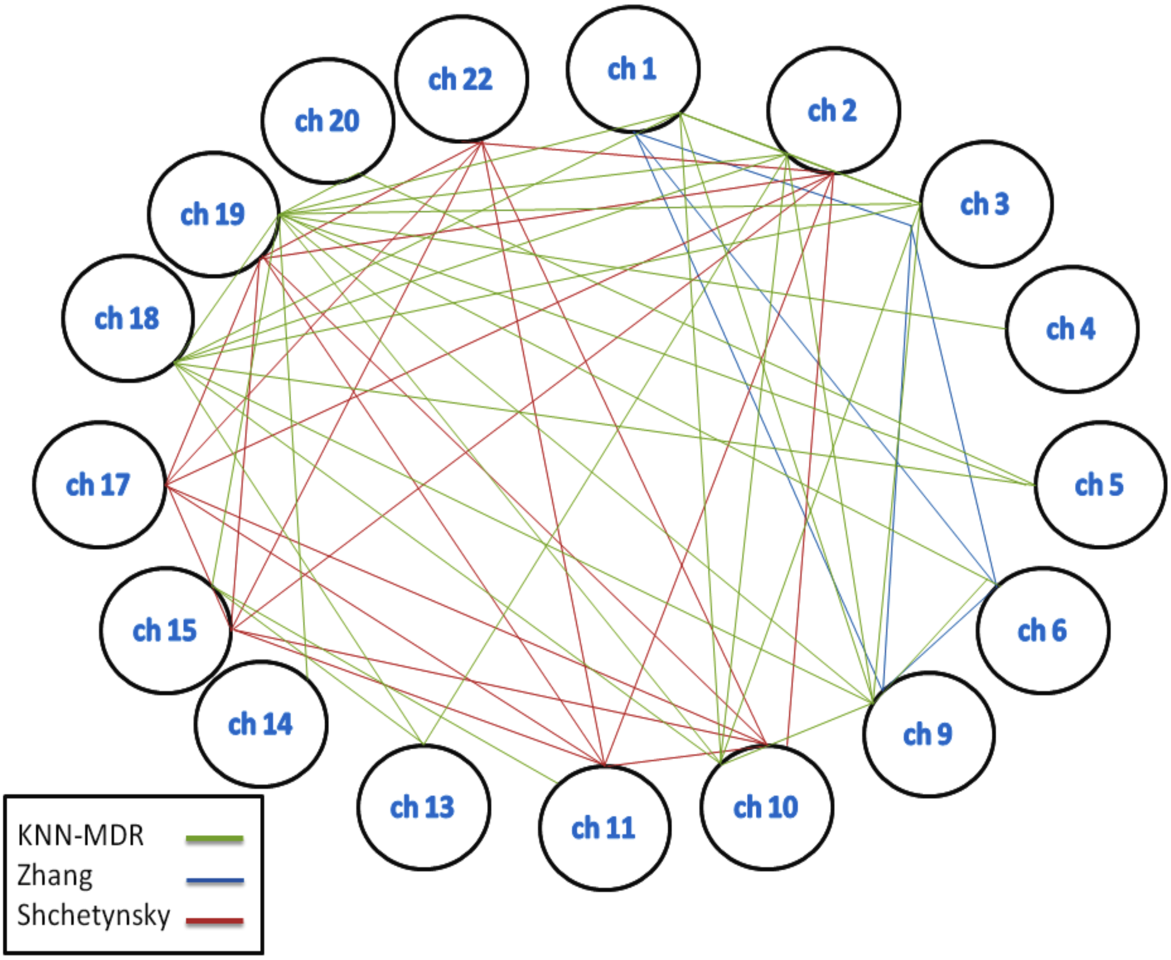

Figure 11 - Comparison of the inter-chromosomal interactions detected on the RA dataset by KNN-MDR and other interaction methods using this same dataset as example (Shchetynsky et al. (2015); Zhang et al. (2012))

### False positive rates

Our simulations have shown that, as reported in other studies, results are often penalized by high false positive rates (Table 7). One obvious reason is multiple testing: the large number of performed tests necessitates that the significance threshold be properly adapted, which is not always easy to do. Another reason in our study is the way we have performed the simulations. Indeed, we have managed to have epistatic interactions with little marginal effects in order to avoid the easier situations where individual loci can be identified in a first step, followed by the identification of interactions between these loci identified first in a second step. To obtain these situations, we have used multi-locus prevalences, which has led to some kind of genetic heterogeneity: a same multi-locus genotype could simultaneously be present in cases and in controls, making it harder to identify these loci. These complicating factors have been associated to higher false positive rates in other studies, along with other design factors such as the number of cross-validation subsets (Winham et al. 2010, Cattaert et al. 2011, Li et al. 2014). Our model might be less sensitive to these factors: looking for neighbors might allow selecting the individuals sharing the relevant features in a heterogeneous set of individuals.

Also, decreasing the number of tests (in comparison to MDR, for example), might also lead to somehow relaxing the penalty arising from multiple testing.

## Power and corrected power

The reason for the drop in the power of the alternative methods when considering the accuracy is not completely clear, but we can suggest a tentative explanation.

As can be seen from Table 7, all methods show high rates of false positive results, while KNN-MDR seems to behave reasonably well from that point of view. Although this is no definite proof, this is an indication that the high power observed in the simulations for the alternative methods is probably due to false positive results. Correcting for the accuracy (using “corrected power”) therefore eliminates most of these false positive results, so drastically reducing the observed power.

A potential criticism on our “accuracy measure” is that using windows sets makes it more likely to cover the culprit regions, and so this “accuracy” measure is biased in favor of KNN-MDR. For that reason, and to make the comparison fair between the methods, we have chosen the parameters to end up with similar number of markers in the finally selected markers sets in each approach. Note nevertheless that the resolution of KNN-MDR could eventually be increased in these analyses, for example using the strategy described for large datasets in the material and methods section.

Figure 11 shows that no combination at the chromosome level is consistent across our study and two other similar studies on the same dataset ((Zhang et al. 2012) and (Shchetynsky et al. 2015)) while other significant results are specific to one or two methods. Some results from KNN-MDR are consistent with those obtained by Shchetynsky, others are consistent with those of Zhang while no corresponding results between Zhang and Shchetynsky studies could be found. Power and false positive issues might potentially explain these discrepancies, although no definite proof can be put forward based on these preliminary analyses.

So, in our study as in the other ones, statistically significant SNP interactions have been identified using KNN-MDR and MDR in a genome-wide association study. Their biological relevance is obviously not clear at this stage and needs more investigations in the future. We can nevertheless say that some of our results are consistent with other results in the domain of Rheumatoid Arthritis ((Zhang et al. 2012), (Wan et al. 2010), (Hua et al. 2012)) and that, in addition, new candidates contributing to the etiology of this disease have potentially been identified. This result shows that, as suggested in the simulations, differences in the approaches and potential differences in the respective powers of the used methods might lead to new insights in the etiology of the disease. This observation should trigger more research on the use of composite methods, combining the qualities of several approaches.

### Computer resources

In our results, the comparisons between (MegaSNPHunter, AntEpiSeeker , BOOST, MDR) and KNN-MDR in terms of computer resources has not been fully addressed. Nevertheless, it has been shown how and why KNN-MDR decreases the computer load with respect to MDR, making it a potential candidate to analyze large datasets, as shown for the RA data. To be fair, it should be mentioned that computing nearest neighbors is more computer intensive than a majority vote in the subset sharing the same multi-locus genotype. Nevertheless, as shown in the simulations, and as can be understood from the previous discussion, computations remain more affordable in KNN-MDR than in MDR and the other methods for similar scans. Furthermore, strategies could also be devised to make KNN-MDR efficient, such as pre-computing distances for windows and using distance additivity properties to compute distance over several windows.

Another point that might be worth adding is that, although KNN is natively a classification method, we have used it here in a detection context. KNN-MDR could nevertheless as well be used as a classification tool: to that end, the best model (i.e. the best set of markers) could be used to compute the neighborhood of a new individual and classify the latter in one or the other category.

## Conclusions

In summary, KNN-MDR is an alternative to existing methods for detecting epistatic interactions, with interesting features. Among these, we have demonstrated that KNN-MDR is more computationally efficient than other exhaustive strategies, facilitating the analysis of large-scale data sets with potentially genome-wide SNPs. The method is also capable to detect high-order interactions and to take into account linkage disequilibrium (LD). Another advantage is that it is able to detect interactions between SNPs even in the absence of marginal effects. Also, the method is non-parametric: no prior distribution is assumed, unlike many parametric-statistical methods. Nevertheless, parameters (distances, number of neighbors, windows definition) are available to allow some flexibility in the search strategies, which could help to render the method useful in other classification contexts.

Although KNN-MDR is potentially beneficial for epistasis detection, several aspects would nevertheless deserve more investigations. For example, the burden associated to the computation of the K nearest neighbors could become an issue when the dataset is very large. Since the load increases quadratically with the number of individuals, and linearly with the number of markers, improving the computational performances of the method could necessitate some code optimization to make the program more efficient. Another point necessitating more work is the tuning of the parameters allowing an optimal detection power. This includes the optimal sizes of the windows - which should be dependent on the studied population, the markers density, the LD pattern, the optimal size of the neighborhoods to be considered, the pre-selection of markers in the early phase of large dataset analyses, the distance measure or the adaptative selection scheme for the selection of markers in large studies, among others.

## *Abbreviations*

**GWAS:** Genome-wide association study

**MDR:** Multifactor dimensionality reduction

**BOOST:** Boolean operation-based screening

**LD:** Linkage disequilibrium

**KNN:** K-nearest neighbors

**SNP:** Single-nucleotide polymorphism

**MAF:** Minor allele frequency

**HSA9:** Human chromosome 9

**FAM-MDR:** Flexible family-based multifactor dimensionality reduction

**MB-MDR:** Model-based multifactor dimensionality reduction

**GMDR:** Generalized multifactor dimensionality reduction

**WTCCC:** Wellcome trust case control consortium

**BHIT**: Bayesian High-order Interaction Toolkit

## *Declarations*

### Acknowledgements

None.

### Funding

No specific funding was provided for this study.

### Availability of data and software

The software developed to implement the KNN-MDR program is available on http://www.fmv.ulg.ac.be/cms/c_1802261/fr/publiclyavailable-softwares

and the simulations data are available in additional files (1 to 5). A users' guide for the software is provided as Appendix 1 (Additional file 6). This study makes use of data generated by the Wellcome Trust Case-Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. All data from WTCC have been used according to the terms of the WTCCC data access agreement. Relevant publication for the used dataset is (Wellcome Trust Case Control Consortium, 2007): Nature 2007;447;7145;661-78 (PUBMED: 17554300; PMC: 2719288; DOI: 10.1038/nature05911). Access to the data from WTCCC needs to be obtained from the Consortium.

Authors' contributions

AS participated in the design of the study and performed the statistical analysis and helped to draft the manuscript. FF conceived the study, wrote the software, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Consent for publication

Not applicable.

### Ethics and consent to participate

Not applicable

### Open Access

## References

Aci, M., C. İnan and M. Avci (2010). "A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm." Expert Systems with Applications 37(7): 5061-5067.

Ban, H.-J., J. Y. Heo, K.-S. Oh and K.-J. Park (2010). "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine." BMC Genetics 11(1): 26.

Boos, D. D. and L. A. Stefanski (2011). "P-Value Precision and Reproducibility." Am Stat 65(4): 213-221.

Calle, M., V. Urrea, N. Malats and K. Van steen (2008). MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. Genètica general. Citogenètica general. Immunogenètica. Evolució. Filogènia. Universitat de Vic, 2008-02-05.

Cattaert, T., M. L. Calle, S. M. Dudek, J. M. Mahachie John, F. Van Lishout, V. Urrea, M. D. Ritchie and K. Van Steen (2011). "Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise." Ann Hum Genet 75(1): 78-89.

Cattaert, T., V. Urrea, A. C. Naj, L. De Lobel, V. De Wit, M. Fu, J. M. Mahachie John, H. Shen, M. L. Calle, M. D. Ritchie, T. L. Edwards and K. Van Steen (2010). "FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals." PLoS One 5(4): e10304.

Chen, L., G. Yu, C. D. Langefeld, D. J. Miller, R. T. Guy, J. Raghuram, X. Yuan, D. M. Herrington and Y. Wang (2011). "Comparative analysis of methods for detecting interacting loci." BMC Genomics 12: 344.

Collins, R., T. Hu, C. Wejse, G. Sirugo, S. Williams and J. Moore (2013). "Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis." BioData Mining 6(1): 4.

De los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel and J. Crossa (2010). "Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods." Genet Res (Camb) 92(4): 295-308.

Fang, G., M. Haznadar, W. Wang, H. Yu, M. Steinbach, T. R. Church, W. S. Oetting, B. Van Ness and V. Kumar (2012). "High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions." PLoS One 7(4): e33531.

Fang, Y. H. and Y. F. Chiu (2012). "SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies." Genet Epidemiol 36(2): 88-98.

Gori, A. S., E. Théâtre, B. Charloteaux, Y. Momozawa, V. Deffontaine, D. Baurain, M. Mni, F. Crins, N. Ahariz, C. Oury, C. Lecut, C. Reenaers, P. Gast, C. Van Kemseke, P. Leclercq, E. Louis and M. Georges "Fine-mapping and functional analysis of the 5p13.1 risk locus for Crohn's disease." Am J Hum Genet.

Gui, J., A. S. Andrew, P. Andrews, H. M. Nelson, K. T. Kelsey, M. R. Karagas and J. H. Moore (2011). "A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility." Ann Hum Genet 75(1): 20-28.

Gunther, F., N. Wawro and K. Bammann (2009). "Neural networks for modeling gene-gene interactions in association studies." BMC Genet 10: 87.

Hall, P., B. U. Park and R. J. Samworth (2008). "Choice of neighbor order in nearest-neighbor classification." The Annals of Statistics 36(5): 2135-2152.

Howie, B. N., P. Donnelly and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet 5(6).

Hua, L., H. Lin, D. Li, L. Li and Z. Liu (2012). "Mining Functional Gene Modules Linked with Rheumatoid Arthritis Using a SNP-SNP Network." Genomics, Proteomics & Bioinformatics 10(1): 23-34.

Koo, C. L., M. J. Liew, M. S. Mohamad and A. H. Salleh (2013). "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology." BioMed Research International 2013: 432375.

Li, C. F., F. T. Luo, Y. X. Zeng and W. H. Jia (2014). "Weighted risk score-based multifactor dimensionality reduction to detect gene-gene interactions in nasopharyngeal carcinoma." Int J Mol Sci 15(6): 10724-10737.

Lishout, F. V., F. Gadaleta, J. H. Moore, L. Wehenkel and K. V. Steen (2015). "gammaMAXT: a fast multiple-testing correction algorithm." BioData Mining 8(1).

Lou, X. Y., G. B. Chen, L. Yan, J. Z. Ma, J. Zhu, R. C. Elston and M. D. Li (2007). "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence." Am J Hum Genet 80(6): 1125-1137.

Mahachie John, J. M., F. Van Lishout and K. Van Steen (2011). "Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data." Eur J Hum Genet 19(6): 696-703.

Manceau, M., V. S. Domingues, C. R. Linnen, E. B. Rosenblum and H. E. Hoekstra (2010). "Convergence in pigmentation at multiple levels: mutations, genes and function." Philos Trans R Soc Lond B Biol Sci 365(1552): 2439-2450.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature 461(7265): 747-753.

Manuguerra, M., G. Matullo, F. Veglia, H. Autrup, A. M. Dunning, S. Garte, E. Gormally, C. Malaveille, S. Guarrera, S. Polidoro, F. Saletta, M. Peluso, L. Airoldi, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, D. Trichopoulos, A. Kalandidi, D. Palli, V. Krogh, R. Tumino, S. Panico, H. B. Bueno-De-Mesquita, P. H. Peeters, E. Lund, G. Pera, C. Martinez, P. Amiano, A. Barricarte, M. J. Tormo, J. R. Quiros, G. Berglund, L. Janzon, B. Jarvholm, N. E. Day, N. E. Allen, R. Saracci, R. Kaaks, P. Ferrari, E. Riboli and P. Vineis (2007). "Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions." Carcinogenesis 28(2): 414-422.

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

Millstein, J. (2013). "Screening-testing approaches for gene-gene and gene-environment interactions using independent statistics." Front Genet.

Musani, S. K., D. Shriner, N. Liu, R. Feng, C. S. Coffey, N. Yi, H. K. Tiwari and D. B. Allison (2007). "Detection of gene x gene interactions in genome-wide association studies of human population data." Hum Hered 63(2): 67-84.

Park, M. Y. and T. Hastie (2008). "Penalized logistic regression for detecting gene interactions." Biostatistics 9(1): 30-50.

Prabhu, S. and I. Pe'er (2012). "Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease." Genome Res 22(11): 2230-2240.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet 81(3): 559-575.

Ritchie, M., L. W. Hahn and J. H. Moore (2003). "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity." Genet Epidemiol 24(2): 150-157.

Ritchie, M., W. Hahn, N. Roodi, L. Bailey, D. Dupont, F. Parl and H. Moore (2001). "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer." Am J Hum Genet 69(1).

Ritchie, M. and A. Motsinger (2006). "Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene – gene interactions in human genetics and pharmacogenomics studies." Hum Genomics 2(5).

Schwarz, D. F., I. R. König and A. Ziegler (2011). "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data." Bioinformatics 27(3): 439-439.

Shang, J., J. Zhang, Y. Sun, D. Liu, D. Ye and Y. Yin (2011). "Performance analysis of novel methods for detecting epistasis." BMC Bioinformatics 12: 475.

Shchetynsky, K., D. Protsyuk, M. Ronninger, L. M. Diaz-Gallo, L. Klareskog and L. Padyukov (2015). "Gene-gene interaction and RNA splicing profiles of MAP2K4 gene in rheumatoid arthritis." Clin Immunol 158(1): 19-28.

Suguna, N. and K. Thanushkodi (2010). "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm." International Journal of Computer Science 7(4).

Upstill-Goddard, R., D. Eccles, J. Fliege and A. Collins (2013). "Machine learning approaches for the discovery of gene-gene interactions in disease data." Brief Bioinform 14(2): 251-260.

Usai, M. G., A. Carta and S. Casu (2012). "Alternative strategies for selecting subsets of predicting SNPs by LASSO-LARS procedure." BMC Proceedings 6(Suppl 2): S9.

Ver Hoef, J. M. and H. Temesgen (2013). "A comparison of the spatial linear model to Nearest Neighbor (k-NN) methods for forestry applications." PLoS One 8(3): e59129.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five years of GWAS discovery." Am J Hum Genet 90(1): 7-24.

Wan, X., C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang and W. Yu (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." Am J Hum Genet 87(3): 325-340.

Wan, X., C. Yang, Q. Yang, H. Xue, N. L. Tang and W. Yu (2009). "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study." BMC Bioinformatics 10: 13.

Wang, J., T. Joshi, B. Valliyodan, H. Shi, Y. Liang, H. T. Nguyen, J. Zhang and D. Xu (2015). "A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies." BMC Genomics 16(1).

Wang, Y., X. Liu, K. Robbins and R. Rekaya (2010). "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm." BMC Res Notes 3: 117.

Wei, W. H., G. Hemani and C. S. Haley (2014). "Detecting epistasis in human complex traits." Nat Rev Genet 15(11): 722-733.

Wellcome Trust Case Control, C. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145): 661-678.

Winham, S., A. Slater and A. Motsinger-Reif (2010). "A comparison of internal validation techniques for multifactor dimensionality reduction." BMC Bioinformatics.

Winham, S. J., C. L. Colby , R. R. Freimuth, X. Wang, M. d. Andrade, M. Huebner and J. M. Biernacka (2012). "SNP interaction detection with Random Forests in high-dimensional genetic data." Bioinformatics 13(164).

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter and X. Lin (2010). "Powerful SNP-set analysis for case-control genome-wide association studies." Am J Hum Genet 86(6): 929-942.

Wu, X., H. Dong, L. Luo, Y. Zhu, G. Peng, J. Reveille and M. Xiong (2010). "A Novel Statistic for Genome-Wide Interaction Analysis." PLoS Genet 6(9).

Yang, C., Z. He, X. Wan, Q. Yang, H. Xue and W. Yu (2009). "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies." Bioinformatics 25(4): 504-511.

Zhang, J., Z. Wu, C. Gao and M. Zhang (2012). "High-Order Interactions in Rheumatoid Arthritis Detected by Bayesian Method using Genome-Wide Association Studies Data " American Medical Journal 3(1).

## *Additional files*

**Additional file 1** – The data set(s) supporting the results of Table 5.

**Additional file 2** – The data set(s) supporting the results of Table 6.

**Additional file 3** – The data set(s) supporting the results of Table 7.

**Additional file 4** – The data set(s) supporting the results of Table 8.

**Additional file 5** – Table 9 complete.

**Additional file 6** – KNN MDR user's guide.

**Additional file 7** – Competitor methods.

**Additional file 8** – Computing multi-locus penetrances.

**Note**: the additional files (1-5) can be found on:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1599-7
The other additional files are available as Appendices 1-3 of the present work.

# Study 2: Aggregation of experts: an application in the field of “interactomics” (detection of interactions on the basis of genomic data)

# *Preamble*

The main idea of this study, on the aggregation of methods, is a more general concept: grouping various methods results might lead to improvements over the individual results. We have illustrated this concept in the field of interactions mapping and obtained results somehow confirming these improvements. In a field such as genetic mapping, where thousands of parallel tests are performed, false positives control is an important issue. False positives are likely to be numerous in these studies, meaning a potential waste of time, energy and money on non-reproducible results, which of course shed doubts on the utility of such studies. We hope that our work is again a step in the direction of an improvement in the perception of interaction studies: we have shown that using a small set of methods in a very simple aggregation strategy led to an increase in the detection power while properly controlling for the false positive rate. Above providing a framework for a joint - i.e. made with several methods - analysis of real datasets, we hope that such results will stimulate interest in the development of new methods: this would be beneficial for the field - new performing methods would be welcome - and for the aggregation strategy - adding more performing methods should enhance the performances of the aggregation strategy. The feasibility of using such methodology on real genome-wide datasets has been demonstrated on an example. This also calls for improvement in the future methods to be developed, because many methods in use today would not be able to manage large genome-wide datasets, which questions on their ability to detect interactions involving variants distantly located on the genome.

# Study 2: Aggregation of experts: an application in the field of "interactomics" (detection of interactions on the basis of genomic data)

*BMC Bioinformatics under review*

Sinan Abo Alchamlat and Frédéric Farnir

# Abstract

## Background

The last decades have seen major developments in the field of genetics. These advances have led to the emergence of a wealth of biological information, allowing new strategies to be applied in many fields of the biological research, such as new methods aiming at the genomic mapping of genes involved in traits of interest. Although these approaches have been successful, many elements in the genetic landscape of the studied traits are still unknown and uncharacterized so far. A potential track to new findings, potentially leading to a better understanding of complex determinisms, is the detection of interactions between regions rather than single region associations. While many methods aiming at the detection of such interactions have been proposed, and despite the success of these methods to solve some problems and to detect genetic interactions, there is currently no gold standard method able to detect interactions in all situations, and the relative performances of these methods remain largely unclear. Our work, which is an attempt to try to benefit from the various advantages of a set of methods, is a contribution to this field**.**

## Results

Experimental results, based on simulated data and real genome-wide data show that the aggregated predictor can produce better performances in the detection of genetic interactions than each individual predictor can.

## Conclusions

The presented methodology - an aggregation of experts as a tool to detect genetic interactions- is a potentially useful addition to the arsenal used in complex traits analyses.

## Keywords

Gene-gene interaction - Epistasis - Single Nucleotide Polymorphism - Genome-wide association study - Multifactor Dimensionality Reduction - K-nearest neighbors.

## Background

The last years have witnessed an explosion in the availability of large datasets in various fields. This phenomenon, related to the rapid evolution in the communication facilities offered to a large share of the population, is also due to the technological developments in many other disciplines, such as satellite imaging or routine access to molecular data in biology laboratories. In the field of molecular biology, the wealth of nowadays available information - including genomic, transcriptomic, proteomic, metabolomic data among others - has led to shift the attention from the data acquisition processes to the understanding of the actual meaning of the individual elements of this data ocean: the exponential growth of data creates great challenges, not only in finding out how to store and access the data but also, and more importantly, how to process and make sense of it. Therefore, the development of efficient algorithms for processing big data and making good use of it is very important (Ulfarsson et al. 2016).

In genetics, instances of this ’big data’ revolution can be found in the mapping activities: the researchers aiming at establishing the connections between the molecular information and the observed phenotypes (such as diseases, production traits in animals and plants, or morphological traits) have in their hands an extensive (and sometimes exhaustive) repertoire of the variations observed at the molecular level, and have to engage into the process of inventing new methods and new strategies to extract the relevant variants (Paixão et al. 2016). Since a large portion of the genetic landscape underlying many traits of interest in various organisms, including the human, is still unknown and uncharacterized, this field is a very active field of research (JYoun et al. 2016). In this domain, a classical approach is to use genome-wide association studies (GWAS), where the goal is to scan the whole genome using molecular markers densely populating the whole genome (most often, large sets of Single Nucleotide Polymorphisms (SNP), but not exclusively) to look for associations with the trait of interest. Although successful in many studies (Stranger et al. 2010), this approach has not been successful in many other cases, even when complete genomic information (i.e. sequence data) was available. Several reasons might be invoked to explain this situation, such as a small power to detect effects of modest size or oversimplified statistical models (Bashinskaya et al. 2015). If increasing the cohorts sizes used for mapping is difficult or useless, a possible track to tackle this “missing heritability” problem might be to fit more elaborate models, such as those introducing epistatic or gene-environment interactions (Lon et al. 2001, Phillips 2008). Genes interactions are interplays between two or more genes that have an impact on the expression of an organism's phenotype. They are thought to be particularly important to discover the genetic architecture underlying some genetic diseases (Lon et al. 2001, Phillips 2008). Consequently, there has been an increased interest in discovering combinations of markers that are strongly associated with a phenotype even when each individual marker has little or even no effect (Chen et al. 2011). This approach has to face at least two problems: first, modeling and identifying every (or even any) interaction is a potentially very challenging task in today situations where very large sets of markers (up to several millions) are available. Second, from a more statistical point of view, fully modeling the complexity leads to models with large dimensionality, leading to the well-known ‘curse of dimensionality’ problem (Abo Alchamlat et al. 2017): in rough words, the accurate estimation of an increased number of parameters is hampered by the reduced sizes of the tested cohorts. Many methods (such as multifactor dimensionality reduction approach using K-Nearest Neighbors (KNN-MDR) (Abo Alchamlat et al. 2017), multifactor dimensionality reduction (MDR) (Ritchie et al. 2001), MegaSNPHunter (Wan et al. 2009), AntEpiSeeker (Wang et al. 2010), BOolean Operation-based Screening and Testing (BOOST) (Wan et al. 2010), Bayesian epistasis association mapping (BEAM) (Zhang et al. 2007), BHIT (Wang et al. 2015), Random forest (RF) (Breiman 2001), among others) have nevertheless been proposed for detecting such interactions. Despite some successes of these methods to unravel some genetic interactions (Bashinskaya et al. 2015), no unique method was able to detect most of the interactions so far. Furthermore, the relative performances of these methods remain largely unclear and necessitate more investigations. As a step in that direction, we propose using a method based on the principle of the aggregation of experts, where the experts would be a set of popular published methods. In parallel, we highlight some of the features of the individual methods and discuss possible aggregation strategies.

## Methods

Methods of aggregation are not new and have been used extensively to improve classification (Gerardi et al. 2009, Tsyganok 2010). They are one of the hot research topics in supervised learning and seen as methods useful for constructing good ensembles of classifiers (Dietteric 2000). In our study, we aim to combine the results of various popular methods to potentially obtain improved performances in the field of gene-gene interactions mapping. The idea of the method is to extract information from a few experts in order to create new knowledge. When knowledge is generated from multiple experts, it is necessary to combine the various sources of expertise in order to arrive at a consensual knowledge base (MAK et al. 1996). Aggregation of experts, an example of the larger class of ensemble methods where aggregation is the technique allowing to combine information from multiple sources, has been shown to yield more accurate and robust predictions than individual experts on a variety of classification problems (Titov et al. 2010). Using this approach, it is often possible to decrease the amount of redundant data, to filter the wrong results, which include false positive and false negative, and to increase the accuracy of the result (Choi et al. 2014). In this paper, we investigate the aggregation of experts, using published gene interactions mapping methods (described below) as the experts. As can be found in the literature, each of the potential methods that could be used has pros and cons, and no unique method is uniformly better than the others to detect genetic interactions. Our objective when turning to the aggregation of experts was therefore to obtain a comprehensive method able to detect more interactions than each individual method by combining the strengths of these individual approaches while potentially better avoiding false positive results. The very simple idea is therefore to let each method run independently, to finally come up with a final decision based on a consensus obtained from the individual methods results. An easy example within the possible approaches to this consensus is to use the most frequent opinion as the aggregated expert's opinion. We have used this approach in our experiments. One of the objectives of the aggregation is to increase the power to detect real interactions with respect to the individual methods embedded in the aggregated expert. We can obtain, using some assumptions, a rough estimate of the power as follows.

| | Aggregated power | | | | |
|---|---|---|---|---|---|
| $p_i$ | Q = 2 | Q = 3 | Q = 4 | Q = 5 | Q = 6 |
| 0.1 | 0.010 | 0.028 | 0.052 | 0.081 | 0.114 |
| 0.2 | 0.040 | 0.104 | 0.181 | 0.263 | 0.345 |
| 0.3 | 0.090 | 0.216 | 0.348 | 0.472 | 0.580 |
| 0.4 | 0.160 | 0.352 | 0.525 | 0.663 | 0.767 |
| 0.5 | 0.250 | 0.500 | 0.687 | 0.812 | 0.891 |
| 0.6 | 0.360 | 0.648 | 0.821 | 0.913 | 0.959 |
| 0.7 | 0.490 | 0.784 | 0.916 | 0.969 | 0.989 |

Table 10 - Aggregated power as a function of the individual methods power pi (assumed identical) and the number Q of methods.

Assume runs are performed on Q ($\geq$ 2) methods, where each method has a power $p_i$, i = 1, ..., Q. If we assume that the methods are independent (in the sense that results obtained using one method gives no indication on what can be expected from another one; this assumption will be discussed below):

- the probabilities $p_i$ can be multiplied to model situations where two or more methods correctly identify a combination underlying the phenotype,
- it is unlikely that 2 or more independent methods would identify the same false positive combination, given that the number of potential combinations is huge in most practical situations.

Using the second assumption, we will then consider that an interaction is detected as soon as at least 2 of the Q methods detect the same combination. Next, if we consider that 2 results are possible for each method (correct identification of a causative combination = 1, incorrect identification of the causative combination = 0), $2^Q$ situations are possible for the aggregated expert: (0, 0, ..., 0), (1, 0, ..., 0), ..., (1, 1, ..., 1). Each of these k situations ($s_1$, $s_2$, ..., $s_Q$) has a probability and the power of the aggregated method is obtained by summing these Pi over the set of all situations where at least 2 methods are successful:

(1)

Table 10 illustrates this result in (theoretical) situations where all the individual methods have the same power.

In this table, the independence assumption penalizes the aggregated expert in situations where the number Q of methods is low and the individual powers are low (these situations correspond to the grayed cells). On the other hand, adding methods increases substantially the power, especially when the individual powers are high. It can be anticipated that if a supplementary method is not independent from the previous set, the power gain could be different, but it is not clear how this would change the results. Consequently, the performance of the aggregated method will depend on the individual methods performances, on the number of methods but also on the correlation between the methods results. These correlations can be assessed using simulations, either directly – by counting situations where methods provide concordant results above what is expected by chance only – or indirectly – by comparing the simulations results to what is expected under the hypothesis of independent methods (Table 10). A correlation measure could be based on Cohen's kappa measure(McHugh 2012):

where is the number of simulations where methods i and j simultaneously provide a positive result, is the number of simulations where methods i and j provide a non-positive result, N is the number of simulations, and and are the powers of methods i and j, respectively.

In order to cover a range of situations where aggregation could be useful (see Table 10), our work is based on six methods that have been published and used to detect interacting genetic loci involved in the genetic determinism of a trait. A short description of each of these methods is given below, and details can be found in the corresponding publications:

1- MDR: The Multi-Dimensional Reduction (MDR) method is designed to replace large dimension problems with reduced dimension ones, allowing to make inferences based on a smaller set of variables (Ritchie et al. 2001).
2- KNN-MDR is an approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) for detecting gene-gene interactions as a possible alternative, especially when the number of involved determinants is high (Abo Alchamlat et al. 2017).
3- BOOST (Boolean Operation-based Screening and Testing), is a two-stage method (screening and testing) using Boolean coding to improve the computational performances (Wan et al. 2010).
4- MegaSNPHunter (MSH)uses a hierarchical learning approach to discover multi-SNP interactions (Wan et al. 2009).

5- AntEpiSeeker (AES) is an heuristic algorithm derived from the generic Ant Colony Optimization family (Wang et al. 2010).

6- BHIT uses a Bayesian model for the detection of high-order interactions among genetic variants in genome-wide association studies (Wang et al. 2015).

Of course, more and/or other methods could have been used.

Since we wanted to assess the performances of the aggregation method and compare them to the individual methods, we have performed simulations that will be described next.

**Simulations**

One of the aims of our study was to assess the performances of the methods to unravel gene-gene (or gene-environment) interactions in the absence of large marginal effects. The reason for that choice was that many methods are able to detect such large marginal effects and to infer interactions within a limited set of loci selected on that basis. Accordingly, we wanted to devise an approach that is able to detect interactions even in the absence of marginal effects. For that reason, efforts have been devoted to generate datasets with interacting genes in the absence of significant marginal effects. Furthermore, heterogeneity between samples has been shown to be a major source for the non-reproducibility of significant signals (Can et al. 2009). We have modeled heterogeneity by associating penetrances (i.e. Pen = probabilities of a phenotype given a genotype) to the multi-locus genotypes underlying the simulated binary trait. Consequently, individuals carrying the causal alleles could be affected (with a probability equal to Pen) or not.

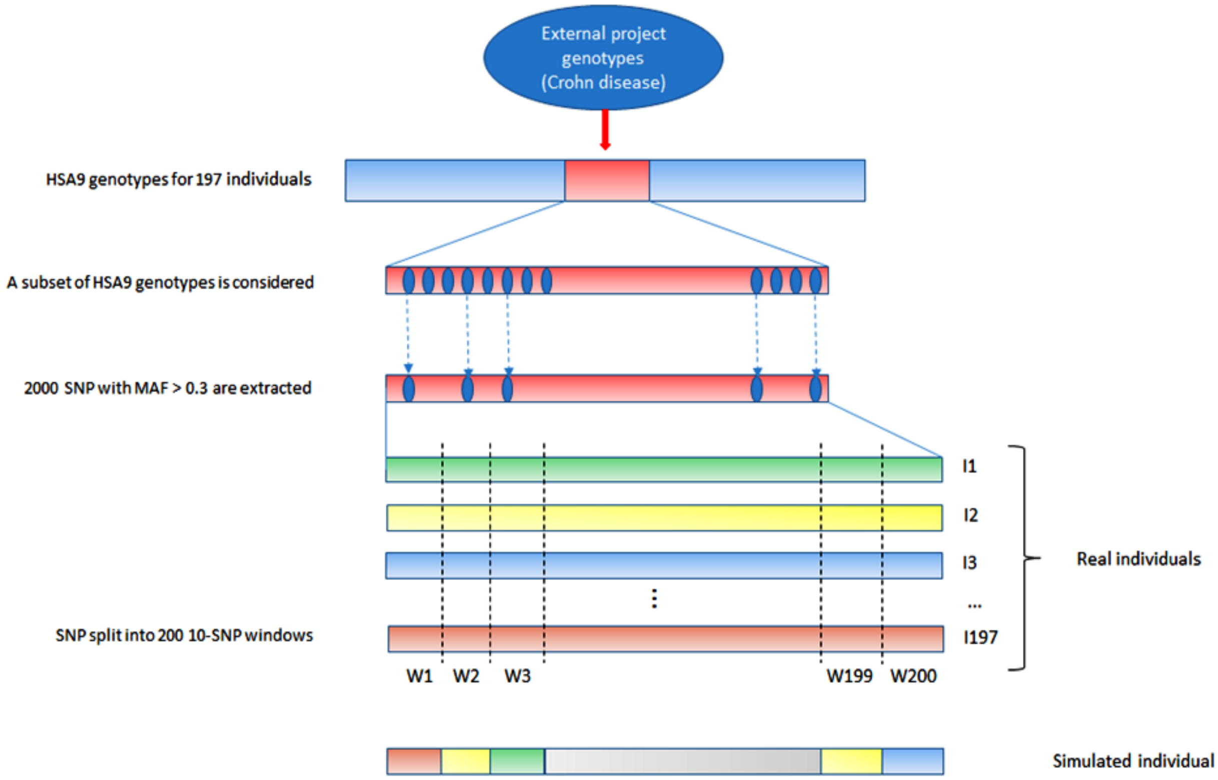
External project genotypes (Crohn disease)
HSA9 genotypes for 197 individuals
A subset of HSA9 genotypes is considered
2000 SNP with MAF > 0.3 are extracted
SNP split into 200 10-SNP windows
I1
I2
I3
...
I197
Real individuals
W1
W2
W3
W199
W200
Simulated individual

Figure 12 - Genotypes generation using a real dataset.

The process can be split into 4 steps:

1. Genotypes generation (see Figure 12).

(1) Genotyping data from a study on Crohn disease in Caucasians (Gori et al. In Press) has been obtained for 197 individuals.

(2) SNPs spanning a combination on chromosome 9 (HSA9) have been extracted, and, to decrease the computational requirements of the simulations, a subset of 2000 informative markers has been selected for our simulations. In order to recover a large part of the information lost in subselecting markers, only markers with a MAF > 0.3 and no missing genotypes have been selected. Subsequent tests (Hardy-Weinberg equilibrium, recovery of a significant linkage disequilibrium) have been carried on to validate the finally used subset (data not shown).

(3) Since many different individuals are needed in the simulations, we have used a trick similar to (Chen et al. 2011) to generate new individuals based on the few (i.e. 197) available genotypes: each individual genotype was chopped into 10 SNP windows, leading to 200 windows. Consequently, each window has (maximum) 197 different 10-loci genotypes. We then built each simulated individual genotype by randomly sampling one of the 197 possible 10-loci genotype for each of the 200 windows and concatenating the 200 10-loci genotypes into a new complete genotype with 2000 markers. This technique allows for $197^{200}$ potentially different individuals while conserving some LD.

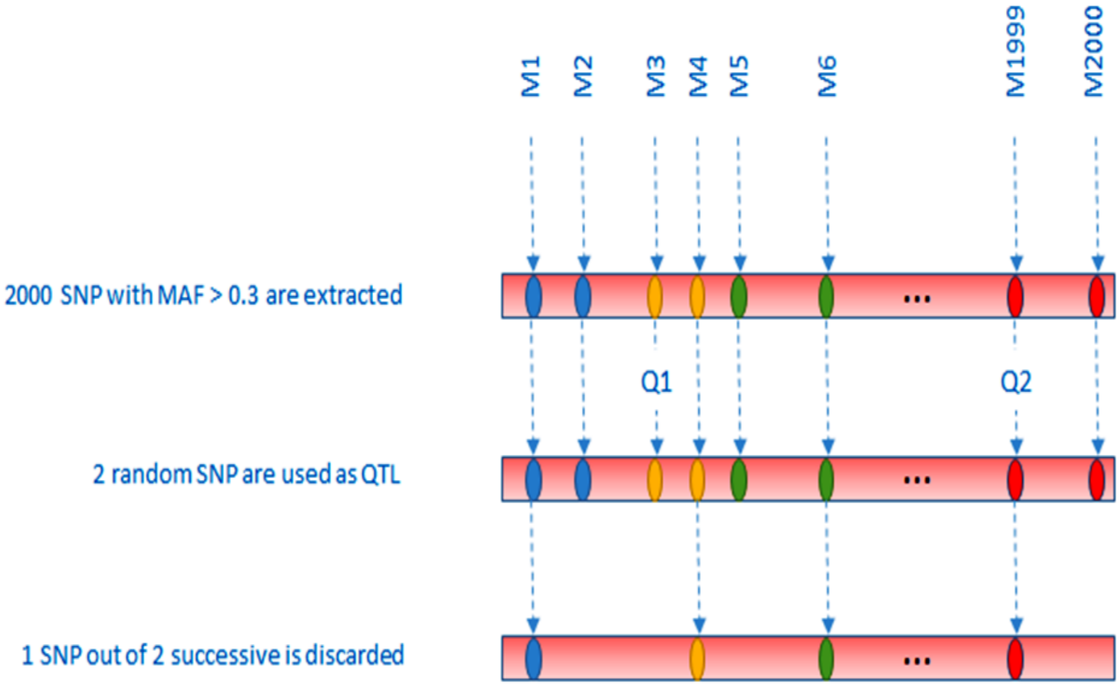
M1
M2
M3
M4
M5
M6
M1999
M2000
2000 SNP with MAF > 0.3 are extracted
...
Q1
Q2
2 random SNP are used as QTL
...
1 SNP out of 2 successive is discarded
...

Figure 13 - QTL (Q1, Q2) used as a basis to generate the interaction.

2. Phenotypes generation (see Figure 13).

(4) 2 SNP were then randomly chosen as having an effect on the simulated phenotype, more complex patterns, i.e. with more involved loci, might be tested but this was not considered in this study. Note that, since SNP selection was random, SNP could be linked or not.

(5) Selected SNP genotypes were used to generate the binary phenotypes. The details of the algorithm are given in (Abo Alchamlat et al. 2017), but, in summary, after generating 2-locus penetrances (Pen) leading to approximately no marginal effect, a uniformly distributed random number R is sampled between 0 and 1 and compared to the penetrance Pen of the simulated 2-locus genotype: if R < Pen, the simulated individual is supposed to be a case (1). If not, it is a control (0)

(6) One SNP out of 2 consecutive SNPs was then randomly discarded, leaving 1000 markers genotypes for the analyses. The rationale of this selection is that causative mutations might nowadays be present or not in the genotyped variants. This will also be the case in our simulations (see Figure 13).

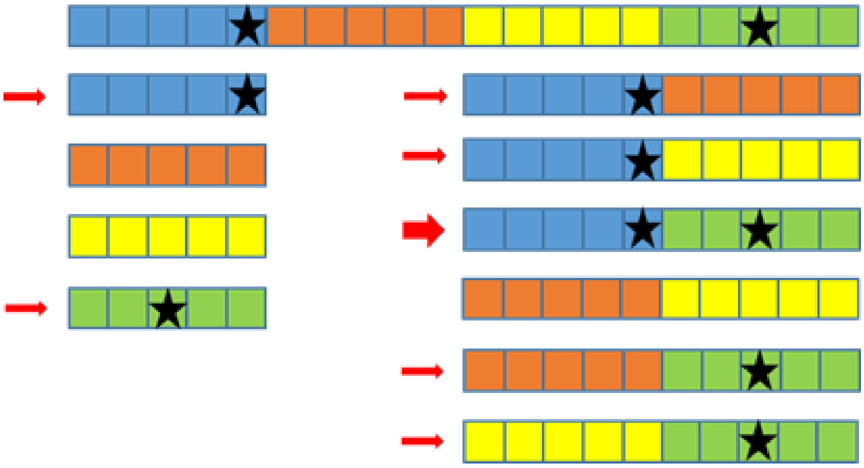

Figure 14 - An example with 20 SNP (represented by squares) partitioned into 4 groups (represented by the colours) of 5 SNP. The causative are marked with a black star. All combinations of one or two groups are then shown, those harbouring one of the causative mutations are signalled using a small red arrow and the (optimal) one harbouring both mutations is shown using a big red arrow.

3. Statistics computation and significance assessment.

(7) The genotypes and corresponding phenotypes were then studied using all 6 methods.

   a. KNN-MDR splits the 1000 SNP into 100 sets of 10 consecutive markers and measures the association between each combination of 1 (100 tests) or 2 sets (4950 tests) with the phenotype using balanced accuracy (Abo Alchamlat et al. 2017). Among all possible combinations, the one considered as optimal is the one containing both causative SNP (see Figure 14). The other approaches use their own statistics to rank the tested combinations associations with the phenotype from strongest to weakest (see (Ritchie et al. 2001), (Wan et al. 2009), (Wang et al. 2010), (Wan et al. 2010), (Wang et al. 2015) for details).

(8) We assessed significance using 100 permutations of the phenotypes for each simulation. Permutation of the phenotypes with respect to the genotypes breaks the potential relationship between phenotypes and genotypes. Accordingly, analyses on permuted data correspond to analyses under the null hypothesis of no association. We kept the highest value of the statistic obtained in each permutation to build the distribution under the null hypothesis, and then compared the statistics obtained with the real (i. e. non permuted) data to this distribution to obtain a p-value for the tested combinations. Although this number of permutations is too low for routine work, it was used to reduce the computing burden and help us to discriminate between results clearly non-significant (i.e. $p > 0.05$) and those potentially significant (i.e. $p < 0.05$). When a higher precision was needed for the p-values (see below for real data), an adaptative permutations scheme was used, in which windows not reaching a pre-determined p-value threshold are progressively abandoned in the permutations scheme since these windows are very unlikely to finally reach a significant result (Purcell et al. 2007).

4. Aggregation of the results.

(9) After completing the simulation and the permutations for each method, we performed a majority vote among the obtained optimal combinations. If one combination obtained the majority, it became the aggregated method’s chosen combination. When no majority could be obtained, the aggregated method failed to obtain a solution (see simulation in Table 11 as an example).

In each simulation, we generated genotypes and phenotypes to obtain 500 cases and 500 controls and analyzed the simulated data using the approach described above.

This whole process was repeated 1000 times in order to obtain an accurate estimator of the corrected power, where the “corrected power” is estimated as the proportion of situations where the methods (including the “aggregated expert”) identify the correct combination. Table 11 illustrates the decision scheme using 6 individual methods and the aggregation method on the few first simulations.

|  | Causal SNP |  | Causal combination | Individual methods best combination |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Sim # | 1 | 2 |  | BHIT | KNN-MDR | MDR | BOOST | MSH | AES | Aggregated methods |
| 1 | 75 | 541 | 819 | 110 |  | 819 | 3474 |  | 819 | 819 |
| 2 | 587 | 811 | 4212 | 4212 | 4212 | 4212 | 4212 | 3161 | 4212 | 4212 |
| 3 | 845 | 964 | 4942 | 4942 | 4942 | 4942 | 4942 | 3943 | 3128 | 4942 |
| 4 | 488 | 626 | 3738 |  |  |  |  |  |  |  |
| 5 | 124 | 982 | 1308 | 1308 | 1308 | 1308 | 1308 | 1308 | 1308 | 1308 |
| 6 | 434 | 794 | 3490 | 2013 | 3490 | 3490 | 3490 | 1923 | 2060 | 3490 |
| 7 | 483 | 962 | 3772 |  |  |  |  |  |  |  |
| 8 | 229 | 460 | 2070 | 602 | 2070 | 241 | 803 | 164 | 218 |  |
| 9 | 268 | 348 | 2357 | 1245 |  |  | 2357 | 2357 | 104 | 2357 |
| 10 | 464 | 675 | 3640 | 3640 | 3640 | 2514 | 2752 | 3640 | 2487 | 3640 |
| Estimated corrected power |  |  |  |  | 0.60 | 0.50 | 0.50 | 0.30 | 0.30 | 0.70 |

Table 11 - A sketch of the results from ten simulations.

## Real data

Analyses using real data have been performed on a Rheumatoid arthritis (RA) genotype dataset involving 1999 cases and 1504 controls obtained from WTCCC (C 2007). Genotypes from the Affymetrix GeneChip 500K Mapping Array Set have been filtered using the usual quality controls tests on DNA quality (percentage of genotyped marker for any given individual above 90 %), markers quality (percentage of genotyped individuals for any given marker above 90 %), genotypes frequencies (markers with a p-value below a Bonferroni adjusted 5% threshold under the hypothesis of Hardy-Weinberg equilibrium in the controls cohort have been discarded). Missing genotypes for the GeneChip markers have been imputed using impute2 software (Howie et al. 2009). This procedure led to 312583 SNP to be analyzed for the 2 cohorts. Working with such a large panel remains quite challenging for several of the methods we have been using in this study. Therefore, we decided to reduce the number of SNP to about 52.000 by roughly considering the SNP with the highest MAF in each window of 6 successive SNP. Of course, in future studies, when more performant methods will be available (such as KNN-MDR, among others), the complete set of SNP could be considered again. Alternatively, after targeting some combinations with the reduced set of SNP, the discarded SNP could be reintroduced in order to refine the location of the combinations of interest.

Next, we used each method described above on this dataset as follows:

- MDR tested all combinations of 2 SNP (i.e. more than 1.350.000.000 combinations) and sorted the results by decreasing balanced accuracies. To obtain significance, we used a Bonferroni correction as is done in the MDR package: we kept the first 5000 highest balanced accuracy results, and used the corresponding 5000 combinations to perform the permutations. The number of permutations was conservatively based on the total number of tests, leading to a corrected p-value equal to 3.698225× $10^{-11}$. This necessitated to perform $10^{11}$ permutations.

- KNN-MDR has been used first on 1000-SNP windows, leading to 1326 tests involving 2 windows. Using an adaptative permutations scheme as is done in (Abo Alchamlat et al. 2017), and progressively decreasing the windows sizes, we ended up with a set of 33 windows containing 50 SNP each. Finally, a MDR approach was performed involving all combinations of 2 SNP from this set of 1650 SNP (i.e. 1.360.425 combinations).
- MegaSNPHunter has been used with the same parameters and using the same approach as has been done in a previous GWAS study (Wan et al. 2009), and the results have been sorted by decreasing $\chi^2$ values. To obtain significance, we performed a Bonferroni correction for the first 5000 results, similarly to what has been done for MDR.
- AntEpiSeeker has also been used with the same parameters and using the same approach as been done in a previous GWAS study (Wang et al. 2010), and the 5000 larger $\chi^2$ were kept to perform the simulations as done in MDR.
- BOOST has also been used with the same parameters and using the same approach as as been done in a previous GWAS study (Wan et al. 2010), with the results sorted by decreasing values of Kirkwood superposition approximation (KSA). To obtain significance, we performed a Bonferroni correction for the first 5000 results, and then used the same permutations approach as for the other methods.

# Results

## Results on simulated data

### Power

Figure 15 shows the estimations of the corrected power as a function of the simulation number. After a few hundreds simulations, the estimations stabilize and the relative ranking of the methods in terms of corrected power becomes fixed. The aggregation method is more powerful than any of the 6 other methods in our simulations. Another more detailed representation of the results is provided in Figure 16. Since the representation of more than 5 simultaneous methods is difficult and of no visual help, we have omitted the results involving MegaSNPHunter in the figure.

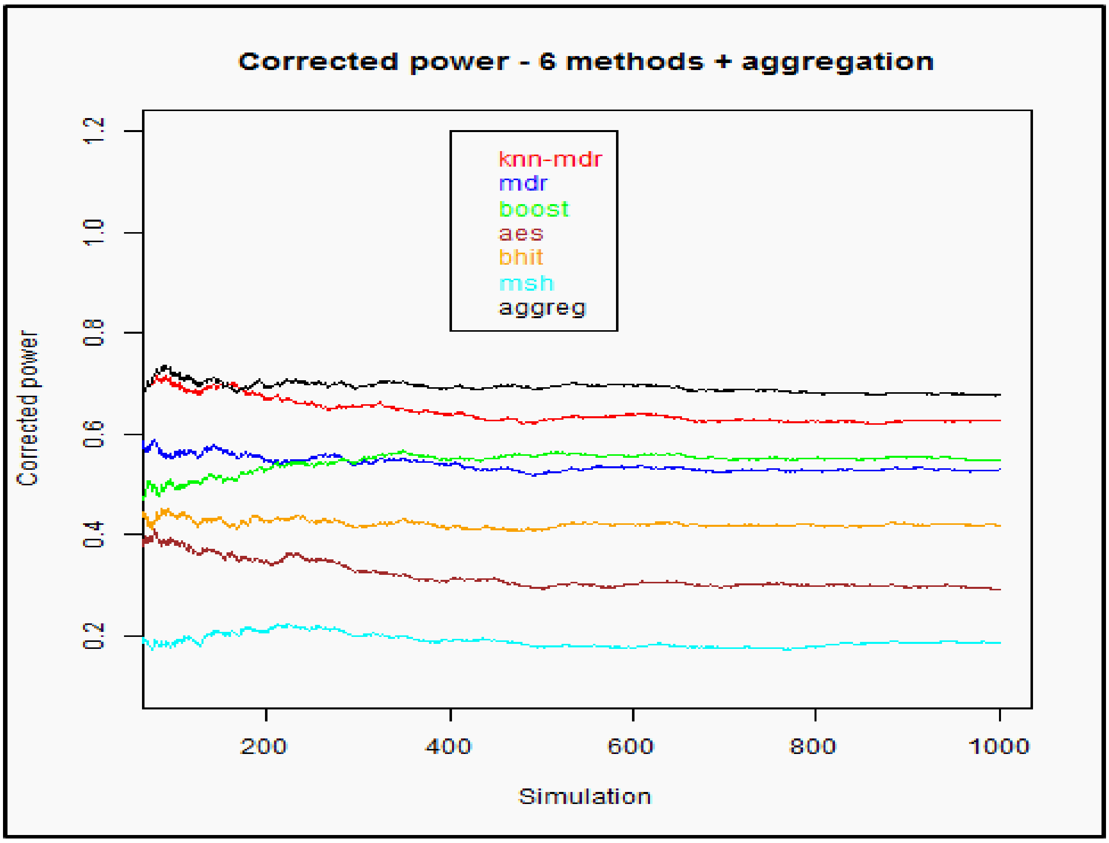

Figure 15 - Estimations of the corrected power for the 6 individual methods and the aggregation method.

In the setting used to obtain the Figure 16 results (i.e. using 5 individual methods), the highest empirical power (0.664) is obtained for the aggregation expert involving the 5 methods. The power of the individual methods used in the theoretical predictions obtained using (1) are the empirical powers of these methods, explaining why these are equivalent in the two graphs. It can also be observed that all powers of the aggregated methods involving only two methods are higher than expected. When three methods are involved, the powers are sometimes higher, sometimes lower than expected under independence. For four or five methods, the powers are constantly lower than expected, although higher than for any individual method when the five individual methods are aggregated (and even higher for six methods, 0.678, as mentioned on Figure 15).

Figure 6 17 shows the number of simulations (within the 1000 simulations) where only single method discovered the proper combination. Consequently, for these few (12 among the 1000 simulations) simulations, aggregation strategies performed less efficiently than stand-alone methods, especially KNN-MDR.

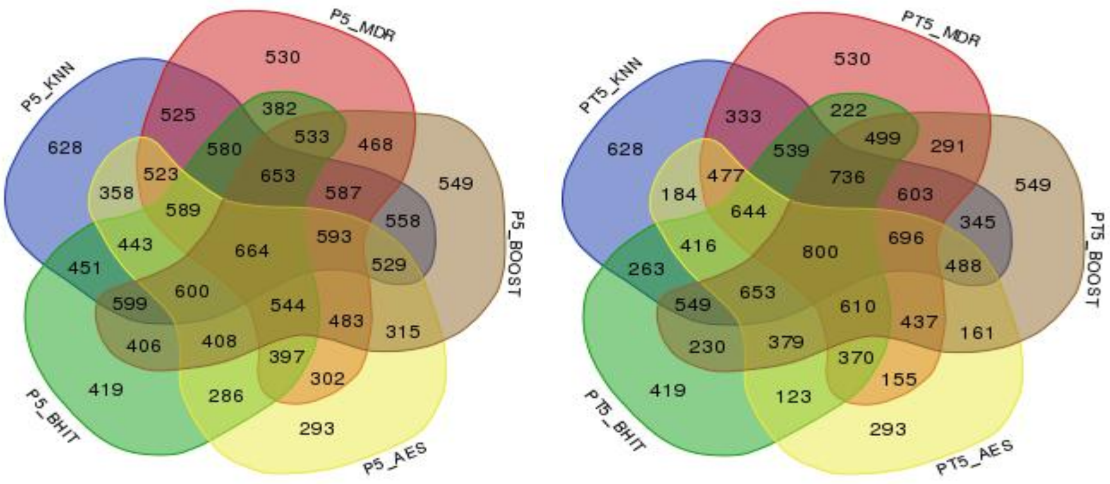

Figure 16 - Power (in ‰) of 5 individual methods (KNN, MDR, BOOST, AntEpiSeeker, BHIT) and of the 26 possible combinations of aggregated methods. The left diagram shows the results obtained in the simulations, while the right diagram shows the expected results under the hypothesis of methods independence (i.e. using formula (1) given above, where the Pi are the empirical powers of the individual methods. Note that the later does not necessarily correspond to a majority vote.

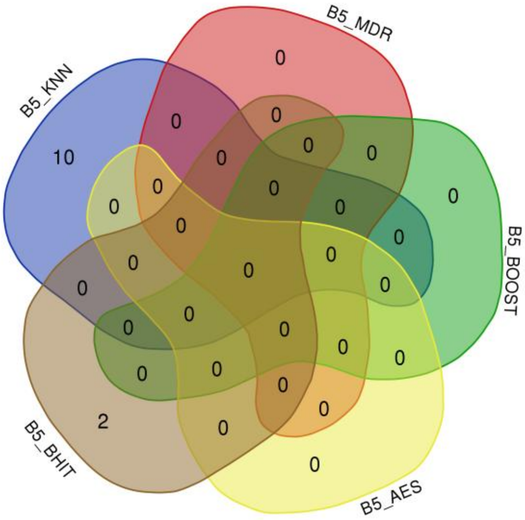

Figure 6  17  - Positive results (in 1000 simulations) obtained using single methods, but not detected using aggregation. . Only stand-alone KNN-MDR (for 10 simulations) and BHIT (for 2 simulations) led to discoveries that combinations could not detect.

## False positive rates

A second incentive for using aggregation is that false positive rates are likely to decline due to the use of a majority vote among parallel results: false positive results obtained using one method might not be obtained using a different method, with a different rationale. In our work, we have assessed two different kinds of false positive results:

- Either the methods identified an incorrect combination (note that these incorrect results are not included in the previous results on “corrected” power), generating an incorrect positive result.
- Either they identified a combination when no combination had been simulated (i.e. found a “false positive” result).

To test the first type of incorrect results, we have used the same set of simulations as for the power results and counted the number of incorrect positive results for each scenario. The combination identified as the most significant, if any, was taken as the solution for each of the methods, and the one with a majority vote, if any, for the aggregated method. The results are reported in Figure 18.

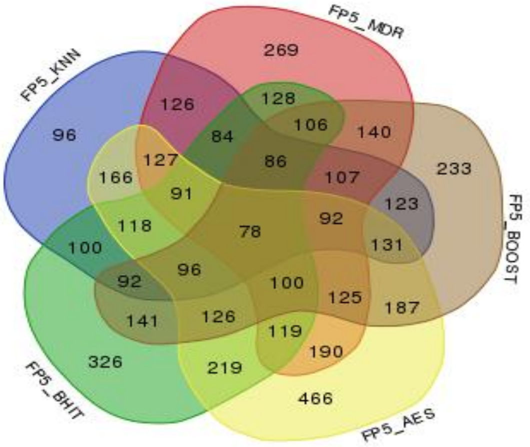

Figure 18 - Incorrect positive results in 1000 simulations at the 5% threshold.

To estimate the false positives rate, we have simulated 200 situations where no SNP was involved to generate the phenotype. Results are reported in figure 19.

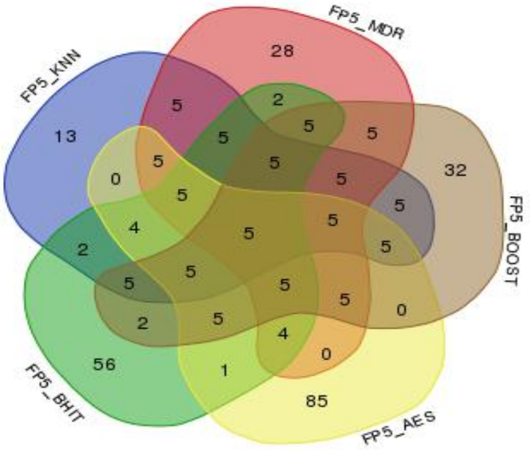

Figure 19 - Number of false positive results (significance threshold = 5%) in a set of 200 simulations.

A second set of 500 simulations has been carried out. In these analyses, we kept up to the 5 most significant combinations to see whether checking more than “the best” combination allows improving the (corrected) power without harming too much the false positives rate. Figures 20 and 21 present the results of these simulations.

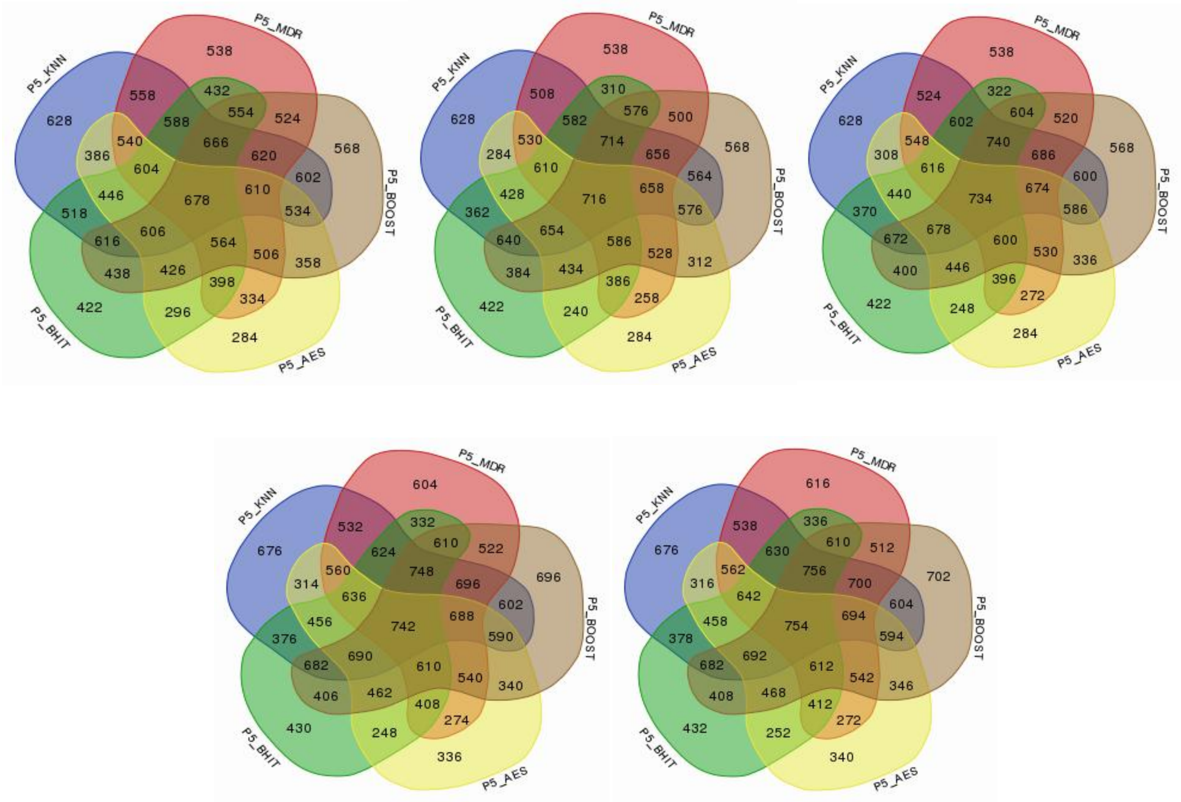

Figure 20 - Powers (in ‰) of 5 individual methods (KNN, MDR, BOOST, AntEpiSeeker, BHIT) and of the 26 possible combinations of aggregated methods when the number of kept significant combinations varies from 1 (top left) to 5 bottom right.

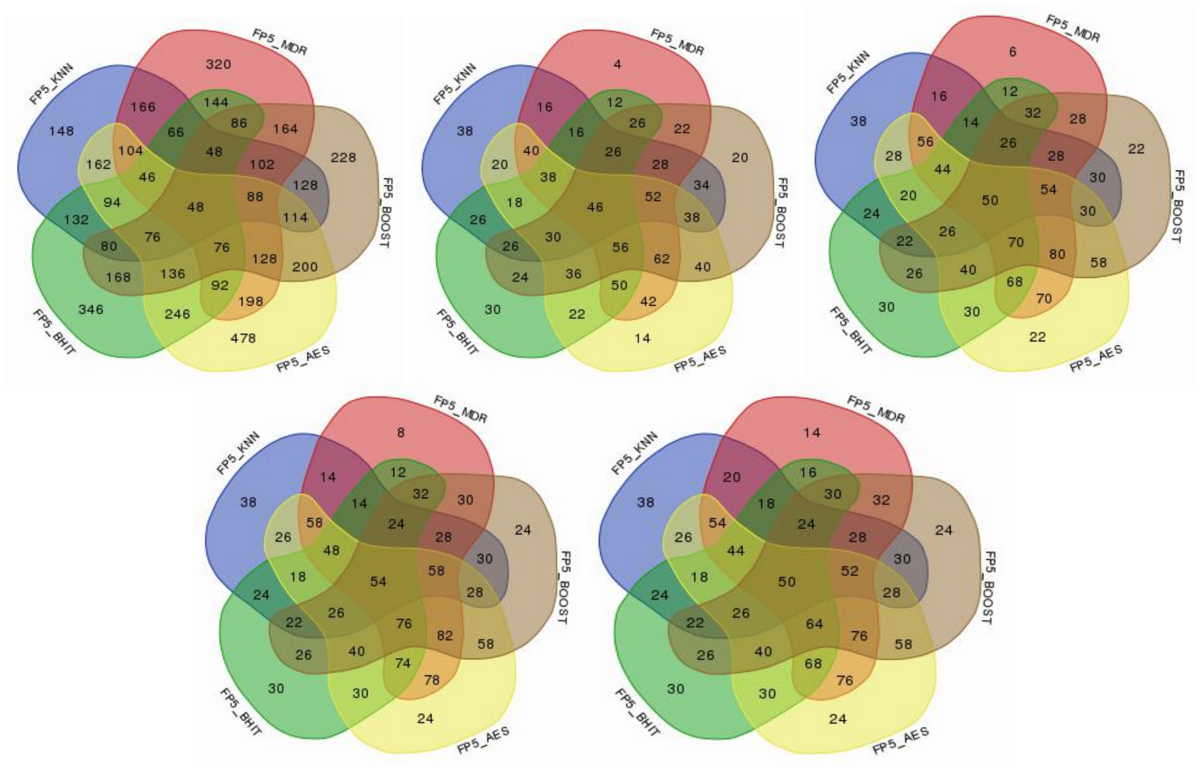

Figure 21 - False positive rates (in ‰) of 5 individual methods (KNN, MDR, BOOST, AntEpiSeeker, BHIT) and of the 26 possible combinations of aggregated methods when the number of kept significant combinations varies from 1 (top left) to 5 (bottom right).

## Correlation

Correlations between the methods results have been computed using the Cohen’s Kappa approach described above. The results are presented in Table 12. The correlations have been computed for each combination of 2 methods, and for 1 to 5 kept top-ranked combinations. We have assessed the significance of these measures by permuting 1000 times the results (success or failure) for each method and computing the corresponding values of kappa. For all combinations of methods and sets of combinations, no permuted kappa reached the value obtained with the real data, indicating that all p-values are lower than 0.001. Consequently, even when the methods show a slight agreement ($\kappa < 0.200$), the methods are very significantly correlated.

| Cohen κ | 1 combination | 2 combinations | 3 combinations | 4 combinations | 5 combinations |
|---|---|---|---|---|---|
| KNN-MDR | 0.587 | 0.537 | 0.518 | 0.515 | 0.511 |
| KNN-BOOST | 0.485 | 0.459 | 0.502 | 0.493 | 0.486 |
| KNN-AES | 0.244 | 0.230 | 0.246 | 0.247 | 0.245 |
| KNN-BHIT | 0.177 | 0.186 | 0.192 | 0.200 | 0.203 |
| KNN-MSH | 0.179 | 0.159 | 0.152 | 0.147 | 0.156 |
| MDR-BOOST | 0.608 | 0.512 | 0.497 | 0.507 | 0.469 |
| MDR-AES | 0.354 | 0.357 | 0.353 | 0.349 | 0.345 |
| MDR-BHIT | 0.238 | 0.255 | 0.263 | 0.270 | 0.275 |
| MDR-MSH | 0.208 | 0.174 | 0.156 | 0.155 | 0.170 |
| BOOST-AES | 0.343 | 0.287 | 0.277 | 0.287 | 0.283 |
| BOOST-BHIT | 0.319 | 0.308 | 0.287 | 0.298 | 0.299 |
| BOOST-MSH | 0.195 | 0.166 | 0.161 | 0.158 | 0.162 |
| AES-BHIT | 0.443 | 0.418 | 0.406 | 0.388 | 0.386 |
| AES-MSH | 0.468 | 0.454 | 0.415 | 0.415 | 0.431 |
| BHIT-MSH | 0.312 | 0.301 | 0.302 | 0.308 | 0.290 |

Table 12 - Cohen kappa coefficients for all combinations of methods using the approach given above. The last 5 columns correspond to situations where the number of considered regions for the majority vote increases from 1 to 5. All obtained correlations are highly significant ($p < 0.001$).

### Results on WTCCC data

Performing genome-wide interaction association studies with several methods on the RA dataset remains a challenge, even after pruning the dataset as described in a previous section. Each of the methods discovered a large number of potential interactions when using the 5% threshold and the correction procedures described in the Methods section (ranging from 1805 for MSH to 3808 for MDR). In total, 1306 significant 2-SNP interactions were discovered by at least 2 methods: 12 by the 5 methods, 19 by 4 methods (see Table 13), 476 by 3 methods and 799 by 2 methods only (see Supplementary material for a complete list). To obtain a ranked list of interactions, and although many sorting criteria could be used, we computed the rank of each interaction among the significant interactions of each method (the most significant interaction found using a given method was ranked 1 for that method, the second was ranked 2, etc. Interactions not present in the list of the given method were ranked (N+1), where N is the number of significant interactions for that method). We then summed up the ranks obtained by each significant pair of SNP and sorted the list according to this sum (the smallest sum corresponding to the “best” interaction). The results for the 31 interactions detected by at least 4 methods are reported in Table 13.

| Methods | | | | | Rankings | | | Interacting markers/genes, positions and references | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MDR | BOOST | AES | MSH | KNN | R1 | R2 | R3 | SNP1 | Position 1 | Ref 1 | Gene 1 | SNP 2 | Position 2 | Ref 2 | Gene 2 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 1 | 53 | 66 | rs12233352 | 22:23273614 | (Reif et al. 2014) | BCR | rs1573767 | 22:26689528 | (Yamamoto et al. 2015) | MIATNB |
| ▨ | ▨ | ▨ | ▨ | ▨ | 2 | 71 | 85 | rs7747393 | 6:108580936 | (Lin et al. 2008) | FOXO3 | rs9285397 | 6:108630452 | (Lin et al. 2008) | FOXO3 |
| ▨ | ▨ | | ▨ | ▨ | 3 | 1 | 22 | rs9836309 | 3:122416348 | | AC083798.2 | rs735728 | 3:122804327 | | DIRC2 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 4 | 425 | 161 | rs759635 | 19:3351892 | | AC007792.1 | rs12151209 | 19:3374122 | (Julia et al. 2012) | NFIC |
| ▨ | ▨ | | ▨ | ▨ | 5 | 141 | 69 | rs1591217 | 1:242764068 | | / | rs4658626 | 1:244345804 | | / |
| ▨ | | ▨ | ▨ | ▨ | 6 | 28 | 43 | rs13032261 | 2:226403875 | (Wan et al. 2009) | / | rs906428 | 15:80850733 | (Wan et al. 2009) | CEMIP |
| ▨ | | ▨ | ▨ | ▨ | 7 | 89 | 48 | rs7172832 | 15:70393164 | (Wan et al. 2009) | / | rs906428 | 15:80850733 | (Wan et al. 2009) | CEMIP |
| ▨ | ▨ | ▨ | ▨ | ▨ | 8 | 129 | 188 | rs4077806 | 2:142449600 | (Lin et al. 2008) | / | rs10861940 | 12:108540404 | (Lin et al. 2008) | SART3 |
| ▨ | ▨ | ▨ | | ▨ | 9 | 340 | 112 | rs1169565 | 2:71196518 | (Abo alchamlat et al. 2017) | PAIP2B | rs571307 | 13:29942173 | (Abo alchamlat et al. 2017) | LINC00544 |
| ▨ | | ▨ | ▨ | ▨ | 10 | 320 | 104 | rs6586516 | 1:17299036 | (Jung et al. 2009) | / | rs3761847 | 9:120927961 | (Jung et al. 2009, Han et al. 2009) | TRAF1 |
| ▨ | ▨ | ▨ | | ▨ | 11 | 187 | 160 | rs1356661 | 3:140856737 | | / | rs2382297 | 9:11528905 | | / |
| ▨ | ▨ | ▨ | | ▨ | 12 | 406 | 163 | rs4979291 | 9:107732763 | (Abo alchamlat et al. 2017) | / | rs10979420 | 9:108634242 | (Abo alchamlat et al. 2017) | / |
| ▨ | ▨ | ▨ | | ▨ | 13 | 423 | 190 | rs17281652 | 15:79759951 | | AC026826.2 | rs10906980 | 15:83728410 | | ADAMTSL3 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 16 | 709 | 322 | rs10422388 | 19:3393466 | (Abo alchamlat et al. 2017) | NFIC | rs778982 | 19:5866574 | (Julia et al. 2012) | FUT5 |
| ▨ | | ▨ | ▨ | ▨ | 24 | 74 | 210 | rs4077806 | 2:142449600 | (Lin et al. 2008) | / | rs3908728 | 10:45043455 | | CEP164P1 |
| ▨ | | ▨ | ▨ | ▨ | 50 | 12 | 299 | rs4077806 | 2:142449600 | (Lin et al. 2008) | / | rs4729973 | 7:104230339 | | / |
| ▨ | | ▨ | ▨ | ▨ | 70 | 76 | 340 | rs4077806 | 2:142449600 | (Lin et al. 2008) | / | rs9369273 | 6:41346376 | (Lin et al. 2008) | NCR2 |
| ▨ | | ▨ | ▨ | ▨ | 93 | 48 | 351 | rs4077806 | 2:142449600 | (Lin et al. 2008) | / | rs11553287 | 16:67871504 | | NUTF2 |
| ▨ | | ▨ | ▨ | ▨ | 163 | 459 | 385 | rs10861940 | 12:108540404 | (Lin et al. 2008) | SART3 | rs8094633 | 18:5349101 | | / |
| ▨ | ▨ | ▨ | ▨ | ▨ | 262 | 901 | 614 | rs1356661 | 3:140856737 | | / | rs2069119 | 8:51773876 | | PXDNL |
| ▨ | ▨ | ▨ | ▨ | ▨ | 298 | 1860 | 633 | rs12493613 | 3:180410500 | | LINC02053 | rs6112264 | 20:19231501 | | SLC24A3 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 348 | 741 | 674 | rs6426569 | 1:227145048 | | CDC42BPA | rs4077806 | 2:142449600 | (Lin et al. 2008) | / |
| ▨ | | ▨ | ▨ | ▨ | 512 | 977 | 661 | rs4776947 | 15:67618350 | | MAP2K5 | rs4965761 | 15:98315690 | | AC015722.2 |
| ▨ | | ▨ | ▨ | ▨ | 514 | 1071 | 662 | rs10209628 | 2:121608436 | | CLASP1 | rs4965761 | 15:98315690 | | AC015722.2 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 522 | 2058 | 776 | rs2313132 | 4:137771096 | (Graham et al. 2009) | / | rs2069119 | 8:51773876 | | PXDNL |
| ▨ | ▨ | ▨ | ▨ | ▨ | 752 | 2097 | 896 | rs2452578 | 4:94552011 | | PDLIM5 | rs17382362 | 20:8913603 | | PLCB1 |
| ▨ | ▨ | | ▨ | ▨ | 810 | 2050 | 915 | rs878249 | 20:1058648 | | / | rs6114758 | 20:2524478 | | ZNF343 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 919 | 2039 | 962 | rs4776947 | 15:67618350 | | MAP2K5 | rs2829098 | 21:24464346 | | LINC01684 |
| ▨ | ▨ | ▨ | ▨ | ▨ | 1006 | 2080 | 1010 | rs2873522 | 20:1438882 | | NSFL1C | rs6112264 | 20:19231501 | | SLC24A3 |
| ▨ | | ▨ | ▨ | ▨ | 1064 | 2112 | 991 | rs4729973 | 7:104230339 | | / | rs7975146 | 12:128146076 | | / |
| ▨ | | ▨ | ▨ | ▨ | 1263 | 2104 | 1142 | rs7975146 | 12:128146076 | | / | rs11553287 | 16:67871504 | | NUTF2 |

Table 13 - List of the significantly interacting SNP in the WTCCC RA data. (The 31 2-SNP interactions found significant by at least 4 methods. The corresponding SNP names and positions are also reported. Ref 1 and Ref 2 (when any) refer to previous studies where the corresponding SNP were already reported. Gene 1 and Gene 2 (when any) are reported when the corresponding SNP are located

in a gene (in intronic, exonic or UTR regions). The methods for which the SNP were reported as significant are indicated by a colored cell. Furthermore, 3 rankings are also reported: the first one is the one used to rank the interactions in the Table and is described in the text. The second is the balanced accuracy computed by KNN-MDR. The third one is the rank of the average rank of the interaction computed over the methods for which this interaction was significant. )

In total, the 31 2-SNP interactions detected by at least 4 methods involve 47 distinct SNP (36 SNP are involved in only one interaction, 10 are involved twice and 1 is present in 6 interactions, see Table 13). Some interactions (12 out of 31) involve SNP on the same chromosome, while 19 involve SNP on distinct chromosomes. For intra-chromosomal interactions, the distance between the SNP ranged from very small (2 are smaller than 50kb), to very large (2 are larger than 10 Mb). This shows that the methods potentially reported interactions involving close windows, such as upstream regulatory regions of genes, as well as much more distant ones, including combinations of windows located on different chromosomes. Several of these interactions have already been reported in previous analyses (see Table 13), while others are new, to our knowledge (for example on chromosome 3), or might potentially be echoes of other more significant ones.

Figure 22 provides another view of the results from this analysis (a supplementary file gives a more complete version of the results). On this figure, chromosomes are reported with a dimension approximatively proportional to their physical size, interacting sites are signaled through dashes corresponding to the location of the interacting SNP on the chromosome and the detected inter-chromosomal interactions are reported using the dashed lines within the circle.

## Discussion

The detection of genetic interactions is a notoriously difficult task, and, although numerous papers have been published in the field, a lot of work remains to be done to propose methodological advances allowing obtaining reliable and reproducible significant results in many gene-mapping studies. Our work aims to be a step in that direction.

A first difficulty is the statistical power issue to detect epistatic interactions: even if epistatic effects are not necessarily more tenuous than main effects, the number of tested hypotheses increases at least quadratically, making multiple testing corrections potentially more penalizing. Therefore, strategies allowing obtaining reasonable power in such studies are desirable. This is one of the features of the approach we propose in this paper. As shown in the Materials and Methods and the Results sections, aggregation strategies provide some potential increases in the detection power. Even if power increases in the current study were rather modest, it has been shown that adding more methods in the aggregation has the potential to increase the overall power. In our study, the theoretical expectations are supported by the simulation results (e.g. Figure 16), with an improved power of the method aggregating the results of the 5 underlying methods with respect to the individual methods and to the methods aggregating less methods, although admittedly smaller than expected under the hypothesis of methods independence. Note that the property of independence mentioned here means that the probability of finding a positive result for one method does not depend on the findings of another method: although this might be arguable for ‘easy to find’ interactions, this might be more plausible for less ‘visible’ interactions, especially when distinct methods rest on very different approaches. Nevertheless, in our study, although we have used methods covering various methodologies (multi-dimensional reduction (MDR, KNN-MDR), exhaustive search (BOOST), empirical (AntEpiSeeker) and Bayesian (BHIT) approaches) and potentially hard to find interactions (small marginal effects, potentially heterogeneous situations, see the simulations description), we obtained significant correlations between the methods results (Table 12). Although the way these correlations affect the power is not clear, the global effect was a reduction of the obtained power compared to the expectation.

Reproducibility is another problem in mapping studies. At least three reasons are at the root of this problem: a first reason is that many published results are probably false positives, partly due to improper correction for multiple testing. Another reason is that not every method is equally likely to detect any type of interaction, making detection not only a function of the variants to be highlighted, but also of the used method. And finally, and more fundamentally, it is to be expected that many phenotypes are under the control of many genes with intricate interaction networks. Consequently, involved interactions in one dataset, or even in subsets, might differ, increasing the heterogeneity of the underlying genetics and making detection of these interactions more complicated. Our approach is of interest for the two first problems. We have indeed shown that aggregating the positive results of various methods helps to control the false positive rates: false positives produced by one of the methods are not necessarily produced by the other used methods, and so will most of the time be discarded from the final results (Figures 18 and 19). On the other hand, positive results produced by a majority of methods - where the way this majority is defined is important - will pop us, allowing combining the detection skills of several methods rather than only considering individual methods results.

Aside of these interesting properties, some difficulties have to be mentioned. An obvious disadvantage of the aggregation strategy is that several methods have to be mastered, installed on the computer facilities and run. This of courses increases the total computing time, which might be an issue when large datasets are considered. A possible solution would be to use the nowadays largely available parallel resources offered to most research centers: using several nodes to perform the tasks (run the programs implementing the various strategies, run the permutations when needed, etc.) should lead to a non-significant increase of the total observed run time, at the cost of the software implementation of this parallelization strategy.

Another difficulty is the aggregation itself: the ranking of the interesting interactions is performed based on their significance. This leads to at least two problems: first, providing a clear ranking might be difficult; for example, when permutations are used, several interactions might easily end up with the same significance, making subsequent ranking almost arbitrary. Next, even when an objective ranking has been obtained, the most interesting might not necessarily be the best-ranked ones. Although this point clearly deserves more investigations, one possible crude approach, used in this study, was to incorporate more than the top-ranked combinations in the aggregation. Figures 20 and 21 show that this simple strategy has some merits, increasing the power while still controlling for the number of false positives when the number of kept top combinations increases from 1 to 5.

In view of the main characteristics of our strategy, it was important to test the approach on real datasets to check whether new clues could be obtained from our analyses. Our results on the WTCCC Rheumatoid Arthritis data provides new information on potential new candidate regions. As shown in Table 13, several previously reported associations and interactions are also found in our study.

Furthermore, interactions between previously identified genes and other genes or regulatory regions are also pointed out, which can possibly provide new and useful information on the molecular mechanisms leading to RA. Finally, entirely new interactions are also found significant in our study, which might point to new target genes to be investigated in future RA studies, although their biological relevance is obviously not clear at this stage.

Figure 22 provides a view of the significant results at the chromosome level. This figure and Table 13 show that some interactions involve SNP on 2 distinct chromosomes while other involve (sometimes closely) linked SNP. Although this might make biological sense (for example, regulatory regions might be close to the genes they influence), a potential bias of our method has to be mentioned. Indeed, BOOST tends to detect much more internal interactions than interactions between different chromosomes segments (Wan et al. 2010). Consequently, adding other (maybe less biased?) methods and/or somehow relaxing the unanimity vote criterion might allow to uncover (and maybe also exclude) other combinations.

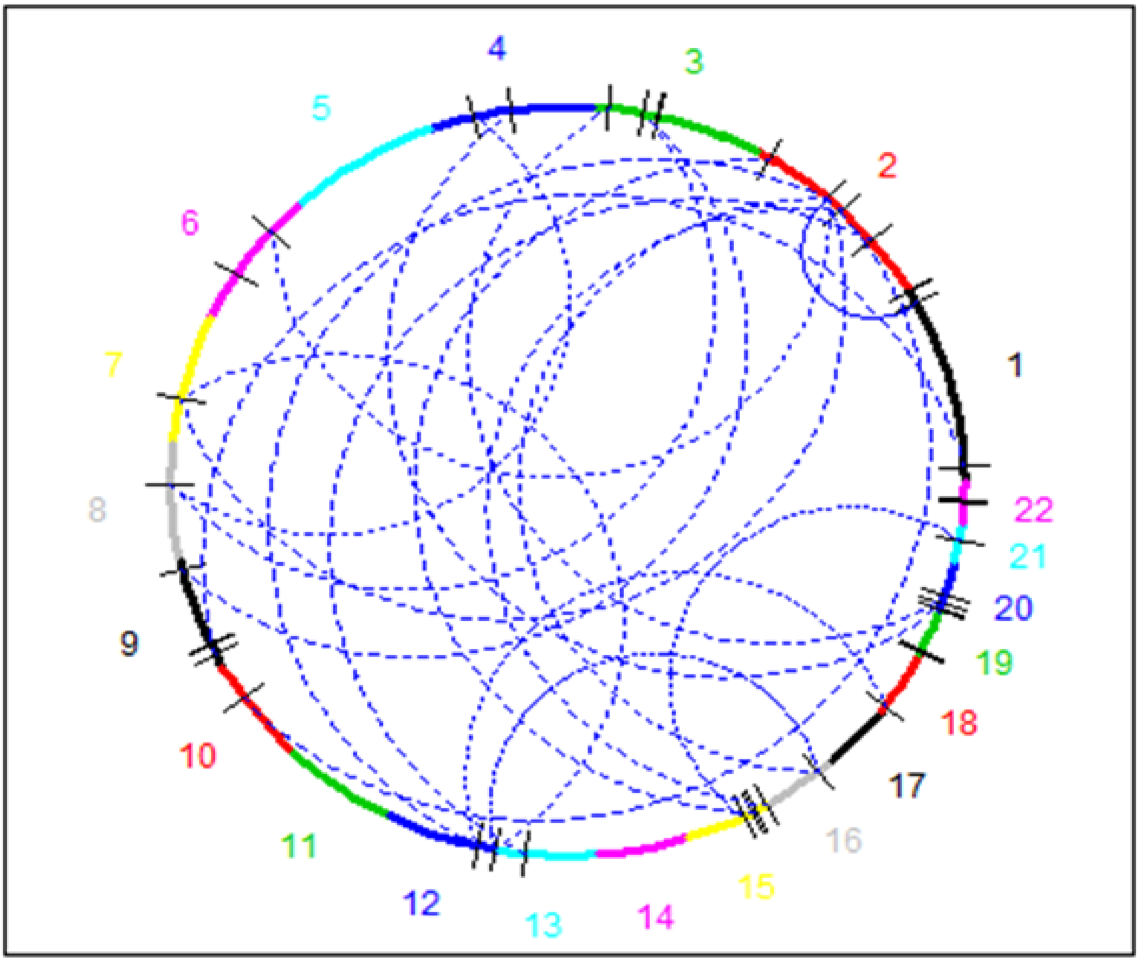

Figure 22 - Results of the analysis on WTCCC data. The figure shows the chromosomes that are involved in interactions, and the approximate location of the involved regions on these chromomses.

## Conclusions

In summary, the aggregation of methods is an approach with interesting features for detecting epistatic interactions. Integrating the results of parallel methods has been shown to increase the corrected power over the power of the individual methods while controlling the false positives rate. The feasibility of using such methodology on real genome-wide datasets has also been demonstrated, providing potential new insights in complex traits analyses.

# *Abbreviations*

**GWAS:** Genome-wide association study

**MDR:** Multifactor dimensionality reduction

**BOOST:** Boolean operation-based screening

**LD:** Linkage disequilibrium

**FAM-MDR:** Flexible family-based multifactor dimensionality reduction

**MB-MDR:** Model-based multifactor dimensionality reduction

**GMDR:** Generalized multifactor dimensionality reduction

**WTCCC:** Wellcome trust case control consortium

**BHIT**: Bayesian High-order Interaction Toolkit

**KNN:** K-nearest neighbors

**SNP:** Single-nucleotide polymorphism

**MAF:** Minor allele frequency

**HSA9:** Human chromosome 9

# *Acknowledgements*

None.

## Funding

No funding was provided for this study.

## Availability of data and software

The software of the KNN-MDR program is available on the following URL:
http://www.fmv.ulg.ac.be/cms/c_1802261/fr/publiclyavailable-softwares
The results on simulations data will be made available as Additional files 9, 10, 11 on the paper's review website. This study makes use of data generated by the Wellcome Trust Case-control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. All data from WTCC have been used according to the terms of the WTCCC data access agreement. Relevant publication for the used dataset is (Wellcome Trust Case Control Consortium, 2007): Nature 2007;447;7145;661–78 (PUBMED: 17554300; PMC: 2719288; doi:10.1038/nature05911). Access to the data from WTCCC needs to be obtained from the Consortium.

## Authors' contributions

AS participated in the design of the study and performed the statistical analysis and helped to draft the manuscript. FF conceived the study and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Consent for publication

Not applicable

### Ethics and consent to participate

Not applicable

### Open Access

## References

Abo Alchamlat, S. and F. Farnir (2017). "KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies." BMC Bioinformatics 18(1).

Bashinskaya, V. V., O. G. Kulakova, A. N. Boyko, A. V. Favorov and O. O. Favorova (2015). "A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches." Human Genetics 134(11-12): 1143-1162.

Breiman, L. (2001). Random Forests. Machine Learning. E. R. Schapire. University of California.

C, W. T. C. C. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145).

Can, Y., H. Zengyou Hkust, W. Xiang , Y. QiangView, X. HongView and Y. WeichuanView (2009). "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies." Bioinformatics 25(4).

Chen, L., G. Yu, C. D. Langefeld, D. J. Miller, R. T. Guy, J. Raghuram, X. Yuan, D. M. Herrington and Y. Wang (2011). "Comparative analysis of methods for detecting interacting loci." BMC Genomics 12: 344.

Choi, Y. and I. Doh (2014). "Hierarchical Aggregation of Uncertain Sensor Data for M2M Wireless Sensor Network Using Reinforcement Learning." International Journal of Distributed Sensor Networks 2014: 1-9.

Dietteric, G. H. (2000). Ensemble Methods in Machine Learning. J. Kittler and F. Roli. Cagliari, Italy. 1857.

Gerardi, D., R. McLean and A. Postlewaite (2009). "Aggregation of expert opinions." Games and Economic Behavior 65(2): 339-371.

Gori, A., E. Théâtre, B. Charloteaux, Y. Momozawa, V. Deffontaine, D. Baurain, M. Mni, F. Crins, N. Ahariz and C. Oury (In Press). "Fine-mapping and functional analysis of the 5p13.1 risk locus for Crohn’s disease." Am J Hum Genet.

Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, Burtt NP, Guiducci C, Parkin M, Gates C (2008). "Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus." Nature Genetics 40(9):1059-1061.

Han T-U, Bang S-Y, Kang C, Bae S-C (2009). "TRAF1polymorphisms associated with rheumatoid arthritis susceptibility in Asians and in Caucasians." Arthritis Rheum 60(9):2577-2584.

Howie, B., P. Donnelly and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet 5(6).

JYoun, G. H. and E. M. Marcotte (2016). "Predictability of Genetic Interactions from Functional Gene Modules." bioRxiv.

Julia Cano A, Marsal Barr S (2012). "Vitro method for the prognosis or prediction of the response in patients with rheumatoid arthritis treated with agents that recognize the cd20 membrane receptor in b lymphocytes." Patent Application Publication 13(380).

Jung J, Song J, Kwon D (2009). "Allelic based gene-gene interactions in rheumatoid arthritis." BMC Proceedings 3(Suppl 7):S76.

Lin H-Y, Desmond R, Louis Bridges S, Soong S-j (2008) "Variable selection in logistic regression for detecting SNP–SNP interactions: the rheumatoid arthritis example." Eur J Hum Genet 16(6):735-741.

Lon, C., R. L. and I. J. Bell (2001). "Association study designs for complex diseases." Nature Reviews Genetics 2: 91-99.

MAK, B., T. BuI and R. BLANNING (1996). "Aggregating and updating experts' knowledge: an experimental evaluation of Five classification techniques." Expert Systems with Applications 10(2).

McHugh, M. L. (2012). "Interrater reliability: the kappa statistic." Biochemia Medica 22(3): 276-282.

Paixão, T. and N. Barton (2016). "The effect of gene interactions on the long-term response to selection." Proc Natl Acad Sci U S A 113(16).

Phillips, P. C. (2008). "Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems." Nat Rev Genet 9(11): 855-867.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." Am J Hum Genet 81(3): 559-575.

Reif DM, White BC, Moore JH (2014). "Integrated analysis of genetic, genomic and proteomic data." Expert Rev Proteomics 1(1):67-75.

Ritchie, M., W. Hahn, N. Roodi, L. Bailey, D. Dupont, F. Parl and H. Moore (2001). "Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer." Am J Hum Genet 69(1).

Stranger, B. E., E. A. Stahl and T. Raj (2010). "Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics." Genetics 187(2): 367-383.

Titov, I., A. Klementiev, K. Small and D. Roth (2010). "Unsupervised Aggregation for Classification Problems with Large Numbers of Categories." Journal of Machine Learning Research 9: 836-843.

Tsyganok, V. (2010). "Investigation of the aggregation effectiveness of expert estimates obtained by the pairwise comparison method." Mathematical and Computer Modelling 52(3-4): 538-544.

Ulfarsson, M. O., F. Palsson and J. Sigurdsson (2016). "Classification of Big Data with Application to Imaging Genetics." Browse Journals & Magazines 104(11).

Wan, X., C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang and W. Yu (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." Am J Hum Genet 87(3): 325-340.

Wan, X., C. Yang, Q. Yang, H. Xue, N. L. Tang and W. Yu (2009). "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study." BMC Bioinformatics 10: 13.

Wang, J., T. Joshi, B. Valliyodan, H. Shi, Y. Liang, H. T. Nguyen, J. Zhang and D. Xu (2015). "A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies." BMC Genomics 16: 1011.

Wang, Y., X. Liu, K. Robbins and R. Rekaya (2010). "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm." BMC Res Notes 3: 117.

Yamamoto K, Okada Y, Suzuki A, Kochi Y (2015) "Genetic studies of rheumatoid arthritis" Proc Jpn Acad Ser B Phys Biol Sci 91(8):410-422.

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

## *Additional files*

**Additional file 9** – The results of 6 individual methods and the aggregation method on the 1000 simulations by taking the first significant combination.

**Additional file 10 –** The results of 6 individual methods and the aggregation method on the 500 simulations by taking the first 5 significant combinations.

**Additional file 11 –** List of the significantly interacting SNP in the WTCCC RA data for the results found by every  method.

# Discussion - Perspectives

## General Discussion

The two main themes of this thesis are the discovery of genetic interactions and the use of non-parametric statistics. The main objective was therefore the development of statistical non-parametric procedures allowing a better detection of the relationship between genomic configurations and phenotypic expressions by taking into account interaction effects while keeping power high by controlling for the complexity of the model.

Although other approaches exist - such as genomic selection, where the impact of genomic sections on the traits of interest is inferred with no particular interest for the causal genes and mutations, or classification methods where the goal is to obtain the more likely status of the tested individuals using similarities with other individuals with a known status and with no specific interest to the regions creating the similarity - we have considered that one of the major goals in genetic studies is the identification and characterization of the genetic interactions contributing to complex traits. Above the fundamental importance of such phenomenon to give clues to the genomic dissection of the traits, applied aspects, such as identification of pharmaceutical targets, should also be considered and generate interest in such techniques. Revealing each variant of importance, either directly or through interactions with other variants or with the environment, seems nowadays an unreachable goal in many cases. Nevertheless, untangling some of the main effects underlying traits of interest might be important for some of these. Furthermore, the corresponding reduction due to new discoveries might lead to an increase in the power to find other ones. In other words, although explaining completely the heritability (in the large) of a trait will be most of the time impossible, it can be hoped that techniques such as the ones presented in this thesis will contribute to a better knowledge in the genetics of various traits. One of the challenges in the epistasis mapping methods is linked to the “curse of dimensionality” problem: revealing interactions involving more than 2 or 3 variants might quickly become infeasible in terms of computing costs, but also in terms of size of the needed cohorts: indeed, the exponential increase in the number of tests to be realized will result into (corrected) thresholds so low that either the detected effects or the cohorts sizes will need to be huge. The detection of multi-ways (i.e. more than 2-ways) interactions has only been slightly studied in this thesis when providing a few results on KNN-MDR. Due to the difficulties cited above, not so many methods are available to tackle this very difficult problem. Large dimensionality problems are often attacked using Bayesian problems and we have grossly described such an approach (BHIT = Bayesian High-order Interaction Toolkit) in the introduction of this document. Although the performances of the program (provided as an add-on with the paper) on our simulated dataset have been disappointing, the used approach might potentially be reexamined and maybe improved to provide new insights in problems where larger dimensionality is needed.

In our work, we have been using methods covering various methodologies (multi-dimensional reduction (MDR, KNN-MDR), exhaustive search (BOOST), empirical (AntEpiSeeker) and Bayesian

(BHIT) approaches), which have been proposed for detecting genetic interactions. Each methodology has its own merits and pitfalls. In the next section, we will discuss more systemically some important features and the contributions of our work to potential improvements for each of these aspects.

## Power

The increase in power is one of the most important goals of all methods in this field. From this point of view, we note that KNN-MDR outperforms the other methods in situations that might be of interest to unravel new regions not previously detected. The reasons underlying that hope are that the method takes into account linkage disequilibrium (LD) - which many methods do not - and is able to detect interactions between SNPs even in the absence of marginal effects, which should provide benefits over the methods using these previously detected marginal effects as a basis for epistasis detection.

The other methodology proposed in this work, the aggregation strategy, could also increase the power more than all the individual methods, although the power increases in this study were rather modest. One reason for the observed small gains in power is that the number of methods we used in our "proof of concept" study was rather modest, but the theoretical development made in the paper suggests that adding more methods should increase the advantage in terms of power. Admittedly, the rather naïve rationale used to provide an idea on the gains in power that could be expected might lead to an overestimation of the gains, but gains could nevertheless be achieved, making this approach potentially useful. A better estimation of the expected gains might maybe be better achieved by considering the correlation between the involved methods, a line of research that has been alluded to in the second paper, but that might deserve more investigations.

## False positive rates

The false positive rate in KNN-MDR has been found to be less than the one obtained using other methods in our study. One obvious reason for that is the way we have performed the simulations. Indeed, we have engaged to have epistatic interactions with little marginal effects in order to avoid the trivial situations where individual loci can be identified in a first step, followed by the identification of interactions between the loci in a second step. To obtain these situations with no (or little) marginal effects, we have used multi-locus prevalences. Furthermore, we have introduced some kind of genetic heterogeneity: a same multi-locus genotype could simultaneously be present in cases and in controls, making it harder to identify these loci. These two features - little marginal effect and heterogeneity - greatly contribute to difficulties in mapping and are quite harmful for several methods. This is less the case for KNN-MDR partly because the use of a classification-like distance measure incorporating information from linkage disequilibrium. This procedure tends to discard spurious stochastic associations.

The nature of the good performances with regard to the false positive rate with the aggregation strategy is different: the rationale here is that differing methods could produce different spurious results. Consequently, a majority vote among the participating methods will lead to rejection of many spurious results, while true positives, hopefully detected by several methods, will be highlighted.

### Number of interacting loci

We have already mentioned some difficulties linked to this feature in this discussion. Some more points can be added. First, as mentioned, methods such as MDR will have trouble to manage more than 2 or 3 interacting loci because of the high dimensionality and the corresponding excessive computing costs. But other difficulties might also appear: we mentioned, when describing the differences between MDR and KNN-MDR, that one of the advantages of KNN-MDR over MDR was to avoid the presence of empty (or with scarce observations) cells. This is ever more the case with more loci, and could lead to difficulties in the status attribution (no status can be given when no similar condition has been met in the learning sample, and status allocation might be unstable when very few observations are present in the concerned cell). Although KNN-MDR can be used with more than 3 interacting loci, only 3 markers have been used in this work. Nevertheless, in practical applications, it is not unlikely that situations involving more than 3 loci might exist. These situations might increase the interest of using methods such as KNN-MDR. Indeed, when more regions are involved in the phenotype, this could decrease the distance measure between individuals sharing some or all of these regions and better cluster individuals sharing the same status. Of course, increasing the number of loci will increase the computing load and strategies, similarly to what has been done in the KNN-MDR paper to work on the real dataset, will need to be designed. For the aggregation strategy, more than 2 interacting loci could be used, but the difficulty would arise from the need to use methods able to cope with the chosen number of loci. Although some methods are already available (KNN-MDR, BHIT, ...), this might be an issue since the performances have been shown to improve when the number of methods is higher.

### Computer resources

Computer time issues are scarcely discussed in this thesis. One obvious reason for the aggregation strategy is that parallel run of the various methods leads to a run-time equal to the run-time of the slowest of the used methods (except for the majority vote, whose run-time is negligible when compared to the individual methods). Consequently, better performances would necessitate improvements in the individual methods, which is clearly not the focus in this work (in the tested situations, several methods turned out to be slower than KNN-MDR).

The obvious parallel nature of the aggregation method makes it suitable for an easy use on today widely spread parallel computers, which is fine (although not required). Note that “easy use” might be overstated: including more methods in the analyses necessitates proper installation and use of a (if possible) large panel of methods, which is not necessarily an easy task.

KNN-MDR is more computationally intensive than the majority vote in the subset sharing the same multi-locus genotype used in MDR. Computation of distance matrices might be costly and increases quadratically with the number of individuals. The relationship between the computing time and the markers number and positions is less clear and depends on population and map-dependent features. As mentioned, windows sizes should be adapted to capture linkage disequilibrium while avoiding taking too large windows in order to minimize the introduced noise. Parallelization of the permutations could also be used to decrease the total run time, as already suggested in the paper.

### Real data

In view of the main characteristics of our work, it was important to test our approaches on real datasets to check whether new clues could be obtained from our analyses. Our results on the WTCCC Rheumatoid Arthritis data provides some results consistent with other results in the domain of Rheumatoid Arthritis and new information on potential new candidate regions, which might point to new target genes to be investigated in future Rheumatoid Arthritis studies, although their biological relevance is obviously not clear at this stage and needs more investigations in the future. Beyond these new results, we have presented a possible strategy to tackle large dataset using the tools we propose. For KNN-MDR, the strategy allows obtaining interesting results and could be applied similarly to other datasets. The problem is of course different for the aggregation strategy, where the limitations come from the used methods rather than from the aggregation strategy itself. Furthermore, the strategy we propose is not the only possible one and other approaches could be used, such as firstly filtering the dataset using one method, followed in a second step by the use of the aggregation strategy on the filtered data. Of course, issues on the choice of the method to be used first and on the filtering strategy have to be addressed before using such an approach.

## Perspectives

This thesis has attempted to provide new approaches for interactions mapping using non-parametric techniques. We are aware that the used approaches could be improved, and that other (parametric and non-parametric) techniques could have been used. This small paragraph lists a few options in these directions.

A first idea would be to write a modified version of KNN-MDR to make it more user-friendly and, maybe, more efficient. Indeed, tuning the programs - i.e. using the best set of parameters - is not an obvious task and can affect the performances (power, false positive rate, speed, ...) of the method. A step in that direction would be to provide a companion tool allowing a definition of the markers windows. Indeed, these windows should capture linkage disequilibrium information, and an optimal size could be obtained from the data. Note that the size depends on various factors: globally, the studied population, and locally, the markers density and the linkage disequilibrium patterns. Accordingly, we can expect that, due to the variability of these factors along the genome, the windows sizes will vary (both in terms of genomic size and of number of markers), and this has to be taken into account in the program as well. Another potential improvement is on the neighborhood to be used (the parameter K in KNN-MDR). In the current version, a fixed value of K is provided and used for all the windows along the genome. New values of K necessitate a new run of the program. Eventually, several runs will lead to a globally optimal value of K, in the sense that this value leads to the best detection performances. An alternative approach would be to have a better strategy to find this optimal value of K. For example, pre-examining the distances between cases and the distances between controls at local scales (we could think about using the windows that have been defined in the linkage disequilibrium analyses) should lead to values of K that vary along the genome and provide a better tool to discriminate that the global values used in the current version. In other words, we would learn the optimal values for the windows sizes and the number of neighbors from the data.

In large dataset analyses, a pre-selection of the markers is used to reduce the computing load, before progressively zooming on the complete set of markers. Although somehow successful in the results presented in the paper, there is no guarantee that the way the first set of studied markers has been selected is optimal, and other strategy selection could be applied: again, the presence of linkage disequilibrium could lead to use so-called “tag SNP” to represent a complete region as faithfully as possible, an obvious pre-filtering. Also, the approach used to better define the windows should be considered when pre-filtering.

Other fields of KNN-MDR would deserve more attention. The way the distance between individuals is measured is an example: the allelic frequencies should be taken into account because obtaining two times a rare allele randomly is less expected than obtaining twice a more common allele, an obvious statement which is not taken into account in the current distance measure. Also, the distance between individuals could be measured on the basis of haplotypes (which would necessitate obtaining firstly the genetic phase of the individuals) because important mutations are likely to be located on similar haplotypes, the similarity arising from the potentially common origin of the mutation in the tested population. Clearly, a distance measure accounting for the diversity of haplotypes would then need to be devised. The same holds true if other types of markers are considered, such as typically microsatellites, for which the number of alleles and of genotypes might be much larger than for SNP.

The methodology of aggregation also offers perspectives. Since we only used a few methods in this thesis, many other methods have been developed for interaction mapping and a more extensive testing could lead to the discovery of sets of methods performing better together. This calls for a continuous development of a tool embracing the previously developed methodologies together with new methods popping up continuously. This development would include the capacity of a smooth integration of the heterogeneous results produced by the included methods: formats are generally different, ranking of the detected interactions might vary from method to method, etc.

Above the two approaches developed in this thesis, other parametric and non-parametric approaches targeting genomic interactions would be necessary. One of the achievements of our work has been to show a way that information could be introduced in well-established methodologies to improve the method performances: introducing windows allowing to capture a part of the information on linkage disequilibrium has allowed to improve the performances of the multi-dimensional reduction (MDR) methods. The same improvements are also possible in other methods and could potentially offer new advantages to previously developed approaches such as support vector machines or random forests.

Another perspective of this PhD work would be to investigate other possible ways to perform large-scale hypothesis, as is the case in interaction mapping. For example, the false discovery rate (FDR) is a nowadays largely embraced statistical inference technique that could be used as a (probably better) alternative to classical testing in this context.

Although the fundamental importance of genetic interactions to give clues to the genomic dissection of the traits, epigenetic causes are also gaining new perspectives in the diseases related to gene deregulation. Therefore, in the future, we must work more in this field either through the approaches in this thesis or through the development of other approaches.

Finally, we can hope that our current developments and the future advances in the field of interaction mapping will yield new insights into real-life applications. Advances in genotyping/sequencing technologies have led to reduced genotyping costs which, in the near future, could provide numerous sufficiently large samples for genetic interaction analyses. Development of robust methods, providing stable and repeatable results, will therefore remain important for a deeper understanding of the diseases and traits underlying genetic mechanisms in the coming years.

# Conclusions

Throughout this thesis work, we have presented and demonstrated the utility of using non-parametric statistical methods as a tool for detecting genetic interactions.

The introduction has hopefully made clear that important challenges remain in the mapping of genetic variants associated to traits of interest, such as diseases in human, animals and plants, or production traits in production animals and in crops. The long-lasting history of mapping methods has led to a profusion of methods, each with pros and cons. The work in this thesis is a contribution to this domain. In view of their robustness and their adaptability to a large set of situations, we have chosen to work on non-parametric methods because it has seemed to us that many advances were possible in that domain for genomic analyses, as is illustrated by the number of proposed approaches. A non-exhaustive panel of methods in the field of genetic interactions mapping is briefly presented in the introduction of this thesis, as an illustration of the vigorous research in that domain.

In the second chapter, we propose the main contribution of this PhD thesis: a novel approach combining K-Nearest Neighbors (KNN) and Multifactor Dimensionality Reduction (MDR) methods for detecting gene-gene interactions as a possible alternative to existing algorithms, especially in situations where the number of involved determinants is high. This method illustrates how taking into account the physical nature of the problem - the markers present on today dense maps are physically linked on a chromosome, introducing a disequilibrium between close markers due to linkage - allows introducing more information in existing methods, and how this can be used to improve these methods. In our case, we have demonstrated that KNN-MDR is more computationally efficient than other exhaustive strategies, using windows of linked markers instead of single markers, which is facilitating the analysis of large-scale data sets with potentially genome-wide SNPs. The improvements on the efficiency of the method make it eligible for the detection of higher-order interactions, although this would admittedly remain a notably challenging task. Another reason making KNN-MDR useful is its ability to detect interactions in the absence of marginal effects. Several methods use marginal effects to pre-filter the data, assuming that only markers showing some effect individually are likely to be involved in interactions. We have considered this as an excessive assumption, and consequently developed a strategy where this assumption is not necessary. Relaxing this assumption in our simulations has proved that KNN-MDR performed generally better than concurrent methods in this context.  Although we have demonstrated some of the advantages of the method, we are aware that improvements are possible, and some ideas in that direction are proposed in the perspectives. These perspectives could render the method more useable for external users and increase its use.

The main idea of the third chapter, on the aggregation of methods, is a more general concept: grouping various methods results might lead to improvements over the individual results. We have illustrated this concept in the field of interactions mapping and obtained results somehow confirming these improvements. In a field such as genetic mapping, where thousands of parallel tests are performed, false positives control is an important issue. False positives are likely to be numerous in these studies, meaning a potential waste of time, energy and money on non-reproducible results, which of course shed doubts on the utility of such studies. We hope that our work is again a step in the direction of an improvement in the perception of interaction studies: we have shown that using a small set of methods in a very simple aggregation strategy led to an increase in the detection power while properly controlling for the false positive rate. Above providing a framework for a joint - i.e. made with several methods - analysis of real datasets, we hope that such results will stimulate interest in the development of new methods: this would be beneficial for the field - new performing methods would be welcome - and for the aggregation strategy - adding more performing methods should enhance the performances of the aggregation strategy. The feasibility of using such methodology on real genome-wide datasets has been demonstrated on an example. This also calls for improvement in the future methods to be developed, because many methods in use today would not be able to manage large genome-wide datasets, which questions on their ability to detect interactions involving variants distantly located on the genome.

In conclusion, we believe that these methods (KNN-MDR and aggregate expert) are valuable in the context of loci and interactions mapping and can be seen as an interesting addition to the arsenal used in complex traits analyses. The output of these methods can enhance the understanding of the biological mechanism of diseases and other traits, and this new knowledge can contribute to the prediction of clinical diseases, to the prevention of most common complex diseases, and to a better understanding of numerous traits.

# References

Abo Alchamlat, S. and F. Farnir (2017). "KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies." BMC Bioinformatics 18(1).

Agrawal, S., Raja, R. and Agrawal, S. (2012). "Support Vector Machine for age classification", International Journal of Emerging Technology and Advanced Engineering, 2.

Aschard, H. (2015). "A Perspective on Interaction Tests in Genetic Association Studies" bioRxiv

Aulchenko, Y. S., D. J. de Koning and C. Haley (2007). "Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis." Genetics 177(1): 577-585.

Ban, H. J., J. Y. Heo, K. S. Oh and K. J. Park (2010). "Identification of type 2 diabetes-associated combination of SNPs using support vector machine." BMC Genet 11: 26.

Barnholtz-Sloan, J. S., X. Guan, C. Zeigler-Johnson, N. J. Meropol and T. R. Rebbeck (2011). "Decision tree-based modeling of androgen pathway genes and prostate cancer risk." Cancer Epidemiol Biomarkers Prev 20(6): 1146-1155.

Basheer, I. A. and M. Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application." J Microbiol Methods 43(1): 3-31.

Binder, H. and M. Schumacher (2008). "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." BMC Bioinformatics 9: 14.

Botta, V., G. Louppe, P. Geurts and L. Wehenkel (2014). "Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies." PLoS One 9(4).

Breiman, L. (2001). "Random Forests." Machine Learning 45(5-32).

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984). "Classification and Regression Trees". Taylor & Francis.

Calle, M., V. Urrea, N. Malats and K. Van steen (2008). MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. Genètica general. Citogenètica general. Immunogenètica. Evolució. Filogènia. Universitat de Vic, 2008-02-05.

Calzone, L., E. Barillot and A. Zinovyeva (2015). "Predicting genetic interactions from Boolean models of biological networks." bioRxiv.

Carlborg, Ö. and Haley, C.S. (2004) "Epistasis: too often neglected in complex trait studies? " Nature Genetics, 5.

Cattaert, T., M. L. Calle, S. M. Dudek, J. M. Mahachie John, F. Van Lishout, V. Urrea, M. D. Ritchie and K. Van Steen (2011). "Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise." Ann Hum Genet 75(1): 78-89.

Cattaert, T., V. Urrea, A. C. Naj, L. De Lobel, V. De Wit, M. Fu, J. M. Mahachie John, H. Shen, M. L. Calle, M. D. Ritchie, T. L. Edwards and K. Van Steen (2010). "FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals." PLoS One 5(4).

Chen, S. H., J. Sun, L. Dimitrov, A. R. Turner, T. S. Adams, D. A. Meyers, B. L. Chang, S. L. Zheng, H. Gronberg, J. Xu and F. C. Hsu (2008). "A support vector machine approach for detecting gene-gene interaction." Genet Epidemiol 32(2).

Chris, W., J. Antony, P. Nikolas, L. P. Marcin, S. B. Oliver, D. C. Jason, R. G. Arcadio, C. F. Ricardo, G. Hui, M. W. Neil, J. S. Deborah, S. R. Stephen, O. Suna, J. S. Stephen, B. Maria, R. Sylvia, A. T. John and S. W. Linda (2015). "Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping." PLoS Genet 11(6).

Cordell, H. J. (2009). "Detecting gene–gene interactions that underlie human diseases." Nature Reviews Genetics 10(6): 392-404.

Costanzo, M., B. VanderSluis, E. N. Koch, A. Baryshnikova, C. Pons., T. Guihong, W. Wen, A. P. Rosebrock, A. A. Caudy, C. L. Myers, B. Andrews and C. Boone (2016). "A global genetic interaction network maps a wiring diagram of cellular function." Science 353(6306).

Curtis, D. (2007). "Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association." BMC Genet 8: 49.

Elster, C., F. Schubert, A. Link, M. Walzel, F. Seifert and H. Rinneberg (2005). "Quantitative magnetic resonance spectroscopy: Semi-parametric modeling and determination of uncertainties." Magn Reson Med 53(6): 1288-1296.

Estrada-Gil, J. K., J. C. Fernandez-Lopez, E. Hernandez-Lemus, I. Silva-Zolezzi, A. Hidalgo-Miranda, G. Jimenez-Sanchez and E. E. Vallejo-Clemente (2007). "GPDTI: a Genetic Programming Decision Tree induction method to find epistatic effects in common complex diseases." Bioinformatics 23(13): i167-174.

Fang, G., M. Haznadar, W. Wang, H. Yu, M. Steinbach, T. R. Church, W. S. Oetting, B. Van Ness and V. Kumar (2012). "High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions." PLoS One 7(4): e33531.

Fang, Y. H. and Y. F. Chiu (2012). "SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies." Genet Epidemiol 36(2): 88-98.

Fang, Y. H. and Y. F. Chiu (2013). "A Novel Support Vector Machine-Based Approach for Rare Variant Detection." PLoS One 8(8).

Forsberg, S. K. G., J. S. Bloom, M. J. Sadhu, L. Kruglyak and Ö. Carlborg (2017). "Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast." Nature Genetics 49(4): 497-503.

Frau, F., D. Crowther, H. Ruetten and K. V. Allebrandt (2017). "Type-2 diabetes-associated variants with cross-trait relevance: Post-GWAs strategies for biological function interpretation." Mol Genet Metab 121(1): 43-50.

Fuxman Bass, J. I., C. Pons, L. Kozlowski, J. S. Reece-Hoyes, S. Shrestha, A. D. Holdorf, A. Mori, C. L. Myers and A. J. M. Walhout (2016). "A gene-centered C. elegans protein–DNA interaction network provides a framework for functional predictions." Mol Syst Biol 12(10): 884.

Ganatra, A. P. and Y. P. Kosta (2010). "Comprehensive Evolution and Evaluation of Boosting." IJCTE 2(6).

Ghahramani, Z. (2012). "Bayesian non-parametrics and the probabilistic approach to modelling." Philos Trans A Math Phys Eng Sci 371(1984): 20110553-20110553.

Gianola, D., R. L. Fernando and A. Stella (2006). "Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures." Genetics 173(3): 1761-1776.

Glickman, M. E. and D. A.D. (2007). Topics in Biostatistics.

Gola, D., J. M. Mahachie John, K. van Steen and I. R. Konig (2015). "A roadmap to multifactor dimensionality reduction methods." Brief Bioinform 17(2)

Goodrich, B. (2012). Estimating the Number of Common Inputs to a Data-Generating Process. Columbia University.

Greene, C.S., Penrod, N. M., Williams, S. M. and Moore, J.H. (2009). "Failure to Replicate a Genetic Association May Provide Important Clues About Genetic Architecture" PLoS ONE, 4(6).

Gui, J., A. S. Andrew, P. Andrews, H. M. Nelson, K. T. Kelsey, M. R. Karagas and J. H. Moore (2011). "A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility." Ann Hum Genet 75(1): 20-28.

Gunther, F., N. Wawro and K. Bammann (2009). "Neural networks for modeling gene-gene interactions in association studies." BMC Genet 10: 87.

Guy, R. T., P. Santago and C. D. Langefeld (2012). "Bootstrap aggregating of alternating decision trees to detect sets of SNPs that associate with disease." Genet Epidemiol 36(2): 99-106.

Hamilton, F., A. Lloyd and K. Flores (2017). "Hybrid Modeling and Prediction of Dynamical Systems." PLoS Comput Biol 13(7).

He, H., W. S. Oetting, M. J. Brott and S. Basu (2009). "Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study." BMC Med Genet 10: 127.

Hill, W.G., Goddard, M.E, and Visscher, P.M.(2008). "Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits", PLoS Genetics 4(2).

Ho, D. E., K. Imai, G. King and E. A. Stuart (2017). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis 15(03): 199-236.

Howard, R., A. L. Carriquiry and W. D. Beavis (2014). "Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures." G3 (Bethesda) 4(6).

Hu, T., N. A. Sinnott-Armstrong, J. W. Kiralis, A. S. Andrew, M. R. Karagas and J. H. Moore (2011). "Characterizing genetic interactions in human disease association studies using statistical epistasis networks." BMC Bioinformatics 12(1): 364.

Huang, L. C., S. Y. Hsu and E. Lin (2009). "A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data." J Transl Med 7: 81.

Isci, S., H. Dogan, C. Ozturk and H. Otu (2014). "Bayesian network prior: network analysis of biological data using external knowledge." Bioinformatics 30(6).

Jackson, C. (2016). "flexsurv: A Platform for Parametric Survival Modeling inR." Journal of Statistical Software 70(8).

Jiang, R., W. Tang, X. Wu and W. Fu (2009). "A random forest approach to the detection of epistatic interactions in case-control studies." BMC Bioinformatics 10 Suppl 1: S65.

Jung, H.-Y., S. Leem, S. Lee and T. Park (2016). "A novel fuzzy set based multifactor dimensionality reduction method for detecting gene–gene interaction." Comput Biol Chem 65: 193-202.

Kadarmideen, H. N. (2014). "Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities." Livest Sci 116: 232-248.

Katsanis, S. H. and N. Katsanis (2013). "Molecular genetic testing and the future of clinical genomics." Nat Rev Genet 14(6): 415-426.

Konomi, M. and G. M. Sacha (2017). "Feedforward neural network methodology to characterize thin films by Electrostatic Force Microscopy." Ultramicroscopy 182: 243-248.

Koo, C. L., M. J. Liew, M. S. Mohamad and A. H. Salleh (2013). "A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology." Biomed Res Int 2013: 432375.

Korte, A. and A. Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review." Plant Methods 9(29).

Li, J., J. Dan, C. Li and R. Wu (2013). "A model-free approach for detecting interactions in genetic association studies." Brief in Bioinform 15(6): 1057-1068.

Li, J., B. Horstman and Y. Chen (2011). "Detecting epistatic effects in association studies at a genomic level based on an ensemble approach." Bioinformatics 27(13): i222-229.

Li, M., J. C. Gardiner, a. NBreslau, J. C. Anthony and Q. Lu (2014). "A non-parametric approach for detecting gene-gene interactions associated with age-at-onset outcomes." BMC Genetics 15(79).

Libioulle, C., E. Louis, S. Hansou, C. Sandor, F. Farnir, D. Franchimont, S. Vermeire, O. Dewit, M. Vos, A. Dixon, B. Demarche, I. Gut, S. Heath, M. Foglio, L. Liang, D. Laukens, M. Mni, D. Zelenika, A. Van Gossum, P. Rutgeerts, J. Belaiche, M. Lathrop and M. Georges (2007). "Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4." PLoS Genet 3(4).

Liu, B., B. McKay and H. A. Abbass (2003). Improving genetic classifiers with a boosting algorithm. Evolutionary Computation 4: 2596 - 2602

Liu, C. (2013). "Baysian statistical methods in gene-environment and gene-gene interaction studies." Doctor of Philosophy (PhD), The University of Texas Graduate School of Biomedical Sciences at Houston.

Liu, C., J. Ma and C. I. Amos (2015). "Bayesian variable selection for hierarchical gene-environment and gene-gene interactions." Hum Genet 134(1): 23-36.

Lubke, G., C. Laurin, R. Walters, N. Eriksson, P. Hysi, T. Spector, G. Montgomery, N. Martin, S. Medland and D. Boomsma (2013). "Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data." J Data Mining Genomics Proteomics 4.

Lunetta, K. L., L. B. Hayward, J. Segal and P. Van Eerdewegh (2004). "Screening large-scale association study data: exploiting interactions using random forests." BMC Genet 5: 32.

Ma, S., L. Yang, R. Romero and Y. Cui (2011). "Varying coefficient model for gene–environment interaction: a non-linear look." Bioinformatics 27(15): 2119-2126.

Mackay, T.F. and Moore, J.H. (2014). "Why epistasis is important for tackling complex human disease genetics", Genome Medicine, 6.

Mahachie John, J. M., F. Van Lishout and K. Van Steen (2011). "Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data." Eur J Hum Genet 19(6): 696-703.

Maity, A. and X. Lin (2011). "Powerful Tests for Detecting a Gene Effect in the Presence of Possible Gene-Gene Interactions Using Garrote Kernel Machines." Biometrics 67(4): 1271-1284.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature 461(7265): 747-753.

Manuguerra, M., G. Matullo, F. Veglia, H. Autrup, A. M. Dunning, S. Garte, E. Gormally, C. Malaveille, S. Guarrera, S. Polidoro, F. Saletta, M. Peluso, L. Airoldi, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, D. Trichopoulos, A. Kalandidi, D. Palli, V. Krogh, R. Tumino, S. Panico, H. B. Bueno-De-Mesquita, P. H. Peeters, E. Lund, G. Pera, C. Martinez, P. Amiano, A. Barricarte, M. J. Tormo, J. R. Quiros, G. Berglund, L. Janzon, B. Jarvholm, N. E. Day, N. E. Allen, R. Saracci, R. Kaaks, P. Ferrari, E. Riboli and P. Vineis (2007). "Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions." Carcinogenesis 28(2): 414-422.

Marcelo, J. and S. Viana (2005). "Dominance, epistasis, heritabilities and expected genetic gains." Genet. Mol. Biol. 28(1).

Martin, E. R., M. D. Ritchie, L. Hahn, S. Kang and J. H. Moore (2006). "A novel method to identify gene-gene effects in nuclear families: the MDR-PDT." Genet Epidemiol 30(2): 111-123.

Maudes, J., J. J. Rodríguez, C. García-Osorio and N. García-Pedrajas (2012). "Random feature weights for decision tree ensemble construction." Information Fusion 13(1): 20-30.

Moore, J. H., F. W. Asselbergs and S. M. Williams (2010). "Bioinformatics challenges for genome-wide association studies." Bioinformatics 26(4): 445-455.

Moore, J. H. and S. M. Williams (2005). "Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis." BioEssays 27(6): 637-646.

Morgante, F., Huang, W., Maltecca, C. and Mackay, T. F. (2018). "Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals" Heredity 120:500-514.

Motsinger-Reif, A. A., S. Deodhar, S. J. Winham and N. E. Hardison (2010). "Grammatical evolution decision trees for detecting gene-gene interactions." BioData Min 3(1): 8.

Motsinger-Reif, A. A., T. J. Fanelli, A. C. Davis and M. D. Ritchie (2008). "Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error." BMC Res Notes 1: 65.

Motsinger-Reif, A. A. and M. D. Ritchie (2008). "Neural networks for genetic epidemiology: past, present, and future." BioData Min 1(1): 3.

Motsinger-Reif, A.A. (2008). "The effect of alternative permutation testing strategies on the performance of multifactor dimensionality reduction" BMC Research Notes 1: 139.

Mountrakis, G., J. Im and C. Ogole (2011). "Support vector machines in remote sensing: A review." ISPRS Journal of Photogrammetry and Remote Sensing 66(3): 247-259.

Mullaney, J. M., R. E. Mills, W. S. Pittard and S. E. Devine (2010). "Small insertions and deletions (INDELs) in human genomes." Hum Mol Genet 19(R2): R131-R136.

Nowak, M. A., M. C. Boerlijst, J. Cooke and J. M. Smith (1997). "Evolution of genetic redundancy." Nature 388(10).

Ozgur, A., T. Vu, G. Erkan and D. R. Radev (2008). "Identifying gene-disease associations using centrality on a literature mined gene-interaction network." Bioinformatics 24(13): i277-285.

Pashova, H., M. LeBlanc and C. Kooperberg (2013). "Boosting for detection of gene-environment interactions." Stat Med 32(2): 255-266.

Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning 1(81-106).

Quintana-Murci, L. (2016). "Understanding rare and common diseases in the context of human evolution." Genome Biology 17(1).

Riancho, J. A. (2012). "Genome-Wide Association Studies (GWAS) in Complex Diseases: Advantages and Limitations." Reumatol Clin 8(2).

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." Am J Hum Genet 69(1): 138–147.

Ritchie, M. D. and A. A. Motsinger (2006). "Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies." Hum Genomic 2(5).

Ritchie, M. D., B. C. White, J. S. Parker, L. Hahn and J. H. Moore (2003). "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases." BMC Bioinformatics 4(28).

Rojas, R. (2009). AdaBoost and the Super Bowl of Classiers. University Berlin.

Sariyar, M., I. Hoffmann and H. Binder (2014). "Combining techniques for screening and evaluating interaction terms on high-dimensional time-to-event data." BMC Bioinformatics 15(1): 58.

Schapire, R. (1990). "The Strength of Weak Learnability." Machine Learning 5(197-227).

Schmutz, S. M. and T. G. Berryere (2007). "The Genetics of Cream Coat Color in Dogs." J Hered 98(5): 544-548.

Sebastiani, P., Y. H. Yu and M. F. Ramoni (2003). "Bayesian machine learning and its potential applications to the genomic study of oral oncology." Adv Dent Res 17: 104-108.

Sheng, X. C., Z. Yu and X. Qu (2014). "Support Vector Machine Optimized by Improved Genetic Algorithm." TELKOMNIKA Indonesian Journal of Electrical Engineering 12(1).

Stanislas, V., C. Dalmasso and C. Ambroise (2017). "Eigen-Epistasis for detecting gene-gene interactions." BMC Bioinformatics 18(1).

Sun, G.-Q. and I. Shuryak (2017). "Advantages of Synthetic Noise and Machine Learning for Analyzing Radioecological Data Sets." PloS One 12(1).

Thapar, A. and M. Cooper (2013). "Copy Number Variation: What Is It and What Has It Told Us About Child Psychiatric Disorders?" J Am Acad Child Adolesc Psychiatry 52(8): 772-774.

Timo, H., D. Wodzisaw, G. Mark and K. Samuel (2011). "Artificial Neural Networks and Machine Learning, Part II - ICANN" 25th International Conference on Artificial Neural Networks

Tomita, Y., S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi and H. Honda (2011). "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma." BMC Bioinformatics 5: 120.

Upstill-Goddard, R., D. Eccles, S. Ennis, S. Rafiq, W. Tapper, J. Fliege and A. Collins (2013). "Support Vector Machine Classifier for Estrogen Receptor Positive and Negative Early-Onset Breast Cancer." PLoS One 8(7).

Upstill-Goddard, R., D. Eccles, J. Fliege and A. Collins (2012). "Machine learning approaches for the discovery of gene-gene interactions in disease data." Briefings in Bioinformatics 14(2): 251-260.

Vieira, M. L. C., L. Santini, A. L. Diniz and C. d. F. Munhoz (2016). "Microsatellite markers: what they mean and why they are so useful." Genet Mol Biol 39(3): 312-328.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five years of GWAS discovery." Am J Hum Genet 90(1): 7-24.

Wan, X., C. Yang, Q. Yang, H. Xue, N. L. Tang and W. Yu (2009). "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study." BMC Bioinformatics 10: 13.

Wang, J., T. Joshi, B. Valliyodan, H. Shi, Y. Liang, H. T. Nguyen, J. Zhang and D. Xu (2015). "A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies." BMC Genomics 16: 1011.

Wang, L., J. Zhu and H. Zou (2008). "Hybrid huberized support vector machines for microarray classification and gene selection." Bioinformatics 24(3): 412-419.

Wang, Z. and C. Wang (2010). "Buckley-James boosting for survival analysis with high-dimensional biomarker data." Stat Appl Genet Mol Biol 9(1).

Winham, S. J., C. L. Colby, R. R. Freimuth, X. Wang, M. Andrade, M. Huebner and J. Biernacka (2012). "SNP interaction detection with Random Forests in high-dimensional genetic data." BMC Bioinformatics 13(164).

Yang, C. H., Y. D. Lin, L. Y. Chuang, J. B. Chen and H. W. Chang (2013). "MDR-ER: balancing functions for adjusting the ratio in risk classes and classification errors for imbalanced cases and controls using multifactor-dimensionality reduction." PLoS One 8(11): e79387.

Yang, Z. and B. Rannala (2012). "Molecular phylogenetics: principles and practice." Nat Rev Genet 13(5): 303-314.

Yee, J., Y. Kim, T. Park and M. Park (2016). "Using the Generalized Index of Dissimilarity to Detect Gene-Gene Interactions in MultiClass Phenotypes." PLoS One 11(8).

Yi, N. (2010). "Statistical analysis of genetic interactions." Genet Res (Camb) 92(5-6): 443-459.

Yi, N., V. G. Kaklamani and B. Pasche (2011). "Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk." Ann Hum Genet 75(1): 90-104.

Yoo, W., B. Ference, M. Cote and A. Schwartz (2012). "A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions." Int J Appl Sci Technol 2(7).

York, T. L., R. T. Durrett, S. Tanksley and R. Nielsen (2005). " Bayesian and maximum likelihood estimation of genetic maps." Genet Res 85(2): 159-168.

Zou, Q., J. R. Haliburton, W. Shao, A. Deutschbauer, A. Arkin and A. R. Abate (2017). "Genetic interaction mapping with microfluidic-based single cell sequencing." Plos One 12(2).

# Appendices (Additional files)

## Appendix 1 (Additional file 6): KNN MDR: user's guide

### Introduction

KNN MDR is a fortran 90 software implementing the KNN MDR methodology detailed in (ABO ALCHAMLAT S. and Farnir F., 2017). The aim of the program is to help obtaining clues about the position of the genes involved jointly in a phenotype. One of the interests of the approach is that it is able to find interacting genes even in the absence of marginal effects. This capability, which was already present in other methods, such as MDR (Multifactor Dimensionality Reduction) is made available through KNN MDR for situations with more markers and more complicated interaction patterns than was feasible computer wise with "simple" MDR. The current version has been written for binary (0/1) traits and for SNP data, but could easily be extended to other traits/attributes. These options will be included in future versions of the software.

### Methodology

This section shortly summarizes how the method works in order to understand the parameters that need to be provided to the software to obtain results. More details can be found in the original publication. Each data point is represented through a phenotype (0/1, where the meaning of these codes is problem dependent) and a set of N attributes. As mentioned above, in the current version of the program, the attributes correspond to SNP genotypes. It will be assumed that these genotypes are available (through direct genotyping or through an imputation method) for all individuals. The idea behind the MDR methods is to reduce the very large multidimensional space faced in situations involving multiple loci (such as genetic interactions) to one dimensional space.

In basic MDR, a status (0/1) is associated to each multi locus genotype through a majority vote performed on the individuals presenting this multi locus genotype in the training set; after that training stage, status (0/1) can be allocated to individuals from the test set on the basis of their multi locus genotype as well (provided similar genotypes were present in the training set). Accuracy of allocation can then be obtained by computing the false positive (i.e. 1) and false negative (i.e. 0) rates in both training and test sets.

KNN MDR uses such a strategy. The difference with the basic MDR is that the allocation phase is performed through a K nearest-neighbors approach: a status is allocated based on the most prevalent status within the set of the K nearest neighbors of the tested individual. The neighborhood is defined using a distance, which, in the current version, is a simple Euclidian distance between the involved genotypes of both individuals for which a distance is computed. In our studies, only SNPs have been used, for which the distance proposed in the Mahalanobis measure makes sense, with D(AA, AB)=D(AB, BB)=0.5* D(AA, BB), where AA, AB and BB are the three possible SNP genotypes

So, for a window embracing M markers, the distance between individuals i and j is:

where and are the number of copies of the A allele at marker k for individuals i and j, respectively.

The advantage of such an approach is double: the distance can be easily (i.e. with not much effort computer wise) computed for any number of markers, and the allocation procedure works even in situation where no other individual in the training set has the same multi locus genotype. Note that both these points become more relevant as the number of involved markers increases, a practically frequent situation.

An issue exists over the definition of the training and test sets. Again, KNN MDR mimics the approach followed in basic MDR using cross validation: the complete dataset is randomly split into V equally sized subsets, and each subset is sequentially considered as a test set, while the (V-1) other sets are used as training sets. Accuracy is computed for every configuration, and the final model accuracy is computed as the average of the obtained accuracies. In order to balance the true positive and true negative rates in the results, we used "balanced accuracy" as our accuracy measurement, where "balanced accuracy" is defined as the average of true positive and true negative rates.

When looking for sets of genes involved in a phenotype, various attributes sets are usually tested in order to find the one best explaining the data, which is, in our approach, the one with the highest balanced accuracy on the test set. This "best" attribute set will be considered as our "best model".

The last problem is to test the significance of the best model. This is done in our software through a permutation procedure: if a specific attribute set is associated to the phenotype, disrupting the association between phenotypes and genotypes should destroy this association. Consequently, by permuting randomly the phenotypes with respect to the genotypes, we create datasets where no association should exist, which corresponds to the null hypothesis we want to test. Comparing the truly obtained balanced accuracy to the ones obtained on the permuted datasets allows one to obtain an estimation of the p-value associated to our best model.

**Parameters**

Several parameters have been defined in the previous and can be transmitted to the program. These parameters are provided through a parameters file, which is invoked while calling the program, as follows:

```
path/knn mdr <analysis name>
```

In this command, "path" represents the eventual path leading to the executable, and "name" represents the name of the analysis. This name is used to provide the parameters file just discussed (named <analysis name>.prm) and to name output files (see below). The parameters file is a text file, where each line is used to specify the various options of the program. These options are:

- **ATT SET FILE file:** this option is used to specify the file containing the list of attributes sets for which an evaluation is demanded. The best model will be chosen among these attributes sets. Attributes sets are specified on distinct lines of the file by providing a comma separated list of the positions of the attributes to be considered in the attributes file. When several consecutive attributes have to be used, the notation using the first and the last attribute separated with an hyphen can be used. For example, "1,3,7-10" means "use first, third, seventh, eighth, nineth and tenth attributes" of the attributes file. No default exists for this parameter.
- **ATTRIB FILE file:** with this option, the file containing all attributes for all individuals in the analysis can be given. Again, "file" is a text file, with one line per individual, and at least as many blank separated columns as the number M of attributes. The attributes file also contains

a column with an individual identifier, and may also contain the (0/1) phenotype. Since attributes, in the current version, are SNP genotypes, these genotypes are assumed to be recoded genotypes: for each SNP, one of the allele is arbitrarily considered as the reference allele, and the recoded genotype is simply the number of occurrences of the reference allele in the genotype. Consequently, the allowed attributes values are either 0, 1 or 2. No default exists for this parameter.

- **HELP:** this option is used to obtain an short reminder of the available options.
- **KLOWn:** this option allows specifying the minimum number of neighbors to be used to allocate status to tested individuals. Default is KLOW = 5.
- **KHIGH n**: this option allows specifying the maximum number of neigh bors to be used to allocate status to tested individuals. Default is KHIGH = 5.
- **MODEL model:** with this option, a model can be specified. In the current version, the only available model is KNN.... Default is 'KNN'.
- **NB ATTRIB n:** this option indicates how many attributes should be found in the data file. Default is NB ATTRIB = 1.
- **NB CROSS V n:** this is used to provide the number V of cross-validation subsets. Default is NB CROSS V = 10.
- **NB INDIV n:** this option indicates how many individuals should be found in the data file. Default is NB INDIV = 1.
- **NB PERM n:** with this option, the number of permutations can be provided. Default is NB PERM = 0.
- **PHENO FILE file:** with this option, the file containing the phenotypes for all individuals in the analysis can be given. As above, "file" is a text file, with one line per individual, one column with the 0/1 phenotype and a column with an individual identifier. This file may be the same as the attributes file. No default exists for this parameter.
- **POS FIRST ATTRIB n:** with this option, the position (column number) of the first attribute to be considered can be provided. Default value is POS FIRST ATTRIB = 1.
- **POS LAST ATTRIB n**: with this option, the position (column number) of the last attribute to be considered can be provided. Default value is POS LAST ATTRIB = 1.
- **POS ID ATTRIB n:** this option allows providing the position (column number) of the individual identifier in the attributes file. Default is POS ID ATTRIB = 1.
- **POS ID PHENO n:** this option allows providing the position (column number) of the individual identifier in the phenotypes file. Default is POS ID PHENO = 1.
- **POS PHENO n:** this option allows providing the position (column number) of the phenotype field in the phenotypes file. Default is POS PHENO = 2.

- **SEED s:** since random choices (cross-validation subsets, permutations) are made, successive invocations of the program will not necessarily result in identical outputs. Identical (different) runs can be performed by specifying identical (different) seeds through this option.

**Example**

In this section, we show the use of the program on a simulated example. Twenty attributes are measured for 500 cases and 500 controls. All data are included in one file, named 'sample.dat'. The individual identifier is the first field, followed by the 0/1 phenotype, and then by 20 attributes. An interaction has been introduced artificially between attributes 4 and 12 as follows: all genotypes at locus 4 are generated randomly, irrespectively of the status of the individuals.

This should ensure that no marginal effect exists for this locus on the trait.

Controls genotypes for locus 12 are also randomly allocated, but cases genotypes for locus 12 are copies of control ones. This creates an interaction between these two loci.

**Attributes and phenotype file**

As mentioned, the attributes and the phenotype are included in the same file, named sample.dat. The first and last 2 lines are provided below as examples of data lines:

```
1 0 1 1 2 1 2 1 1 1 0 2 2 1 1 1 2 2 1 1 1 1
2 0 1 2 1 1 2 1 0 2 1 1 1 0 1 1 0 0 1 2 1 1
...
999 1 0 0 0 1 0 0 1 2 1 0 0 1 0 0 0 0 0 1 2 2
1000 1 2 1 1 1 1 0 0 0 2 0 1 1 0 1 1 1 1 2 2 1
```

**Attributes sets file**

The attributes sets file sample.set has been created to span the 20 attributes using 5 markers-wide windows. So 4 single windows are first tested (markers 1-5, 6-10, 11-15 and 16-20). After testing single windows, all sets of 2, 3 or 4 windows are considered. This leads to the following file:

```
1-5
6-10
11-15
16-20
1-10
1-5,11-15
1-5,16-20
6-15
6-10,16-20
11-20
1-15
1-10,16-20
1-5,11-20
6-20
1-20
```

**Parameters file**

The parameters file sample.prm is as follows:

```
ATT_SET_FILE sample.set
ATTRIB_FILE sample.dat
KLOW 4
KHIGH 5
MODEL KNN
NB_ATTRIB 20
NB_CROSS_V 10
NB_INDIV 1000
NB_PERM 100
PHENO_FILE sample.dat
POS_FIRST_ATTRIB 3
POS_LAST_ATTRIB 22
POS_ID_ATTRIB 1
POS_ID_PHENO 1
POS_PHENO 2
SEED 123
```

**Running the program**

To execute a run of the program, simply type:

```
knn mdr sample
```

The program starts running and show intermediate results on the screen. All reported results and more) are also reported to knn mdr.log file for further reference if needed.

**Interpreting the output**

Three output files are generated: <analysis name>.log, <analysis name>.cv and <analysis name>.perm. Let's first take a look at the 3 first and last lines of <analysis name>.cv. This file shows how the various individuals in the dataset have been allocated to the cross-validation subsets:

```
Indiv 1 -> Subset 1
Indiv 2 -> Subset 2
Indiv 3 -> Subset 10
...
Indiv 998 -> Subset 10
Indiv 999 -> Subset 9
Indiv 1000 -> Subset 5
```

The next file is the <analysis name>.perm file which shows the individual (i.e. for each cross-validation subset) and average (over the cross-validation subsets) balanced accuracies obtained for the real and permuted data. Again we show the few first and last lines of the file:

```
0 1 0.61317408 0.57258689
0 2 0.59042907 0.61969697
0 3 0.57969087 0.56854343
0 4 0.59012622 0.59803927
```

```
0 5 0.56991446 0.65815413
0 6 0.59432721 0.53693694
0 7 0.57333332 0.55000001
0 8 0.58455396 0.53030300
0 9 0.58904111 0.65914893
0 10 0.58518159 0.61642408
0 AVG 0.58697718 0.59098333
1 1 0.49010193 0.51542205
1 2 0.51209480 0.52012885
1 3 0.51317942 0.50740135
1 4 0.51018775 0.50146198
1 5 0.48608297 0.51298702
... ... ... ...
100 5 0.51373267 0.47380954
100 6 0.49153537 0.45833331
100 7 0.50000000 0.47355768
100 8 0.50441492 0.57211542
100 9 0.52405810 0.48684210
100 10 0.50578344 0.46470588
100 AVG 0.49897560 0.50332409
```

The first column represents the permutation number (permutation 0 corresponds to the real not permuted data), the second represents the cross-validation subset (AVG represents the average over all the subsets), the third and the fourth are the training and test balanced accuracies, respectively. Finally, the third file (<analysis name>.log) is the most important one, providing details on the execution of the program along with the main results. The content is given below:

```
Starting the program...
Time is 15:41:05 on January 09,2014
Step 0: name of the analysis: sample
Step 1: obtaining parameters
KNN_MDR will be launched with following options:
NB_INDIV 1000
NB_ATTRIB 20
ATTRIB_FILE sample.dat
ATT_SET_FILE sample.set
PHENO_FILE sample.dat
POS_ID_ATTRIB 1
POS_ID_PHENO 1
POS_FIRST_ATTRIB 3
POS_LAST_ATTRIB 22
POS_PHENO 2
```

```
NB_CROSS_V 10
NB_PERM 100
MODEL KNN
KLOW 4
KHIGH 5
SEED 123
Step 2: reading data
=> 1000 phenotypic values have been read
=> 1000 attributes sets have been read
Step 3: defining cross-validation subsets
=> Allocation of CV subsets reported to sample.cv
Step 4: defining attributes sets
=> Number of attributes sets: 15
Step 5: looping through attributes sets
==> Attribute set 1: 1,2,3,4,5
===> Average balanced accuracy for set 1 = 0.50148523 0.50684011
===> Best balanced accuracy after set 1 = 0.50148523 0.50684011
==> Attribute set 2: 6,7,8,9,10
===> Average balanced accuracy for set 2 = 0.49123150 0.50559229
===> Best balanced accuracy after set 2 = 0.50148523 0.50684011
==> Attribute set 3: 11,12,13,14,15
===> Average balanced accuracy for set 3 = 0.47580791 0.49732471
===> Best balanced accuracy after set 3 = 0.50148523 0.50684011
==> Attribute set 4: 16,17,18,19,20
===> Average balanced accuracy for set 4 = 0.49637920 0.50071424
===> Best balanced accuracy after set 4 = 0.50148523 0.50684011
==> Attribute set 5: 1,2,3,4,5,6,7,8,9,10
===> Average balanced accuracy for set 5 = 0.50194442 0.52629578
===> Best balanced accuracy after set 5 = 0.50194442 0.52629578
==> Attribute set 6: 1,2,3,4,5,11,12,13,14,15
===> Average balanced accuracy for set 6 = 0.66471612 0.67562711
===> Best balanced accuracy after set 6 = 0.66471612 0.67562711
==> Attribute set 7: 1,2,3,4,5,16,17,18,19,20
===> Average balanced accuracy for set 7 = 0.51523346 0.52965009
===> Best balanced accuracy after set 7 = 0.66471612 0.67562711
==> Attribute set 8: 6,7,8,9,10,11,12,13,14,15
===> Average balanced accuracy for set 8 = 0.51154667 0.51678431
===> Best balanced accuracy after set 8 = 0.66471612 0.67562711
==> Attribute set 9: 6,7,8,9,10,16,17,18,19,20
===> Average balanced accuracy for set 9 = 0.49743909 0.51876813
```

```
===> Best balanced accuracy after set 9 = 0.66471612 0.67562711
==> Attribute set 10: 11,12,13,14,15,16,17,18,19,20
===> Average balanced accuracy for set 10 = 0.50408757 0.50820243
===> Best balanced accuracy after set 10 = 0.66471612 0.67562711
==> Attribute set 11: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
===> Average balanced accuracy for set 11 = 0.63244808 0.63826907
===> Best balanced accuracy after set 11 = 0.66471612 0.67562711
==> Attribute set 12: 1,2,3,4,5,6,7,8,9,10,16,17,18,19,20
===> Average balanced accuracy for set 12 = 0.50658184 0.52353173
===> Best balanced accuracy after set 12 = 0.66471612 0.67562711
==> Attribute set 13: 1,2,3,4,5,11,12,13,14,15,16,17,18,19,20
===> Average balanced accuracy for set 13 = 0.62913483 0.64112699
===> Best balanced accuracy after set 13 = 0.66471612 0.67562711
==> Attribute set 14: 6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
===> Average balanced accuracy for set 14 = 0.51725930 0.52325708
===> Best balanced accuracy after set 14 = 0.66471612 0.67562711
==> Attribute set 15: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
===> Average balanced accuracy for set 15 = 0.58697718 0.59098333
===> Best balanced accuracy after set 15 = 0.66471612 0.67562711
Step 6: reporting the most significant attributes set
p-value = 0.0000
List of attributes: 1,2,3,4,5,11,12,13,14,15
Now ending the program...
Time is 15:45:20 on January 09,2014
```

The content of this file is easily understandable. Step 6 reports that set 6, containing markers 1-5 and 11-15, is significantly associated to the phenotype, which is good news given the way the dataset has been generated... Note also that the p value equal to 0 is obtained through permutations and is only an estimator of the true one. Computing a confidence interval for this p value would lead to show that p is within [0; 0:036] with a 95% confidence level, so this seems to be a really significant signal!

# Appendix 2 (Additional file 7): Competitor methods

### A short description of the use of each method

In our study, the parameters have generally been set to their default values when using the various methods.

Here we explain how the results are obtained in every method.

**KNN-MDR**

The best combinations, containing (10 or 20 or 30) SNPs, are selected as the solution.

The raw power corresponds to the power obtained when the best combination(s) is (are) significant, regardless of if it (they) contains or not the causal SNPs.

The corrected power is the power obtained when the best combination is significant and contains the causal SNPs.

**MegaSNPHunter**

There are four main parameters in the models, including the depth of trees, the threshold for selecting SNPs from trees, the subgenome size and the overlap between subgenome.

1. The depth of trees indicates the depth of SNP interaction. Since most significant interactions are depth 2, so as long as the depth of trees is above 2, the results would not be changed. MegaSNPHunter uses 5 as default setting.

2. The size of subgenome depends on the density of SNP data. Each subgenome should cover the genomic area of possible haplotype effects in practical. Before we start the experiment, we collect some statistics on how many SNPs are genotyped for one gene. This number will be used as the size of subgenome.

3. The overlap between subgenomes is used to solve the boundary problem between genes. Half of the size of subgenome is the best choice. Both the size of subgenome and the overlap between subgenomes depend on the priori knowledge on epistatic interactions.

4. The threshold for selecting SNPs from trees is a very critical parameter to the method.

MegaSNPHunter could rank the importance of SNPs in each subgenome. A cut-off threshold can be used to choose the top ones. The selected SNPs from all subgenomes will first merge together and then compete with each other in the same way at the next level. By having all SNPs compete with each other in training classifiers, MegaSNPHunter reduces the large number of relevant SNPs into a very small set.

The small set contains between 10 and 40 SNPs.

The raw power corresponds to the power obtained when this small set is significant, regardless of if it contains or not the causal SNPs.

The corrected power is the power obtained when the small set is significant and contains the causal SNPs.

**AntEpiSeeker**

In AntEpiSeeker, a two-stage design of ACO (Ant Colony Optimization) is proposed. The first stage of AntEpiSeeker searches SNP sets of sufficient size (larger than the number of SNPs in a given epistatic interaction) using the above ACO, which results in a pre-defined number of highly suspected SNP sets determined by $\chi 2$ scores, and another SNP set of a pre-defined size, determined by pheromone levels. The second stage of AntEpiSeeker conducts exhaustive search of epistatic interactions within the highly suspected SNP sets, and within the reduced set of SNPs with top ranking pheromone levels.

For our comparison, we took suspected SNP sets containing more or less 30 SNPs.

The raw power corresponds to the power obtained when this suspected set is significant, regardless of if it contains or not the causal SNPs.

The corrected power is the power obtained when the suspected set is significant and contains the causal SNPs.

**BOOST**

This method examines all two-locus interactions in a screening stage and the ones over a user-specified threshold are then tested in the testing stage.

In the testing stage, two statistic tests, i.e., likelihood ratio test and chi-squared test are conducted to determine whether the interactive effect of a SNP pair is significant.

For our comparison, we took the first 20 SNPs pair, leading to more or less 30 different SNPs.

The raw power corresponds to the power obtained when this SNP set is significant, regardless of if it contains or not the causal SNPs.

The corrected power is the power obtained when the SNP set is significant and contains the causal SNPs.

## Appendix 3 (Additional file 8): Computing multi-locus penetrances.

Computing multilocus penetrances leading to the absence of marginal effects. This document describes the procedure when 2 interacting markers are considered. It can be extended to situations where more than 2 markers interact. We will use the following notations:

- is the disease prevalence,
- denotes a "multilocus genotype" where the first locus (noted ) has genotype i and the second locus (noted ) has genotype j,
- is the frequency of the multilocus genotype ,
- is the "multilocus penetrance", which is the probability that an individual carrying the multilocus genotype be affected.

An objective in the simulations is to choose the multi-locus penetrances to obtain no marginal effect at either locus:

for all i and j.

To that end, we can write:

In this expression, the are the unknown penetrances and the are known constants, dependent on the 2 loci involved in the interaction. The penetrance of any genotype can then be expressed as a function of the other genotypes frequencies and penetrances as:

where the double sum corresponds to the double sum given above, but excluding the multilocus genotype . It is assumed that is different from 0 (when = 0, no prevalence needs to be computed for that genotype). Note that:

- The maximum value of (written ) is obtained when all other penetrances are equal to their minimal value (written ):

  In the absence of other constraints, these minimal values are equal to 0, which leads to a maximum value of:

  .

When this value is larger than 1, the maximum penetrance is 1.

- The minimum value of (written ) is obtained when all other penetrances are equal to their maximal value (written ):

In the absence of other constraints, these maximal values are equal to 1, which leads to a minimum value of:

.

When this value is lower than 0, the minimum penetrance is 0.

**Example**: assume a prevalence of 0.20 and the following frequencies configuration:

| **A\B** | **0** | **1** | **2** | **Total** |
|---|---|---|---|---|
| **0** | 0.04 | 0.32 | 0.14 | 0.50 |
| **1** | 0.03 | 0.15 | 0.12 | 0.30 |
| **2** | 0.03 | 0.13 | 0.04 | 0.20 |
| **Total** | 0.10 | 0.60 | 0.30 | 1.00 |

-----------------------------------------------------------------------------------------------------

The algorithm to obtain the multilocus penetrances leading to the absence of marginal effects is based on this last formula. It proceeds as follows:

1. Set a range of allowable values for each genotype penetrance. This can be done using the formula above.

**Example (continued)**: this leads to the following tables of minimal and maximal prevalences:

| $p_m$ | **0** | **1** | **2** |
|---|---|---|---|
| **0** | 0.000 | 0.000 | 0.000 |
| **1** | 0.000 | 0.000 | 0.000 |
| **2** | 0.000 | 0.000 | 0.000 |

| $p_M$ | **0** | **1** | **2** |
|---|---|---|---|
| **0** | 1.000 | 0.625 | 1.000 |
| **1** | 1.000 | 1.000 | 1.000 |
| **2** | 1.000 | 1.000 | 1.000 |

-----------------------------------------------------------------------------------------------------

2. Next, the algorithm iterates over each genotype, considering first the ones with the smallest allowable penetrance range. A penetrance value is then randomly chosen for the selected genotype, and the minimal and maximal penetrance for that genotype is set to that value. The range for the remaining genotypes are then recomputed using the minimal and maximal values as described above.

**<u>Example (continued)</u>**: the smallest range is for genotype . Assume a value of 0.5 has been "randomly" chosen for that penetrance. If we want to compute, for example, , the formula becomes:

and for , the formula becomes:

Consequently, the range of the allowable penetrances for this genotype is left to [0,1]. The ranges for the other genotypes are computed similarly, leading to:

| **$p_m$** | **0** | **1** | **2** |
|---|---|---|---|
| **0** | 0.000 | 0.500 | 0.000 |
| **1** | 0.000 | 0.000 | 0.000 |
| **2** | 0.000 | 0.000 | 0.000 |

| **$p_M$** | **0** | **1** | **2** |
|---|---|---|---|
| **0** | 1.000 | 0.500 | 0.286 |
| **1** | 1.000 | 0.267 | 0.333 |
| **2** | 1.000 | 0.308 | 1.000 |

Based on this table, the next penetrance to be set is for genotype (1,1).

**----------------------------------------------------------------------------------------------------**

The algorithm ends after all genotypes have been considered and all penetrances have been obtained. If the procedure fails (because no allowable value remains for one or several genotypes when using the sampled penetrance values for the preceding genotypes in the algorithm), the procedure can be restarted to obtain new penetrance values until a complete set of multilocus penetrances have been obtained.

## Appendix 4 (Additional file 12): Definitions of some terms used within the text

True positive rate (TP) = the proportion of subjects correctly identified as affected (“case”).

False positive rate (FP) = the proportion of subjets incorrectly identified as affected (“case”).

True negative rate (TN) = the proportion of subjects correctly identified as healthy (“control”).

False negative rate (FN) = the proportion of subjects incorrectly identified as healthy (“control”).

Sensitivity (SE): probability that an affected subject is classified as a case by the method.,Mathematically:SE = TP/(TP + FN)

Specificity (SP): probability that a healthy subject is classified as a control by the method. Mathematically: SP = TN/(TN + FP)Accuracy (AC): probability that subjects are classified correctly. Mathematically: AC = (TP + TN)/(TP + TN + FP + FN)

Power: probability that the best combination is significant. In the literature, this value is most often referred to as “rate of positive predictions” and is of limited utility because this rate integrates significant combinations not containing the causal mutations.

Corrected power: probability that the best combination is significant and contains the causal SNPs. In the literature, this power is often called “the true statistical power”.