

Stable feature selection for multi-locus Genome-Wide Association Studies

Asma Nouira

▶ **To cite this version:**

Asma Nouira. Stable feature selection for multi-locus Genome-Wide Association Studies. Bioinformatics [q-bio.QM]. Université Paris sciences et lettres, 2022. English. NNT : 2022UPSLM024 . tel-03850681

HAL Id: tel-03850681
https://pastel.hal.science/tel-03850681

Submitted on 14 Nov 2022

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

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

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à MINES ParisTech

Stable feature selection for multi-locus Genome-Wide Association Studies

Sélection stable de variables pour les études d'association génome entier

Soutenue par

Asma Nouira

Le 13 Juillet 2022

École doctorale n°621

Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Bio-informatique

Composition du jury :

Christophe Ambroise Professeur, Université d'Évry Val d'Essonne	*Président*
Joseph Salmon Professeur, Université de Montpellier	*Rapporteur*
Nataliya Sokolovska Maîtresse de conférence, Université de Sorbonne	*Rapporteuse*
Marylyn Ritchie Professeur, Université de Pennsylvanie	*Examinatrice*
Chloé-Agathe Azencott Maîtresse assistante, MINES Paristech	*Directrice de thèse*

MINES ParisTech | PSL

Acknowledgements

This thesis has been achieved after a work of three years presented in this dissertation. This research will not be done without a great deal of support and assistance of many people.

I would like to start by thanking my supervisor **Chloé**. Thank you for your direction during these years and for the big amount of help you offered. I was very lucky to be supervised by you and learn some of your impressive expertise in GWAS and Machine Learning. Thank you for always finding the time to correct and improve the slides, the reports, the papers and even the posters of our research work. I appreciate your sense of detail. Also, thank you for your kindness, your trust and your support in stressful moments.

Secondly, my thanks are directed to the other members of GWAS team of CBIO: **Héctor**, **Vivien**, **Lotfi**, **Adeline** and **Gwenäelle**. Thank you for the scientific discussions that helped me a lot to accomplish this work, for our random chats and jokes and for all informations and knowledge you shared with me.

I am very grateful to be a part of the CBIO team, I deeply enjoyed the supportive and friendly environment. The divergence of our research topics enriched a lot my knowledge in different bioinformatics fields. I will deeply miss the CBIO meetings, the Book club sessions and especially lunches in the canteen, it was always a pleasure to share a cup of coffee for a short break with them. I would like to thank all my lab mates for making my PhD life enjoyable: **Benoit**, **Peter**, **Joe**, **Judith**, **Romain**, **Vivien**, **Arthur**, **Maguette**, **Matthieu N.**, **Tristan**, **Elise**, **Marc**, **Vincent**, **Mélanie**, **Thomas D.**, **Anne**, **Gwen**, **Philippe**, **Matthieu C.**, **Thomas B.** and all the newcomers.

I would like to acknowledge **Thomas**, the director of CBIO, for his support, advices and help with the administrative procedures. He always find the time to answer my questions even when he do not have any. Also, thank you **Véronique** for all the enjoyable discussions we shared, **Florian** for your deep collaborative sense and your precious help.

I will not forget to thank all the members of U900 in Institute Curie for the constructive environment they offered. I want to extend my special thanks to **Emmanuel Barillot**, the director of the unit, **Caroline Belliere Dahan** and **Christine Lonjou**.

My warmest thanks goes to my family for all support. This dissertation is dedicated to you.

To my mother, thank you for your unconditional love, your advices and your positivity. Thank you for being a friend before being a mom, you are an example of the strong women that I would like to be.

To my father, thank you for always believing in me even when I did not believe in myself. Thank you for showing me what a hardworking successful person looks like, I respect how much you gave to me.

To my brother and my partner in crime during my childhood, thank you for always being there for me. You are the most generous and supportive brother anyone could ask for.

To my awesome sister, thank you for always listening with attention to all my thoughts and random chats. The friendly connection we share is very valuable to me.

To my husband, thank you for all the times you shared your strength with me. Thank you for being a part of this PhD adventure, for supporting me and bringing positivity into our daily life. I look forward to welcoming other adventures with you.

Contents

List of Figures **9**

List of Tables **14**

1 Context **1**
1.1 Introduction . . . 3
1.2 Genome-Wide Association Studies . . . 4
1.2.1 Association analysis . . . 4
1.2.2 Missing heritability . . . 7
1.2.3 The curse of dimensionality . . . 8
1.2.4 Population stratification . . . 9
1.2.5 Linkage disequilibrium . . . 9
1.2.6 Microarray data . . . 11
1.3 Feature selection methods . . . 12
1.3.1 GWAS as a Machine Learning problem . . . 12
1.3.2 Lasso . . . 13
1.3.3 Group Lasso . . . 13
1.3.4 Sparse group Lasso . . . 14
1.3.5 Multitask Lasso . . . 15
1.3.6 Elastic Net . . . 15
1.4 Stability of the feature selection . . . 16
1.4.1 Major problem of feature selection models . . . 16
1.4.2 Measurement of the stability of the selection . . . 16
1.4.3 Desirable properties . . . 16
1.5 Data studied in this thesis . . . 17
1.5.1 Breast cancer datasets . . . 17
1.5.2 Wellcome Trust Case Control Consortium 1 . . . 18
1.5.3 Arabidopsis thaliana . . . 20
1.5.4 Simulated data . . . 20
1.6 Contributions . . . 21

2 Population stratification adjustment in case-control studies for Genome-Wide Association Studies **23**
2.1 Introduction . . . 25
2.2 Correcting associations analysis . . . 26
2.2.1 Genomic control . . . 26
2.2.2 PCA-based methods . . . 26
2.2.3 Linear Mixed Models (LMM) . . . 30
2.3 Population structure and Linkage Disequilibrium . . . 30

2.4 Data and implementation details . . . 31
2.4.1 Data . . . 31
2.4.2 Preprocessing and quality control . . . 32
2.4.3 Implementation details . . . 35
2.5 Results . . . 35
2.5.1 LD pruning helps to capture the population structure in ROOT and DRIVE datasets Principal Components 35
2.5.2 Population stratification adjustment methods decrease the inflation factor . . . 36
2.5.3 Performance of adjustment methods in correcting for population stratification under simulated data . . . 36
2.5.4 Population stratification adjustment in real data . . . 42
2.6 Discussion and conclusion . . . 45

3 Multiscale genomic evaluation of the stability of the selection for Genome-Wide Association Studies 47
3.1 Introduction . . . 50
3.2 Methods . . . 51
3.2.1 Association analysis and feature selection models . . . 51
3.2.2 Measuring the stability at different genomic scales . . . 52
3.2.3 Linkage Disequilibrium blocks clustering . . . 54
3.2.4 FUMA for functional mapping and annotation of genes 55
3.2.5 Stability selection . . . 55
3.2.6 Related work . . . 58
3.3 Experiments . . . 59
3.3.1 Data . . . 59
3.3.2 Preprocessing . . . 59
3.3.3 Implementation details . . . 59
3.4 Results . . . 60
3.4.1 Clustering the SNPs to LD-blocks and mapping the SNPs to genes . . . 60
3.4.2 The stability of the selection in classical GWAS . . . 60
3.4.3 Lasso and Elastic Net lead to better biological interpretation for biomarker discovery . . . 61
3.4.4 Stability selection methods increase the stability index of Lasso . . . 63
3.5 Discussion and conclusion . . . 64

4 Multitask group Lasso for Genome-Wide Association Studies in diverse populations 69
4.1 Introduction . . . 72
4.2 Methods . . . 73
4.2.1 Population stratification . . . 73

4.2.2 Linkage disequilibrium groups . . . 74
4.2.3 Multitask group Lasso . . . 74
4.2.4 Stability selection . . . 77
4.3 Experiments . . . 77
4.3.1 Data . . . 77
4.3.2 Preprocessing . . . 78
4.3.3 Comparison partners . . . 78
4.4 Results . . . 79
4.4.1 MuGLasso draws on both LD-groups and the multitask approach to recover disease SNPs . . . 79
4.4.2 MuGLasso provides the most stable selection . . . 80
4.4.3 MuGLasso selects both task-specific and global LD-groups 81
4.5 Discussion and Conclusions . . . 81

5 Sparse multitask group Lasso for Genome-Wide Association Studies in diverse populations 85
5.1 Introduction . . . 88
5.2 Methods . . . 89
5.2.1 Population structure . . . 89
5.2.2 Linkage disequilibrium groups clustering . . . 90
5.3 Sparse multitask group Lasso . . . 90
5.3.1 Notations . . . 90
5.3.2 Related work . . . 90
5.3.3 General framework and problem formulation . . . 91
5.3.4 Gap safe screening rules . . . 92
5.4 Experiments . . . 92
5.4.1 Data . . . 92
5.4.2 Preprocessing . . . 93
5.4.3 Comparison patterns . . . 96
5.5 Results . . . 96
5.5.1 SMuGLasso and MuGLasso rely on both LD-groups and the multitask approach to recover disease SNPs . . . 96
5.5.2 SMuGLasso and MuGLasso outperform the other methods in terms of stability . . . 98
5.5.3 The selection of both task-specific and shared LD-groups 98
5.6 Discussion and conclusion . . . 99

6 Conclusions and perspectives 103
6.1 Introduction . . . 104
6.2 Chapters summary . . . 104
6.3 Future of GWAS . . . 106
6.4 Final thoughts . . . 109

A GWAS data **111**
A.1 GWAS data . . . 111
A.1.1 1000 Genome Project . . . 111
A.1.2 International HapMap Project . . . 111

B PLINK files format **115**

C Lasso and Elastic Net stability evaluation for the phenotypes T1D and T2D **117**
C.1 State-of-the-art stability of the selection methods . . . 117
C.2 Results of Lasso for T1D phenotype . . . 118
C.3 Results of Lasso for T2D phenotype . . . 119
C.4 Results of Elastic Net for T1D phenotype . . . 121
C.5 Results of Elastic Net for T2D phenotype . . . 122

D Multitask group lasso (MuGLasso) supplementary materials **125**
D.1 Data availability . . . 125
D.1.1 Simulated data . . . 125
D.1.2 DRIVE . . . 126
D.2 Supplementary Methods . . . 127
D.2.1 LD groups across populations . . . 127
D.2.2 Multitask group lasso . . . 127
D.2.3 Gap safe screening rules . . . 128
D.2.4 Measuring selection stability . . . 129
D.3 Supplementary Results . . . 130
D.3.1 PCA of the genotypes . . . 130
D.3.2 Runtimes . . . 130
D.3.3 Breast cancer risk loci detected by MuGLasso on DRIVE 132

E Sparse Multitask group Lasso (SMuGLasso) supplementary materials **135**
E.1 Population stratification adjustment in Arabidopsis thaliana dataset . . . 135
E.2 Breast cancer risk loci detected by SMuGLasso and MuGLasso on DRIVE . . . 136
E.3 DTF3 loci detected by SMuGLasso and MuGLasso on Arabidopsis thaliana dataset . . . 136

Bibliography **139**

List of Figures

1.1 An example of a Manhattan plot for DRIVE dataset presented in Section 1.5.1. 6
1.2 An example of a Q-Q plot for DRIVE dataset presented in Section 1.5.1. 7
1.3 An example of PCA plot capturing the population structure in the 1000 Genome Project data. Five ancestral populations were detected using the first three Principal Components (PCs): African, Hispanic, East-Asian, Caucasian and South Asian . . 10
1.4 LD blocks of CYP7A1 gene in different populations (CEU, YRI, JPT and CHB) from the HapMap data. Bright red color corresponds to very strong LD, white color to no LD, pink red and blue to intermediate LD [Nakamoto *et al.*(2006)] 11

2.1 Q-Q plots obtained for three simulated datasets (no PS, **moderate** PS, **strong** PS) before population stratification adjustment 33
2.2 For ROOT and DRIVE datasets, PCA plots before LD pruning 36
2.3 PCA for ROOT dataset after performing LD pruning of $r^2 > 0.1$, colors coding corresponds to subpopulations and symbols denotes races presented in Table 2.2. AfAm is African American, AfBB is African Barbadian, AfNG is African from Nigeria 37
2.4 PCA plots on DRIVE on the left and simulated data on the right 38
2.5 Q-Q plots obtained for DRIVE dataset before adjustment and after adjustment using the following methods: Genomic control(GC), EIGENSTRAT, Logistic Regression with top PCs as covariates (LogReg1), Logistic Regression for phenotype adjustment (LogReg2) and FastLMM. 43
2.6 Q-Q plots obtained for ROOT dataset before adjustment and after adjustment for population stratification using the following methods: Genomic control(GC), EIGENSTRAT, Logistic Regression with top PCs as covariates (LogReg1), Logistic Regression for phenotype adjustment (LogReg2) and FastLMM. . 44

3.1 For Lasso, number of selected SNPs, LD-blocks and genes against values of lambda 63
3.2 For Lasso, the average error and stability index for different values of lambdas 64
3.3 For Lasso, the stability index at different genomic scales (SNP, LD-block and gene levels) against the average error 65

3.4 For Elastic Net, the average error and stability index for different values of lambda 66
3.5 For Elastic Net, number of selected SNPs, LD-blocks and genes against the values of lambda 67
3.6 For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene levels) against the average error . . 67

4.1 On simulated data, ability of different methods to retrieve causal disease SNPs as a ROC plot (4.1a), and stability index of MuGLasso as a function of the number of bootstrap samples (4.1b). On the ROC plot, the black dot indicates the performance of the stratified GWAS at the Bonferonni-corrected significance threshold. 80
4.2 On DRIVE, runtimes of the different Lasso approaches (4.2a) and stability index of MuGLasso as a function of the number of bootstrap samples (4.2b). 81
4.3 For simulated data, precision and recall of MuGLasso and the stratified approaches on the populations-specific SNPs 84

5.1 PCA plots in Arabidopsis thaliana, we identify 5 subpopulations from 46 countries 94
5.2 K-means clustering for Arabidopsis thaliana 95
5.3 On simulated data, ability of different methods to retrieve causal disease SNPs as a ROC plot 96
5.4 Runtimes of Lasso approaches for simulated, DRIVE and Arabidopsis thaliana datasets 97
5.5 For simulated data, precision and recall of MuGLasso and the stratified approaches on the populations-specific SNPs 102

6.1 GWAS advantages and challenges 109

C.1 For Lasso: number of selected SNPs, LD-blocks and genes against lambdas in T1D phenotype 118
C.2 For Lasso, the average error and stability index for different values of lambdas in T1D phenotype 119
C.3 For Lasso, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T1D phenotype 119
C.4 For Lasso: number of selected SNPs, LD-blocks and genes against lambdas in T2D phenotype 120
C.5 For Lasso, the average error and stability index for different values of lambdas in T2D phenotype 121
C.6 For Elastic Net, number of selected SNPs, LD-blocks and genes against lambdas in T1D phenotype 122

C.7 For Elastic Net, the stability index is given for different values of lambda in T1D phenotype 122
C.8 For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T1D phenotype 123
C.9 For Elastic Net, the stability index is given for different values of lambda in T2D phenotype 123
C.10 For Elastic Net, number of selected SNPs, LD-blocks and genes against lambdas in T2D phenotype 124
C.11 For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T2D phenotype 124

D.1 Choice of shared LD-groups choice after adjacency-constrained hierarchical clustering for each population 127
D.2 Multitask group Lasso architecture 127
D.3 PCA for simulated and real datasets 131
D.4 Runtimes of the different Lasso approaches 131

E.1 For Arabidopsis thaliana, Q-Q plots before and after population stratification adjustment 137

List of Tables

1.1 Estimation of missing heritability for several complex diseases [Manolio *et al.*(2009)] 8

2.1 For each simulated dataset, the ratio of cases and controls is given for both subpopulations, the predefined disease loci are presented for all subpopulations in chromosomes 12, 19, 21 and 22 33

2.2 Samples number per subpopulation for ROOT dataset. AfAM is African American and AfBB is African Barbadian 34

2.3 Samples number per country for DRIVE dataset 34

2.4 For each dataset, the chosen number of included PCs is presented for each dataset 35

2.5 For each dataset, the inflation factor is given after population stratification adjustment obtained by the tested methods . . . 39

2.8 Under **moderate** PS simulated data, the following metrics are given: **TR**: true positive, truly selected SNPs; **FP**: false positive, wrongly selected SNPs; **FN**: false negative, wrongly non-selected SNPs; **TN**: true negative, truly non-selected SNPs; **FPR**, false positive rate = $\frac{FP}{FP+TN}$; **FNR**: false negative rate = $\frac{FN}{TP+FN}$; **Precision** = $\frac{TP}{TP+FP}$; **Recall** = $\frac{TP}{TP+FN}$; and **Accuracy**= $\frac{TP+TN}{TP+FP+FN+TN}$ 39

2.6 Under **moderate** PS simulated data, the estimated odds ratios (OR) for 10 predefined causal SNPs. Between parenthesis, the percentage of absolute change from true OR. In green, the less-biased method giving the lowest % absolute change from true OR is highlighted for each SNP. In red, the more-biased method giving the highest % absolute change from true OR is highlighted for each SNP. 40

2.7 Under **strong** PS simulated data, the estimated odds ratios (OR) for 10 predefined causal SNPs. Between parenthesis, the percentage of absolute change from true OR. In green, the less-biased method giving the lowest % absolute change from true OR is highlighted for each SNP. In red, the more-biased method giving the highest % absolute change from true OR is highlighted for each SNP. 41

2.9 Under **strong** PS simulated data, the following metrics are given: **TR**: true positive, truly selected SNPs; **FP**: false positive, wrongly selected SNPs; **FN**: false negative, wrongly non-selected SNPs; **TN**: true negative, truly non-selected SNPs; **FPR**, false positive rate = $\frac{FP}{FP+TN}$; **FNR**: false negative rate = $\frac{FN}{TP+FN}$; **Precision** = $\frac{TP}{TP+FP}$; **Recall** = $\frac{TP}{TP+FN}$; and **Accuracy**= $\frac{TP+TN}{TP+FP+FN+TN}$ 42

3.1 For each dataset, the number of SNPs and their corresponding LD-blocks and genes obtained after the clustering and the positional mapping respectively 60

3.2 For single-marker analyses, the average number of selected SNPs, LD-blocks and genes across all bootstraps 61

3.3 For single-marker analyses, the stability indexes at different genomic scales: SNP level, LD-block level and gene level . . . 61

3.4 For Lasso, the stability indexes at different genomic scales: SNP level, LD-block level and gene level 62

3.5 For Elastic Net, the stability index values at different genomic scales: SNP level, LD-block level and gene level 63

3.6 For Lasso, the stability index values obtained at different genomic scales (SNP level, LD-block level and gene level) after stability selection using [Meinshausen and Bühlmann(2009)] method. In brackets, the number of selected/mapped features is given for each studied level 64

3.7 For Lasso, the stability index values obtained at different genomic scales(SNP level, LD-block level, gene level) after stability selection using [Shah and Samworth(2013)] method. In brackets is given the number of selected/mapped features at each studied level 65

4.1 For each subpopulation of both datasets (simulated and real), LD-groups number is given and the shared LD-groups number after combination 78

4.2 Stability index and number of selected features for different methods, on simulated data 82

4.3 Stability index and number of selected features for different methods, on DRIVE 83

4.4 For MuGLasso, number of selected LD-groups/SNPs, across and per population 83

5.1 For simulated data, number of predefined causal SNPs 93

5.2 For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 12, 19, 21 and 22 93
5.3 For each subpopulation of the studied datasets (simulated, DRIVE and Arapidopsis thaliana) LD-groups number is given and the shared LD-groups number after combination across subpopulations 95
5.4 Stability index and number of selected features for different methods, on simulated data 98
5.5 Stability index and number of selected features for different methods, on DRIVE 99
5.6 Stability index and number of selected features for different methods, on Arabidopsis thaliana 100
5.7 For SMuGLasso, number of selected LD-groups/SNPs, across and per population 100
5.8 For MuGLasso, number of selected LD-groups/SNPs, across and per population 101

A.1 Population samples classification in the 1000 Genome Project data 112
A.2 Population samples and SNPs for genotyping in HapMap 3 release 113

C.1 Satisfied properties for each stability measurement [Nogueira *et al.*(2018)] 117

D.1 For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 2, 12, 19, 21 and 22 125
D.2 For simulated data, number of predefined causal SNPs 126
D.3 The 32 potential breast cancer risk genes within 10kb of loci identified by MuGLasso and not the adjusted GWAS, together with information as to their biological relevance 133

E.1 For Arabidopsis thaliana, inflation factor values are given before and after adjustment for population stratification 136
E.2 For DRIVE dataset, list of risk genes associated with breast cancer selected by SMuGLasso, MuGLasso and Adjusted GWAS. In bold are genes selected by Adjusted GWAS. CEU-specific selected genes are highlighted in blue and YRI-specific selected genes are highlighted in red. The others (in black) are risk genes shared across all populations 137

E.3 For Arabidopsis thaliana dataset and for Adjusted GWAS, SMuGLasso and MuGLasso, list of selected genes associated with DTF3 trait. In bold are the genes selected by Adjusted GWAS. In blue are genes selected for specific populations. The others are shared genes selected across all populations 138

Chapter 1
Context

Abstract: *Genome-Wide Association Studies have been spread over last 15 years to become an interesting approach for biomarker discovery by finding association between the genotype presented with Single Nucleotide Polymorphisms and the phenotype that denotes a particular disease or trait of interest. However, it is necessary to deal with many challenges such as the missing heritability, the curse of dimensionality, population stratification and linkage disequilibrium. Consequently, machine learning techniques have been adapted to address such issues. For example, feature selection models based on regularization terms have been proven to be efficient to identify candidate genes associated with diseases. Unfortunately, these models suffer from the lack of robustness, they are sensitive to small perturbations in the input dataset. Several measurements have been proposed to estimate the stability of the feature selection. Further methods have been suggested to improve the stability in identifying consistently the same features over different input subsamples.*

Résumé: *Les études d'association pangénomiques se sont émergés au cours des 15 dernières années pour devenir une approche intéressante pour la découverte de biomarqueurs en trouvant une association entre le génotype présenté par des polymorphismes nucléotidiques et le phénotype qui correspond à une maladie ou un trait d'intérêt particulier. Cependant, il est nécessaire de faire face à de nombreux défis tels que l'héritabilité manquante, le fléau de la dimension, la structure de la population et le déséquilibre de liaison. Par conséquent, les techniques d'apprentissage automatique ont été adaptées pour résoudre ces problèmes. Par exemple, les modèles de sélection des variables basés sur des termes de régularisation se sont avérés efficaces pour identifier les gènes candidats associés aux maladies. Malheureusement, ces modèles manquent de robustesse, ils sont sensibles aux petites perturbations dans le jeu de données d'entrée. Plusieurs mesures ont été proposées pour estimer la stabilité de la sélection des variables. D'autres méthodes ont été suggérées pour améliorer la stabilité dans l'identification cohérente des mêmes variables sur différents sous-échantillons d'entrée.*

Contents

1.1 Introduction 3
1.2 Genome-Wide Association Studies 4
1.2.1 Association analysis 4
1.2.2 Missing heritability 7
1.2.3 The curse of dimensionality 8
1.2.4 Population stratification 9
1.2.5 Linkage disequilibrium 9
1.2.6 Microarray data 11
1.3 Feature selection methods 12
1.3.1 GWAS as a Machine Learning problem 12
1.3.2 Lasso 13
1.3.3 Group Lasso 13
1.3.4 Sparse group Lasso 14
1.3.5 Multitask Lasso 15
1.3.6 Elastic Net 15
1.4 Stability of the feature selection 16
1.4.1 Major problem of feature selection models 16
1.4.2 Measurement of the stability of the selection 16
1.4.3 Desirable properties 16
1.5 Data studied in this thesis 17
1.5.1 Breast cancer datasets 17
1.5.2 Wellcome Trust Case Control Consortium 1 18
1.5.3 Arabidopsis thaliana 20
1.5.4 Simulated data 20
1.6 Contributions 21

1.1 Introduction

Since the accomplishment of the human genome project, many studies have shown that the risk of diseases development can be explained from the human genome. The progress of genetic research allows then the identification of common genetic factors associated with a disease.

Variations in the genome are called Single Nucleotide Polymorphisms (SNPs) and they represent the most of the genetic material between individuals. However, this task is complex as the human genome contains around 15 million Single Nucleotide Polymorphisms (SNPs) [Tak and Farnham(2015)]. To link variation in the human population to the risk of disease, researchers have developed Genome-Wide Association Studies (GWAS) which find association between the human genome (the genotype) and a studied disease (the phenotype) by comparing data from people suffering from the disease (cases) and healthy people (controls). This design is defined as case-control study. Unfortunately, datasets used in GWAS analysis usually contain a huge number of features (SNPs) compared to the number of participants. This problem is known as the curse of dimensionality and implies the statistical power of classical GWAS analysis that has remained limited because of the small number of samples. High-dimensional data therefore requires the use of appropriate methods.

In addition, in the case of complex diseases, SNPs identified by GWAS do not always provide the whole information about the phenotype variability. This problem is known as the missing heritability in genomic studies [Manolio *et al.*(2009), Zuk *et al.*(2012), Nolte *et al.*(2017)].
In order to find adapted solutions, recent studies [Bermingham *et al.*(2015)] were oriented towards the exploration of feature selection models such as Lasso to reduce the dimensionality of data by keeping only the relevant features associated with disease and excluding irrelevant associations. Consequently, many contributions have been developed in the last 15 years suggesting several improvements for feature selection procedure by proposing efficient designs dedicated for GWAS analysis. One way to improve feature selection for GWAS analysis is to add prior knowledge about biological environment in a graph or a group structure. On the one hand, feature selection models can incorporate connectivity in graph constraints from biological networks, in addition to regularization terms [Azencott *et al.*(2013)]. On the other hand, Linkage Disequilibrium (LD), presented by high correlation between nearby SNPs, can be incorporated also in feature selection models based on group structure such as the group Lasso [Yuan and Lin(2006), Jacob *et al.*(2009)]. Indeed, SNPs in strong LD can be clustered in groups [Ambroise *et al.*(2019)]. However, despite the promising results of feature selection models, they are still sensitive to the slightest variation in the input dataset and lack of stability [Haury *et al.*(2011)]. To evaluate the robustness of these

methods, several stability of the selection measurements were proposed to quantify the variability that occurs down to the smallest perturbation in the dataset [Kalousis *et al.*(2007), Kuncheva(2008), Lustgarten *et al.*(2009), Wald *et al.*(2013), Nogueira *et al.*(2018)]. Unfortunately, disappointing results show that although best hyper parameters was chosen to obtain the highest stability index, feature selection methods still lack of robustness [Haury *et al.*(2011)].

Further works have proposed new procedures of combining feature selection along with subsampling. These methods are known as, stability selection, and aim to improve the stability of the selection of any existing sparsity enforcing method [Meinshausen and Bühlmann(2009), Shah and Samworth(2013)].

In this chapter, we start by defining fundamental biological notations about GWAS and its challenges in Section 1.2. Then, we present in Section 1.3 the main Machine Learning models applied to GWAS, and more precisely feature selection methods based on regularization: Lasso which shrinks the less important features to zero using an ℓ_1-norm regularization, Group Lasso which allows some predefined groups of covariates to be jointly selected, Multitask Lasso which provides joint tasks learning procedure and Elastic Net which uses a mixed ℓ_1-norm and ℓ_2-norm penalties. We also detail their loss functions and the use cases of each of them. Section 1.4 details the concept of the stability of the selection quantification. The desirable properties that a stability measure should possess will be presented. Next, we describe in Section 1.5 the datasets that we use to conduct our analysis during this thesis. First, two case-control breast cancer datasets: *DRIVE Breast Cancer OncoArray Genotypes Distribution set* (**DRIVE**) and *CIDR-GWAS of Breast Cancer in the African Diaspora - the ROOT study* (**ROOT**). Second, three case-control datasets of three different diseases from the *Wellcome Trust Case Control Consortium 1 data* (**WTCCC1**): Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). Also, we work with *Arabidopsis thaliana* dataset which represent a genotype plant data composed of five chromosomes. We choose to work on the flowering duration time as a phenotype (DTF3). Finally, we also use simulated datasets generated by `GWAsimulator` software to evaluate the developed methods. In the last section, we outline the contributions achieved in this dissertation.

1.2 Genome-Wide Association Studies

1.2.1 Association analysis

Genome-Wide Association Studies (GWAS) have rapidly developed over the last 15 years, becoming an interesting approach to identify candidate genomic regions associated with complex diseases in human medicine [Visscher *et al.*(2017)]. In other words, the aim of GWAS is to determine the associa-

tion between genotype and phenotype. The achievements of GWAS projects come after the completion of the Human Genome Project with the goal of determining the sequence of individual human genomes. It provides a way to decode the whole human genome with a good mapping at the level of Single Nucleotide Variant (SNVs). A SNV consists of a single base-pair variation at a specific position in the genome. In order to be considered as Single Nucleotide Polymorphism (SNP), a SNV is required to be present at a frequency of at least 1% or more of all chromosomes. A SNP (pronounced *"snip"*) includes two alleles denoted by Adenine [A] Cytosine [C], Guanine [G], or Thymine [T]. These new genetic markers are the most common genetic variations in the genome. The human genome contains 3 billions base pair divided in 22 pairs of autosomes and one pair of sex chromosomes (XX or XY) [WS and JH(2012)]. Autosomes are ordered roughly in relation to their sizes from chromosome 1 to chromosome 22. It was estimated that more than 17 million SNPs in the human genome have been cataloged in the SNP Database dbSNP[1] [Naidoo *et al.*(2011)].

Categorical studies

One common design used in GWAS is case-control studies, where the estimation of a SNP corresponds to variation on the observed phenotype on a binary 0-1 scale. GWAS compare the genotypes of two groups of participants: samples with the phenotype of interest, called cases with a particular disease, and similar samples without the phenotype called controls. The aim of GWAS is to identify SNPs present only in cases or only in control subjects and that are thus said to be associated to the disease.

Statistical test

Traditional GWAS methods are based on single-marker analyses, that consist in conducting an independent statistical test for each marker. For a classic GWAS approach, let us define n to be the number of samples (participants) and p to be the number of features (SNPs), $y = (y_1, y_2, ..., y_n)$ denotes the vector of phenotype values for each sample. The genotype matrix is presented by x_{ij} for an individual i and a SNP j, where $i = 1...n$ and $j = 1...p$.

To perform GWAS analysis, a linear regression model is considered for each SNP from p SNPs:

$$y_i = \beta_0 + \beta_j x_{ij} + \epsilon_i,$$

where β_0 corresponds to the model intercept, x_{ij} is the genotype vector for the SNP j, β_j denotes the SNP j effect and $\epsilon_i \sim N(0, \sigma^2)$ is Gaussian error term. The significance is evaluated SNP by SNP individually. Following a

[1] http://www.ncbi.nlm.nih.gov/projects/SNP/

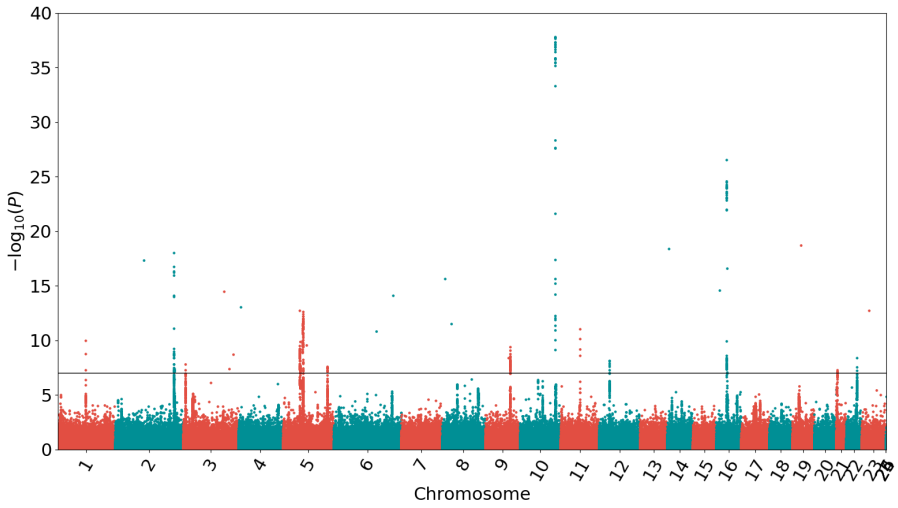


Figure 1.1: An example of a Manhattan plot for DRIVE dataset presented in Section 1.5.1.

t-test or χ^2-test, their p-values are computed against the null model: $H_0 = \beta_j = 0$.

The obtained p-values can be visualized to identify the regions of interest determined by the used statistical test.

On the one hand, Manhattan plots show the p-values of the entire GWAS on a genome scale. The p-values are represented in order by chromosome (X-axis). The Y-axis corresponds to the $-log_{10}$ of the p-values. Each point in the Manhattan plot is a SNP across the human chromosomes from left to right and the heights correspond to the strength of the association with the phenotype (the disease). The strongest associations represented by the peaks in the plot as shown in Figure 1.1.

On the other hand, the Quantile-Quantile plots (Q-Q plots) represent the expected distribution of association test statistics (X-axis) across the SNPs compared to the observed values highlighted in Y-axis. A deviation from $X = Y$ line describes a relevant difference between cases and controls in the genome. In other words, the Q-Q plot is a representation of the deviation of the observed p-values from the null hypothesis. If the observed values are similar to the expected values, all dots are on the diagonal between the X-axis and the Y-axis. Otherwise, if the observed values seem to be relevant than expected, dots in the graph will move toward the Y-axis. An example of a Q-Q plot is presented in Figure 1.2.

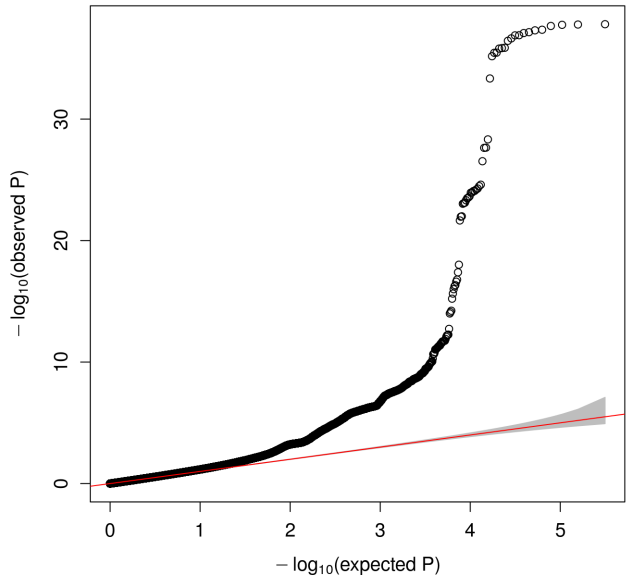


Figure 1.2: An example of a Q-Q plot for DRIVE dataset presented in Section 1.5.1.

1.2.2 Missing heritability

Factors including genetics, environment and random chance can all contribute to the variation between individuals in their phenotypes. Hence, genetic susceptibility to disease depends also on additional environmental risk factors. As an example, some genes are known and identified to play a role in obesity, but they strongly rely on other environmental factors such as smoking withdrawal, pregnancy and antidepressant medication [Mayhew and Meyre(2017)]. Heritability can be defined then as the contribution of genetics to describe the studied phenotype. It is quantified with a genetic measure that finds the observable differences in a trait due to genetic factors between individuals within a population. The heritability in complex diseases can be measured using different statistics such as: Sibling recurrence risk [Rybicki and Elston(2000)], Genetic risk [Jr. *et al.*(2019)] and Phenotypic variance [Byers(2008)]. We give in Table 1.1 the estimation of the heritability explained for some complex disease.

The term "missing heritability" refers to the lack of information about the overall genetic component and risk of common diseases detected from GWAS [Manolio *et al.*(2009)]. GWAS data explain only a modest fraction of heritability. Common variants account for only a small proportion of genetic components, and the missing heritability lies in the huge class of rare genetic variants that GWAS do not consider. Many explanations have been suggested,

missing heritability occurs because of joint genetic effects of common SNPs acting additively. This additive effect is computed as the sum of the effect of each allele at all loci that influence the phenotype.

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Residual phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenotypic variance
Fasting glucose	4	1.5%	Phenotypic variance

Table 1.1: Estimation of missing heritability for several complex diseases [Manolio *et al.*(2009)]

1.2.3 The curse of dimensionality

Designing methods for GWAS analysis must take into account the complexity of the high dimensionality in SNPs microarrays. Hundreds-of-thousands, or even millions, of SNP markers per individual are common in GWAS which implies a huge number of tests to find association with the phenotype of interest. Hence, this leads to a lack of statistical power because of the small samples size. The problem is known as, the curse of dimensionality, i.e., small n, big p. Thus, this leads to mysterious effect and massive computational complexity challenge.

Despite the considerable decrease in sequencing cost and time thanks to geneticists effort, it remains challenging to bridge the large number of SNPs to obtain $n = p$. In fact, the number of possible cases affected with common diseases in a given population is still low and not sufficient to resolve high dimensionality of SNP arrays.

Although Machine Learning community suggested some proposals to alleviate the effect of high dimensionality in genomic data using parallelization models [O'Brien and Szu(2017)], most algorithms still suffer from the lack of

robusteness in identifying causal regions of interest, model over-fitting and local convergence.

1.2.4 Population stratification

In association analysis, population stratification is defined by the presence of different ancestral subpopulations within samples based on their allelic frequency in the same GWAS study. It occurs when the number of samples between cases and controls is different among these subpopulations, this can lead to spurious results especially when the association is found due to the population structure rather than a relationship with the studied disease. The population stratification can also occur in homogeneous populations where individuals belong to the same ancestry or country, as like as heterogeneous populations that include samples from different genetic ancestries.

It is possible to detect whether the data contains a population stratification or not by computing the inflation factor λ. This is computed for each chromosome along the genome as the median of the observed χ^2-test statistics divided by the expected median of the corresponding χ^2 distribution. Empirically, a value larger than 1 implies the presence of population structure confounder. Principal Components Analysis (PCA) can also capture the existence of subpopulations within samples by computing the eigenvectors of the genotype data's covariance matrix. We give in Figure 1.3 an example of PCA plot detecting the population structure of the 1000 Genome Project GWAS data (described in Appendix A.1.1), five populations of different ancestries were captured using the first three Principal Components (PCs). In order to avoid false discoveries due to population stratification issue, we present and compare in Chapter 2 several adjustment methods for case-control phenotype.

1.2.5 Linkage disequilibrium

Alleles of genes on the same chromosome tend to be transmitted together, this is called Linkage disequilibrium (LD). It results in a strong correlation between SNPs in the neighborhood of the same chromosome. LD occurs through recombination events over generations from maternal and paternal chromosomes, it provides information about past event. LD explains how much an allele of one given SNP is correlated and segregated by chance with another allele of another SNP. Many measurments of LD have been proposed in the literature [Devlin and Risch(1995)]. One of the common method relies on using the squared correlation coefficient r^2 that is defined as:

$$r^2\left(f_a, f_b, f_{ab}\right) = \frac{\left(f_{ab} - f_a f_b\right)^2}{f_a\left(1 - f_a\right) f_b\left(1 - f_b\right)},$$

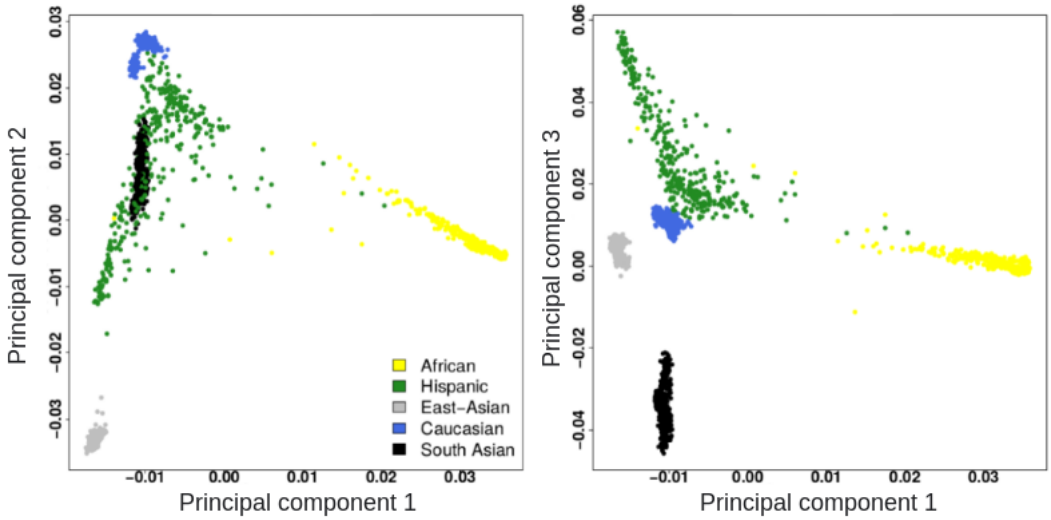


Figure 1.3: An example of PCA plot capturing the population structure in the 1000 Genome Project data. Five ancestral populations were detected using the first three Principal Components (PCs): African, Hispanic, East-Asian, Caucasian and South Asian

where f_{ab} is the frequency of haplotypes having an allele a at locus 1 and an allele b at locus 2. r^2 can be ranged between 0 and 1.

Many factors influence LD such as genetic linkage, population structure, genetic recombination, mutation rate or genetic drift. Population structure as described precedently manifests in LD as differences of the LD regions across population. Indeed, several studies have proved that different ancestral populations present large variation in the LD patterns structure [Nakamoto *et al.*(2006),Teo *et al.*(2009),Park(2019)]. [Nakamoto *et al.*(2006)] have studied LD patterns of human CYP7A1 gene in different populations from HapMap data that we present in Appendix A.1.2. They have shown that the LD-blocks (of strongly correlated SNPs) are different and do not have the same size across the studied populations (CEU, YRI, JPT and CHB). Figure 1.4 exposes an example of their findings in a region of 13 SNPs in the CYP7A1 gene, the Caucasian population (CEU) presents two LD blocks of sizes 14 kb and 2 kb, the African population (YRI) gives two different LD-blocks of sizes 9 kb and 2 kb, the Japanese population (JPT) produces one bigger LD block of 16 kb and finally the Chinese CHB population shows one different LD block of 10 kb size.

In association analysis, the power of the test in detecting the true causal SNPs will greatly rely on the LD strength between the tested SNP and the causal region. In general, the top ranked candidate SNPs are in LD with the real causal ones. Thus, with the growth of SNP arrays, inducing the LD pat-

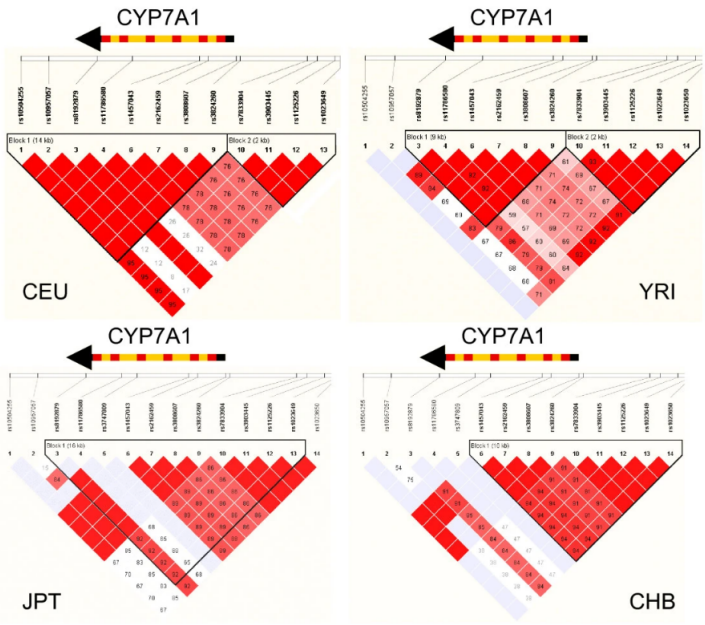


Figure 1.4: LD blocks of CYP7A1 gene in different populations (CEU, YRI, JPT and CHB) from the HapMap data. Bright red color corresponds to very strong LD, white color to no LD, pink red and blue to intermediate LD [Nakamoto *et al.*(2006)]

terns information in GWAS analysis will improve the biological interpretation and alleviate the curse of dimensionality.

1.2.6 Microarray data

In our study, we conduct our analysis using SNP arrays that were manufactured to genotype human DNA at hundred of thousand of SNPs across the whole genome [LaFramboise(2009)]. For each SNP, an individual's genotype is the specific combination of alleles that it possesses. By convention, the most common allele at each SNP is called A and the less common SNP is called B. Therefore, there are three possible pairs of alleles for each SNP: AA, AB and BB. There is a pair of probes for each of the alleles.

The most commonly used SNP array platforms are: Affymetrix platform and Illumina platform. Despite the differences between their technologies, they share the same main components such as their screening arrays contains oligonucleotide probes. Both array technologies call for the hybridization of fragmented single-stranded DNA to arrays containing hundreds of thousands of unique nucleotide probe sequences.

For the genotypes, we use the additive encoding where the minor allele explains the disease prevalence: the major allele in homozygous subjects is

represented by 0, the heterozygous genotype is represented by 1 and the minor allele in homozygous subjects is represented by 2. In practice, PLINK [Purcell *et al.*(2007)] files are widely used for analyzing SNP arrays. Their format are presented in Appendix B.

1.3 Feature selection methods

1.3.1 GWAS as a Machine Learning problem

Thus far, classical GWAS analysis have exponentially grown in power and precision using SNP arrays which cover the whole human genome including millions of SNPs. Despite their advancement, the precision of GWAS approaches remains limited as they are based on statistical tests where the association with the phenotype is evaluated SNP by SNP. In addition, microarrays contain data in high-dimensional spaces with a huge number of features (SNPs) comparing to the number of participants. Such data design requires the exploration of other appropriate methods including notably feature filtering steps.

GWAS have generated a number of important bioinformatics challenges, including the modeling of complex genotype-phenotype relationships using data mining and ML methods. The use of biological knowledge databases helps the interpretation of the genetic association studies by developing powerful models. Therefore, ML models have been developed and successfully applied to avoid the curse of dimensionality in GWAS data. From a ML perspective, the choice of relevant set of attributes is an initial step called feature selection.

A common association study problem can be solved by providing a predictive model for a complex disease in a population from a training dataset of markers genotype and phenotypes of case-control design. The general steps for ML application to GWAS problem are:

- Construction of proper dataset (including preprocessing and quality control analysis)
- Feature selection
- Predictive model construction
- Model validation

In this thesis, we focus mainly in the feature selection step. It remains challenging to construct a predictive model using only GWAS data. Although diseases or some tumor growth are related to the genetic inheritance, there are other factors that explain in higher percentage the affection with a disease, such as environmental factors related to the lifestyle of patients. So far,

the development of an efficient biomarker discovery framework using feature selection models is more valuable than constructing a predictive model.

Therefore, in a GWAS study, the aim is to select features which are fully associated with a given phenotype of interest. Several feature selection algorithms exist within the ML framework.

1.3.2 Lasso

Lasso was first proposed by [Tibshirani(1996)] and it uses an ℓ_1-norm regularization. The model penalizes the coefficients of the regression attributes in a way which shrinks some of them to zero to ensure sparsity. The variables that still have non-zero coefficients are considered selected to be a part of the model. The aim is to minimize the prediction error. Lasso model is the solution to the following equation:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, \beta X) + \lambda \sum_{j=0}^{p} \mid \beta_j \mid,$$

where the penalization parameter λ controls the strength of the penalty. The larger the parameter , the sparser the model. The choice of this parameter has a great importance and it is chosen by cross-validation in practice.

Lasso was succefully applied in GWAS applications. [Wu and Chen(2009)] have evaluated the performance of Lasso penalized with logistic regression in case control design on coeliac disease. In his study, they confirmed their discoveries by retrieving SNPs known for the same disease in previous studies. [Li *et al.*(2010)] later proposed a two-stage procedure by combining supervised PCA with Bayesian Lasso. The approach has shown promising results in identifying body mass index (BMI) SNPs that support previous studies findings. Also, [Waldmann *et al.*(2013)] have conducted a study to analyze GWAS data in different quantitive phenotypes using feature selection model including Lasso. Recently, [Yang and Wen(2020)] have presented a new permutation-assisted tuning procedure in Lasso (called plasso) in order to determine the ammount of shrinkage, the model was succefully applied to real data to find associated SNPs with BMI phenotype.

1.3.3 Group Lasso

The group Lasso deals with problems where the features follow a group structure where it is desirable to yield sparsity (or not) to all coefficients within a group simultaneously [Yuan and Lin(2006)].

For G groups, group Lasso is defined by the following convex optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, \beta X) + \lambda \sum_{j=1}^{G} \sqrt{p_j} \left\| \beta_j \right\|_2,$$

where $\sqrt{p_j}$ scales the penalization factor according to the group size and $\|\beta_j\|_2$ is the Euclidean norm. This approach yields the sparsity criterion at the group level and not among the features within a group.

In a biological context, sometimes features can belong to several groups. The overlap group Lasso [Jacob *et al.*(2009)] allows the features to contribute to more than one group by solving the problem defined as:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, X(\sum_{j=1}^{V} v_j)) + \lambda \sum_{j=1}^{V} \|v_j\|_2 ,$$

where V is a set of groups, and v_j is a group with $v_j \in V_j$.

The loss function is defined by Ω_V as:

$$\Omega_V(\beta) := \inf_{v_j \in V_j} \|v_j\|_2 , \text{where } \beta = \sum_{j=1}^{V} v_j$$

Consequently, this penalty is used for solving the optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, \beta X) + \lambda \Omega_V(\beta).$$

[Silver and Montana(2011)] have proposed to use the group Lasso in which SNPs are grouped into functionally related gene sets or pathways. The authors have shown that the method produces good results in Alzheimer's disease comparing to existing methods. Few studies have used the group Lasso where the groups correspond to SNPs that are in LD [Liu *et al.*(2012), Dehman *et al.*(2015), Ambroise *et al.*(2019)].

1.3.4 Sparse group Lasso

The sparse group Lasso yields sparsity of both groups and features within a group. That is, the model considers two regularizers. The first corresponds to the penalty of group Lasso and enforces sparsity at the group level. The second corresponds to a second penalty that yields sparsity within each group. The group Lasso is defined by the following convex optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, \beta X) + \lambda_1 \sum_{j=1}^{G} \sqrt{p_j} \|\beta_j\|_2 + \lambda_2 \sum_{j=1}^{p} \|\beta_j\|_1 .$$

[Yang *et al.*(2017)] have conducted a GWAS analysis using the sparse group Lasso for both gene groups and network to identify Alzheimer's disease-related risk SNPs.

1.3.5 Multitask Lasso

Multitask Lasso is a derivative method from Lasso, which performs learning with related joint tasks. This setting can be performed for GWAS applications [Sugiyama *et al.*(2014)] as it improves the precision of feature selection by increasing the number of samples among the tasks for multiple phenotypes. In addition, this mapping can be exploited by incorporating one task for each subpopulation within samples for one phenotype.

Several multitask approaches have been proposed [Bellon *et al.*(2016)]. Here, we present the Multitask Lasso proposed by [Obozinski *et al.*(2006)]. The connection between the tasks is given by the ℓ_2-norm regularization, which encourages the shrinkage of coefficients shared between tasks. A ℓ_1-norm regularization term ensures sparsity across all tasks. The aim is to select the same features for all tasks.

We consider T to be the number of tasks to learn and the training set consists of the samples $\{(X^{(tm)}, y^{(tm)})$ for $t = 1...T$ and $m = 1...n_t\}$ where i indexes the i.i.d (independent and identically distributed) samples for each task t. The objective function of the multitask Lasso is then defined as:

$$\min_{\beta \in \mathbb{R}^{T \times p}} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{m=1}^{n_t} \mathcal{L}\left(y^{(tm)}, \left(\beta_0^{(t)} + \sum_{j=1}^{p} \beta_j^{(t)} X_j^{(tm)}\right)\right) + \lambda \sum_{j=0}^{p} \sum_{t=1}^{T} \beta_{tj}^2.$$

1.3.6 Elastic Net

Elastic Net [Zou and Hastie(2005)] is a combination of two approaches: (1) Ridge Regularization using ℓ_2-norm penalty term, and (2) Lasso presented above. The model retains the capacity of feature selection by excluding irrelevant features and replicates the associated features into groups. The loss function of Elastic Net is defined as:

$$\min_{\beta \in \mathbb{R}^p} \mathcal{L}(y, \beta X) + \lambda_1 \sum_{j=0}^{p} \beta_j^2 + \lambda_2 \sum_{j=0}^{p} \mid \beta_j \mid,$$

where λ_1 and λ_2 are the penalization terms that control the strength of both regularization terms.

Elastic Net was successfully applied to GWAS in case-control study to rheumatoid arthritis phenotype [Cho *et al.*(2009)], as well as, to height variation in Korean population [Cho and Kim(2010)].

1.4 Stability of the feature selection

1.4.1 Major problem of feature selection models

In high dimensional data sets, a feature selection procedure is typically applied to obtain a smaller set with reduced dimensionality. Therefore, feature selection is an efficient method to discover the SNPs associated to the trait by removing irrelevant features. In GWAS, it is important to note that the slightest variation in the input dataset leads to different selected subsets by feature selection methods. This is due to the problem of high-dimensional data where we have a small number of samples as compared to the large number of features. Consequently, it is essential to quantify the robustness of feature selection models using an empirical measurement based on several desirable properties. This ensures the selection of meaningful subset of features. In this section, we discuss the concept of the stability of the selection.

1.4.2 Measurement of the stability of the selection

The stability of the selection can be defined as the sensitivity to small perturbations in the input dataset [Kalousis *et al.*(2007)]. Let us repeat a feature selection procedure M-times for M different bootstraps. Doing so, we obtain $\mathcal{Z}$ of M subsets of selected features that we use to measure the stability. The stability is defined by the average of the similarity measure sim between all pairs of subsets. The stability measure is denoted by $\hat{\Phi} : \{0,1\}^{M\times p} \to \mathbb{R}$ and given by:

$$\hat{\Phi}(\mathcal{Z}) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} sim(s_i, s_j),$$

where s_i is i^{th} subset in $\mathcal{Z}$.

Several similarity measurements were proposed relying on a set of desirable properties [Nogueira *et al.*(2018)].

1.4.3 Desirable properties

The observed desirable properties of stability measures in the literature are presented by [Kuncheva(2008), Nogueira *et al.*(2018)] and they are described as follows:

- **Fully defined:** The stability measure should allow the selection of any collection of subsets. Thus, with this property, we avoid returning a constant number of features.

- **Monotonicity:** The stability measure should be a strict decreasing function of the variances of the selection of each feature. [Kuncheva(2008)]

explained that for a fixed subset size, k, and number of features, p, the larger the intersection between the subsets, the higher the value of the stability measure. Therefore, this property ensures that the stability value increases with the size of the intersection of the two sets.

- **Bounds:** The stability measure should be bounded by a constant value which is not related to the number of selected features in the subsets. Some proposed similarities in the literature do not include this property. Then, the stability measure achieves higher value if we select a higher number of features.

- **Maximum:** The stability measure achieves its maximum **if-and-only-if** all selected features are identical within the subsets. Some methods [Wald *et al.*(2013), Somol and Novovicova(2010)], which do not include this property, violate **the forward implication** and the stability measure still achieve its maximum even if we select different features each time. Other methods [Lustgarten *et al.*(2009)] violate **the backward implication**, such that even if the selection is stable and we select the same number of features for all subsets, these methods take different values of stability.

- **Correction for chance:** Under the null model of a feature selection method, the stability measure should be constant. For this property, we assume to correct for random selection by chance which represents a fully unstable case. The stability value in this scenario is supposed to be near to zero, then $E[\Phi] = 0$.

We present in Appendix C.1, the desirable properties for 17 stability indexes proposed in the literature, and for each of the 5 properties, we show which measure satisfies which property.

1.5 Data studied in this thesis

1.5.1 Breast cancer datasets

Breast cancer is the cancer that forms in breast cells tissue. It is the most common diagnosed cancer for women after skin cancer, and it is the second cause of death in women after lung cancer. The tumor usually starts in the inner lining of milk ducts or the lobules that supply them with milk. Breast cancer can affect men, but it is far more common in women. Apart to being women, other risk factors can increase the formation of this disease such as age, obesity, alcohol consumption and the age of first birth. It is also estimated that about 5% to 10% of cases are linked to ihnerited genetic mutation targeting two genes, BRCA1 and BRCA2. Some of their mutations influence the risk

of Breast cancer disease, but they remain unknown.
For this reason, GWAS analysis are involved to identify new breast cancer loci. In this thesis, our analysis are performed using two breast cancer data for case-control phenotype:

DRIVE Breast Cancer OncoArray Genotypes Distribution set

OncoArray dataset [Christopher *et al.*(2016)] is composed of 28 281 individuals that were genotyped for 582 620 SNPs. 13 846 samples from the individuals are cases and 14 435 are controls. The dataset contains data for the following countries: USA, Uganda, Nigeria, Cameroon, Australia and Denmark. In this study, environmental parameters were provided as well such as: age, estrogen rate, study or histological type.

CIDR-GWAS of Breast Cancer in the African Diaspora - the ROOT study

A total of 3 766 study subjects were genotyped on the Illumina HumanOmi2.5-8v1 platform with the GRCh37/hg19 genome build. The genotype data contains 1 681 cases and 2 085 controls. This dataset contains samples for an African population including three different subpopulations as follows: 2 073 American African, 330 African Barbadian and 1 363 African. For each participant, 2 379 855 SNPs were assessed. In addition, this dataset provides information about some environmental variables for all samples participating in the study such as: age group, height, weight, BMI, age of menarche, parity, age of first birth, menopause, age of menopause, alcohol contraceptive and estrogen rate.

The genotype is not the only factor linked to breast cancer risk. The inclusion of these environmental data to the study provides a more complete modeling to this problem.

1.5.2 Wellcome Trust Case Control Consortium 1

The data come from the Wellcome Trust Case Control Consortium 1 (**WTCCC1**) studies [Consortium(2007)]. This study includes seven major diseases from over 2 000 individuals for each disease (cases), 14 000 cases in total and almost 3 000 individuals not affected with the diseases (controls). The participants included in the study live in England, Scotland, and Wales ('Great Britain') and the vast majority identified themselves as white Europeans. This dataset presents a homogeneous population. The genotyping process was performed using Affymetrix 500K chip. In this thesis, we chose to study 3 different phenotypes:

Rheumatoid Arthritis (RA)

Rheumatoid arthritis is a long-term autoimmune disorder that affects lining of the synovial joints [Guo *et al.*(2018)]. Although the cause of rheumatoid arthritis is not clear, it is thought to involve a combination of genetic and environmental factors. The underlying mechanism involves the body's immune system attacking the joints. It was estimated that the disease affected around 24.5 million people in 2015 and resulted in 38 000 deaths in 2013, up from 28 000 in 1990 [GBD2013(2015)].

The dataset contains the mapping of 453 772 SNPs for 3 479 individuals (1 241 males, 2 238 females). 1 999 are cases and 1 480 are controls.

Type 1 Diabetes (T1D)

Diabetes is a prolonged elevation of the concentration of glucose in the blood, called hyperglycemia [Association(2009)]. In the case of type 1 diabetes, this is due to a lack of insulin, a hormone that regulates blood sugar. It is caused by the malfunctioning of T-cells (cells of the immune system) which identify the β-cells of the pancreas as foreign to the patient's body and eliminate them. It is therefore an autoimmune disease. According to the World Health Organization, there were 9 million people with type 1 diabetes in 2019.

The dataset from WTCCC1 contains the mapping of 453 772 SNPs for 3 443 individuals (1 739 males, 1 704 females). 1 963 are cases and 1 480 are controls.

Type 2 Diabetes (T2D)

Type 2 diabetes is caused by a disturbance in carbohydrate metabolism. If it appears gradually and insidiously, the disease has serious, even fatal, consequences in the long term. Hyperglycemia is caused by a decrease in the sensitivity of cells, particularly those in the liver, muscle, and fat tissue, to insulin. The role of this pancreatic hormone is to facilitate the penetration of glucose (their main fuel) into the cells, thus lowering its concentration in the blood. To meet the increased demand for insulin resulting from this insensitivity, the insulin-secreting cells of the pancreas produce more insulin until they run out. Insulin production then becomes insufficient, and glucose accumulates irreparably in the blood. More than 90% of people with diabetes have type 2 diabetes [Cantley and Ashcroft(2015)]. In 2017, type 2 diabetes affects approximetly 462 million people in the world [Khan *et al.*(2020)].

The dataset contains the mapping of 453 772 SNPs for 3 479 individuals (1 903 males, 1 576 females). 1 999 are cases and 1 480 are controls.

1.5.3 Arabidopsis thaliana

Arabidopsis thaliana [Grimm *et al.*(2017)] is one of the most studied genome in plant biology. We use the 1001 Genomes Project set (Build TAIR10) [Weigel and Mott(2009), Consortium(2016)] that contains 1 135 samples for 6 973 565 SNPs divided into 5 chromosomes. We choose to study the flowering time that was scored as days until first open flower (DTF3) as a quantitative phenotype. It contains 923 samples. This dataset groups plants samples coming from 46 countries. We use this dataset as a case of a diverse population data, as it presents a high population stratification inflation factor.

1.5.4 Simulated data

In this thesis, we rely on different simulated datasets to evaluate the developed methods. For real data, the interpretation of the identified candidate SNPs by a given model remains difficult, as we do not have a ground truth about which SNP must be selected. Consequently, the usage of simulated datasets helps to count the false positive rate as the user predefines the disease loci. We use GWAsimulator to simulate retrospectively genotype data for case control samples by following a disease model in a sliding-window algorithm. By using SNP chips as a reference, this tool is able to generate population samples.

For case control design, the program allows the user to choose the disease model parameters such as: disease prevalence, the number of disease loci, its locations, the risk allele, the genotypic relative risks and two-way interaction effects between pairs of disease loci.

Let $x_j = 0, 1, 2$ denotes the number of copies of the risk allele at SNP j and f_j be the risk allele frequency at SNP i. We define r_{i1} and r_{i2} as the risk ratio of genotypes 1, 2 versus 0.

For population simulation, assuming Hardy-Weinberg equilibrium (HWE), the population genotypic frequencies are $Pr(0) = (1-f_i)^2$, $Pr(1) = 2f_i(1-f_i)$ and $Pr(2) = f_i^2$.

For the genotype $X = \{x_1, ..., x_p\}$, the penetrance is given by $f(X) = Pr(\text{affected}|X)$. The software consider the penetrance as a function of genotypes presented as follows:

$$\text{logit}[f(X)] = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where $\text{logit}(g) = \ln[g/(1-g)]$ and the parameters α and β_i are chosen by the user in the input control file.

In order to generate case-control data, using the disease model defined above, GWAsimulator algorithm computes the conditional probabilities $\Pr(X \mid \text{case})$ and $\Pr(X \mid \text{control})$ for all disease loci and samples predefined by the user. For other SNPs, the tool uses a moving-window algorithm to simulate the genotypes assuming that all of them follow Hardy-Weinberg

equilibrium (HWE). GWAsimulator follows also the LD patterns of the population given in the reference data.

In this thesis, we generate different simulations for each study according to our goals and the structure of data we need to conduct our analysis.

1.6 Contributions

The goal of my thesis was the development of a stable framework of feature selection for biomarker discovery in GWAS by dealing with the different challenges in GWAS explained above: the curse of dimensionality, the computational complexity, the population structure, Linkage disequilibrium and the lack of stability of the selection.

The first contribution is presented in Chapter 2, we present several existing population stratification adjustment methods, mainly genomic control, several PCA-based methods and Linear mixed models (LMM). We also compare and evaluate these correction methods in simulated and real datasets.

Chapter 3 focuses on an empirical evaluation of the stability of different widely-used feature selection models: single-marker analyses based on a traditional t-test, Lasso and Elastic Net. The novelty of this work lies on the evaluation of the stability of the selection of these methods at different genomic scales (the SNP level, the LD-group level and the gene level). In addition, to improve the stability in different scales, we implemented two stability selection approaches.

The fourth work (Chapter 4) consists in developing a novel feature selection model, the Multitask Group Lasso (MuGLasso), where the tasks correspond to the ancestral subpopulations and the groups corresponds to the LD-groups. The model relies on an $\ell_{1,2}$-norm regularization term. We incorporate a stability selection procedure to improve the robustness of the model, and we rely on gap safe screening rules in the optimization process to speed up the solvers. The goal was to select shared LD-groups for all subpopulations/tasks and specific LD-groups for some subpopulations/tasks thanks to a post-processing step.

The fifth contribution presented in Chapter 5 is an extension of the MuGLasso, called the Sparse Multitask Group Lasso (SMuGLasso). The particularity of SMuGLasso is to add an ℓ_1-norm penalty to improve the sparsity of specific LD-groups for the subpopulations/tasks. This method is computationally more expensive comparing to MuGLasso, but leads to better interpretable results.

Finally, I made the codes and the algorithms that I developed during my thesis all open online for reproducible science.

The materials of the different contributions presented in this thesis are available in the following github repositories:

- **GWA-skills**: Classic GWAS tools (quality control and preprocessing) and population stratification correction: https://github.com/asmanouira/GWAskills
- **multiscale-stability**: Multiscale genomic evaluation the stability of the feature selection in GWAS: https://github.com/asmanouira/multiscale-stability
- **GWAS-admixed-population-simulator**: Simulating GWAS data in PLINK format with GWAsimulator tool using HapMap 3 data: https://github.com/asmanouira/GWAS-admixed-population-simulator
- **MuGLasso and SMuGLasso**: Multitask group Lasso and Sparse Multitask group Lasso in diverse populations https://github.com/asmanouira/MuGLasso_GWAS

Chapter 2

Population stratification adjustment in case-control studies for Genome-Wide Association Studies

---

Abstract: *Population stratification is one of the major problems in association analysis. It occurs when samples among a dataset come from diverse or admixed populations, presenting genotype data of participants of different ancestries. Thus, the association could be detected due to population structure rather than a true association with the studied phenotype. In addition, different population do not necessary share the same linkage disequilibrium patterns. Many methods have proposed in the literature to adjust for population stratification, such as genomic control, PCA-based models and linear mixed models. In this chapter, we present a deep comparison study of the well-known population stratification adjustment methods. We rely on different simulated datasets and two real breast cancer datasets to conduct our analysis.*

Résumé: *La structure de la population est l'un des problèmes majeurs de l'analyse d'association. Cela se produit lorsque des échantillons d'un ensemble de données proviennent de populations diverses ou mélangées, présentant des données génotypiques de participants d'ascendances différentes. Par conséquent, l'association pourrait être détectée en raison de la structure de la population plutôt qu'une véritable association avec le phénotype étudié. De plus, différentes populations ne partagent pas nécessairement les mêmes structures de déséquilibre de liaison. De nombreuses méthodes ont été proposées dans la littérature pour corriger cette stratification, telles que le contrôle génomique, les modèles basés sur l'ACP et les modèles mixtes linéaires. Dans ce chapitre, nous présentons une étude comparative approfondie des méthodes courantes d'ajustement de la stratification de la population. Nous nous servons de différents jeux de données simulées et sur deux jeux de données réelles de cancer du sein pour effectuer notre analyse.*

Contents

2.1 Introduction 25
2.2 Correcting associations analysis 26
2.2.1 Genomic control 26
2.2.2 PCA-based methods 26
2.2.3 Linear Mixed Models (LMM) 30
2.3 Population structure and Linkage Disequilibrium . . 30
2.4 Data and implementation details 31
2.4.1 Data 31
2.4.2 Preprocessing and quality control 32
2.4.3 Implementation details 35
2.5 Results 35
2.5.1 LD pruning helps to capture the population structure in ROOT and DRIVE datasets Principal Components 35
2.5.2 Population stratification adjustment methods decrease the inflation factor 36
2.5.3 Performance of adjustment methods in correcting for population stratification under simulated data 36
2.5.4 Population stratification adjustment in real data . . . 42
2.6 Discussion and conclusion 45

2.1 Introduction

Population stratification refers to the presence of differences in allele frequencies between subpopulations within the samples, due to different ancestry. The presence of population stratification is one of the major problems in association studies as it increases type I error and leads to ambiguous results. This is particularly true when allele frequency differences in cases and controls are due to differences in ancestry rather than association between SNPs and disease. Several methods have been developed to adjust for population stratification, such as genomic control (Section 2.2.1), PCA-based methods (Section 2.2.2) and Linear mixed models (Section 2.2.3).

To investigate population structure, we chose to use PCA-based methods. They are based on the idea that the first principal components of a set of genomes (that is to say, the eigenvectors of their covariance matrix) map to subpopulations [Price *et al.*(2006), Zeggini *et al.*(2008), Need *et al.*(2009), Yu *et al.*(2008a), Peloso *et al.*(2009), Peloso and Lunetta(2011), Novembre and Stephens(2008), Qizhai and Kai(2008)]

Other methods have been developed based on mixed models [Kang *et al.*(2010), Zhang *et al.*(2010)], where the phenotype is modeled using a mixture of fixed effects and random effects. These fixed effects include the SNPs and additional covariates such as gender or age. The random effects are based on the phenotypic covariance matrix. Nonetheless, PCA-based approaches remain computationally less intensive and provide simpler implementation. In addition, using mixed models often increase the complexity of the problem for the next steps of this study. Indeed, feature selection procedures become more complex using mixed categorical data.

Notations: For the following sections, we denote by x_{ij} the genotype for individual i on SNP j, where $i = 1...N$ and $j = 1...M$. y_i denotes the phenotype of individual i.

In this chapter, we start by presenting in Section 2.2 the most common used methods of population stratification correction in GWAS for case-control study, mainly genomic control, PCA-based models and linear mixed models (LMM). Then, we explain in Section 2.3 the influence of LD at the population structure and the importance of LD-pruning before applying PCA-based models. In Section 2.4, we describe the studied data to conduct our analysis, and we provide details about the implementation and the evaluation of the presented methods for population stratification correction. We present in Section 2.5 the results of the applied methods. Lastly, we discuss and conclude in Section 2.6 the comparison study that we conduct in this first work.

2.2 Correcting associations analysis

In this section, we present the state-of-the art methods adjusting for population stratification in **case-control** studies.

2.2.1 Genomic control

This statistical method has been developed by [Devlin and Roeder(1999)] and it was the first approach proposed for the correction of association analysis. It involves measuring the extent of the inflation of the association test statistics that is caused by population structure and adjusting the test statistics accordingly. In case-control studies, the association between one SNP and the phenotype can be measured using a χ^2 test of association. Under the null hypothesis that there is no association between the SNPs and the phenotype, all association test statistics follow χ^2 distribution with one degree of freedom (denoted χ^2_1). The **genomic inflation factor** λ is defined as the ratio of the median of the observed χ^2 statistics and the expected median of a χ^2_1 and should be equal to 1 under the null. Hence [Devlin and Roeder(1999)] uses a scaling factor of λ to adjust the test statistics:

$$\chi_1^{adj} = \frac{\chi_1^2}{\lambda}.$$

However, this uniform adjustment can lead to overcorrection of causal SNPs because allele frequencies across ancestries can vary from some markers to others. Thus, markers with strong differentiation will suffer from a loss of power [Wu *et al.*(2011), Price *et al.*(2006)].

2.2.2 PCA-based methods

EIGENSTRAT

EIGENSTRAT is a popular software for correcting population stratification, developed by [Price *et al.*(2006)] and implemented in the EIGENSOFT package. The adjustment is based on the top principal components (PCs) computed from the genotype matrix. It uses the eigenvectors (called axes of variation in their article) to correct both the genotype and the phenotype. More specifically, the idea here is to project both the genotypes and the phenotype on a space that is orthogonal to the one spanned by the principal components that capture population structure. EIGENSTRAT is equivalent to including the top PCs as covariates in a linear regression model. A statistical test based on univariate analysis is then performed to the adjusted genotype and phenotype using a generalization of the Armitage trend test [Armitage(1955)] to test whether the correlations between adjusted genotype and phenotype

follow a χ^2_1 distribution.
Let's describe now the components of EIGENSTRAT algorithm:

- **Input format**
  The input of the EIGENSTRAT is a genotype matrix in which the rows are individuals, the columns are SNPs. The cells take values in 0, 1 or 2 which refers to the number of a random selected allele for an individual's genotype at a SNP.

- **Normalization of features**
  The data is normalized by dividing each SNP j by $\sqrt{p_j(1-p_j)}$, where p_j is the allele frequency for a SNP j and defined by:

$$p_j = \frac{1 + \sum_i x_{ij}}{2 + 2M}.$$

- **Calculation of PCs, eigenvectors and eigenvalues**
  Calculation of the covariance matrix C:

$$C_{ij} = \frac{1}{M} \sum_{s=1}^{M} \frac{(x_{is} - 2\hat{p_s})(x_{js} - 2\hat{p_s})}{2\hat{p_s}(1-\hat{p_s})},$$

  where $\hat{p}_s$ is the allele frequency at marker s. The eigenvectors V are then obtained by the spectral decomposition of C as: $V^{-1}CV = D$, where D is the diagonal matrix of eigenvalues of C.

- **Detection of population structure via PCs**
  The key idea behind EIGENSTRAT is that the detection of the ancestry differences between samples is performed using the axes of variations (eigenvectors) as they provide a geographic interpretation. In fact, the range of graduation of the values for an eigenvector is different from a subpopulation to another, depending on the ancestry.

- **Adjustment of genotype and phenotype using the eigenvectors**
  The adjustment is produced continuously depending on the detected population ancestry. Let us call a_{ik} the contribution of individual i to the k^{th} eigenvector and K the total number of eigenvectors considered. The genotype adjustment is given by $x_{ij}^{adj} = x_{ij} - \sum_{k=1}^{K} \gamma_{jk} a_{ik}$ where the regression coefficients are defined by $\gamma_{jk} = \sum_{i=1}^{N} a_{ik} x_{ij}$ and $\sum_i a_{ik}^2 = 1$ if the data do not contain missing values. Hence x_j^{adj} is the projection of SNP x_j onto a space orthogonal to that of the K axes of variation.
  The phenotype correction is performed in the same way.
  Note: This setting is equivalent to the use of the eigenvectors as covariates in a multilinear regression model:

$$y_i = \alpha + \beta_1 a_{i1} + \beta_2 a_{i2} + ... + \beta_K a_{iK} + \eta x_{is} + \epsilon_i,$$

where a_{ik} is the ancestry for an individual i along an axis of variation k, η is the Armitage trend test and used to evaluate $H_0 : \eta = 0$ vs $H_A : \eta \neq 0$.

- **Verification of association using χ^2_1 distribution**
  Finally, the χ^2_1 test statistic is given by:

$$EG = (N - K - 1)Corr^2(x_s^{adj}, y^{adj}),$$

where x_s^{adj} is the adjusted genotype at marker s, and presents the residuals after regressing genotypes on the top PCs K. The phenotype y^{adj} is defined equivalently. $Corr^2$ denotes the statistical test that EIGENSTRAT statistical test EG follows, it corresponds approximately to χ^2 test.

Logistic Regression

Several approaches have been developed using a logistic regression model for the adjustment of population structure in case-control studies [Zeggini *et al.*(2008), Need *et al.*(2009), Yu *et al.*(2008a), Peloso *et al.*(2009), Peloso and Lunetta(2011), Novembre and Stephens(2008), Qizhai and Kai(2008)]. Basically, logistic regression is applied by integrating the top PCs as covariates. The key idea behind this algorithm is to apply a logit function to Generalized Linear Model as the case-control outcome (phenotype) does not follow normal distribution [Wu *et al.*(2011)].

[Need *et al.*(2009)] investigated genotype data in Schizophrenia for case-control study. Their method was performed by using the top PCs coming from EIGENSTRAT software and sex as covariates in a logistic model.

In this study, we will focus on the PCA-L variant described by the following fitted logistic regression:

$$\text{logit}\,(\pi_i) = \alpha + \beta_1 a_{i1} + \cdots + \beta_K a_{iK} + \eta x_{ij},$$

However, in machine learning application, the usage of feature selection models (presented in Section 1.3) by including the top PCs in a feature selection model will not ensure its selection. For this reason, we will also test another variant proposed by [Abegaz and et al.(2021)]. They aim to correct the effect of population stratification by adjusting the phenotype and computing the new one y_i^{adj} by fitting a logistic regression model on the top PCs and subtracting the obtained residuals $\hat{\pi}_i$ from the real phenotype values (1 for cases or 0 for controls). The logistic fitted model on the top PCs $(a_1, \ldots, a_K)$ is defined by the following equation:

$$\text{logit}\,(\pi_i) = \alpha + \beta_1 a_{i1} + \cdots + \beta_K a_{iK}.$$

Then, the adjusted phenotype y_i^{adj} is computed:

$$y_i^{adj} = y_i - \hat{\pi}_i,$$

where the residuals $\hat{\pi}_i$ are obtained as follows:

$$\hat{\pi}_i = \frac{\exp\left(\hat{\alpha} + \hat{\beta}_1 a_{i1} + \cdots + \hat{\beta}_K a_{iK}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\beta}_1 a_{i1} + \cdots + \hat{\beta}_K a_{iK}\right)}.$$

Note that this model has been proposed by analogy with the residual regression model, which is well-known for linear regression. In that case, fitting the features individually is equivalent to fitting the features jointly.

Finally, the quantitative adjusted phenotype is used to conduct association analysis with linear regression.

Choice of the number of principal components

PCA-based methods are widely applied to detect population structure by capturing variation among variables and including PCs as covariates. Numerous studies have examined the choice of the number of PCs to include [Peres-Neto *et al.*(2005)]. Some studies have proposed to use a fixed number for all use cases. For instance, [Price *et al.*(2006)] set a number of PCs equal to 10 by default to run EIGENSTRAT. [Abegaz *et al.*(2018)] and [Pardiñas *et al.*(2018)] mention that a reasonable number of PCs is 5, while other studies use only the 2 first PCs.

[Patterson *et al.*(2006)] have proposed a method based on coupling the Tracy–Widom statistic with PCA method in order to determine the number of components. Nonetheless, [Elhaik(2021)] has claimed that this statistic is sensitive and may produce an overestimated number of PCs.

Another alternative is to pick up the number of PCs that correspond to the number of top eigenvectors. This can be determined according to their corresponding values of eigenvalues that are significant and explain the highest variation of the features. However, [Yu *et al.*(2008a)] have claimed that this technique may lead to ambiguous interpretations, especially if samples between cases and controls are distributed equitably. The authors have suggested another alternative to identify the relevant number of PCs, based on a permutation procedure. The goal is to choose a minimal number of PCs, but results in an efficient adjustment. Note that including a very large number of PCs as covariates may cause numerical instability in association studies [Lin and Zeng(2011)]. But, using fewer PCs than needed may cause in residual bias [Zhao *et al.*(2018)].

In this thesis, we propose to investigate empirically the effect of including different number of PCs on the inflation factor λ. Then, we choose the number of PCs that results in the lowest value of λ close enough to 1. Following this procedure, the chosen number of PCs can differ from one dataset to another.

2.2.3 Linear Mixed Models (LMM)

Linear mixed models (LMM) are widely used for population stratification correction in GWAS [Kang *et al.*(2008), Price *et al.*(2010)]. LMM can capture the population structure by modeling the phenotype with a mixture of fixed effects and random effects. For GWAS, the fixed effects are represented by the genotype (the SNPs) and the covariates, while the random effects correspond to a phenotypic covariance matrix. Thus, a categorical phenotype y_i is presented by the fixed effects x_{ij} (similar to a simple linear model), mixed with the random effect component u as follows:

$$y_i = \eta x_{ij} + u + \epsilon,$$

where $u + \epsilon$ denotes the total noise variance. More precisely, u corresponds to the heritable component of random variation and ϵ corresponds to the non-heritable component of random variation. The component $u + \epsilon$ is modeled with a kinship matrix K.

Note that a kinship matrix is presented by the coefficients modeling the pairwise genetic similarity between samples, its components explain the population structure and other genetic effects such as family structure and cryptic relatedness. The kinship coefficient is the probability that an allele taken randomly from a first population (at a given locus) will be identical by descent to an allele taken randomly from another population at the same locus [Ochoa and Storey(2021)].

The population structure and the relationship between the samples are presented by means of variance components of the random effects:

$$\mathrm{Var}(u) = \sigma_g^2 K.$$

The inflation factor due to population structure can be detected by the coefficient σ_g^2.

LMM are known to be computationally intensive. However, several efficient variants are provided to scale the model for GWAS data, such as FastLMM [Lippert *et al.*(2011)], EMMAX [Kang *et al.*(2010)] and TASSEL [Zhang *et al.*(2010)].

2.3 Population structure and Linkage Disequilibrium

Performing PCA analysis requires some additional steps before computing the eigenvectors using singular value decomposition. These steps include mainly examining the effect of LD in multiple populations.
In fact, PCs can detect the LD patterns instead of the population structure,

adding these particular PCs as covariates can lead to lower power in association analysis [Privé *et al.*(2020)]. One can consider that some regions in the genome are overrepresented by PCs [Abdellaoui *et al.*(2013)] due to the strong correlation of SNPs in LD, decreasing the effect of population structure and the ancestral components.

To address this issue, LD-pruning is common. This procedure consists in removing the SNPs that are in LD by computing the correlation between a pair of SNPs in a window using r^2, and discarding one of the pair if r^2 is greater than a chosen cutoff. The choice of this threshold is critical, [Gusareva and Steen(2014)] recommend setting the filtering threshold to 0.75.

However, [Price *et al.*(2010)] did not recommend pruning because it doesn't affect the top PCs in HapMap populations. They have suggested instead to remove long-range LD regions. Also, [Abdellaoui *et al.*(2013)] have proposed to mix both procedures by discarding long-range LD regions and pruning SNPs in LD as well. In another approach, [Privé *et al.*(2020)] have developed an `R` package called `bigsnpr`, which included a procedure of LD clumping and discarding of long-range LD regions as an optional step. Note that LD clumping has a similar purpose as LD pruning, but it uses a statistical test to compute the p-values of SNPs associated to the phenotype. It takes the first SNP and removes all the SNPs correlated to it in a specific window and with a chosen cutoff for r^2. In the case of PCA, as no p-values are available at this stage, it is recommended to use Minor Allele Frequency (MAF) instead of p-values as the statistic to rank SNPs (in decreasing order), this makes clumping very similar to pruning.

In this thesis, we perform LD pruning to tackle this problem. The choice of the pruning cutoff is estimated by an empirical evaluation according to each dataset case. For some datasets, choosing a less restrictive cutoff is sufficient to resolve the confounding factor caused by LD. Nevertheless, for other datasets an important amount of pruning is needed to be able to capture population structure with PCs. We show examples in Section 2.5.1.

2.4 Data and implementation details

2.4.1 Data

Breast cancer GWAS data

In this chapter, we study two datasets of breast cancer: DRIVE Breast Cancer OncoArray Genotypes Distribution set (**DRIVE**) and CIDR-GWAS of Breast Cancer in the African Diaspora - the ROOT study (**ROOT**). Both of them suffer from population stratification. We have already presented them in Section 1.5.1.

Simulated data

We simulate 3 different case-control datasets using GWAsimulator (described in Section 1.5.4). We rely on HapMap3 [1] data as a reference dataset to obtain two different populations: CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria). The simulated datasets mimic the LD patterns of HapMap3 data for both populations. The simulation procedure of each dataset is performed as given in the following steps:

1. Set the input control files from HapMap3 (for CEU and YRI), and choose mainly the number of samples (females and/or males; cases and controls) and the disease loci.

2. Simulate samples from the two subpopulations (CEU and YRI), varying the case control ratio between samples to model the population stratification confounder.

3. Merge the obtained subpopulations in one dataset and convert it into PLINK format for further analysis.

A step-by-step detailed tutorial was made available online to produce the simulation procedure: `GWAS-admixed-population-simulator`[2].

We generate 3 datasets of 2 000 samples and 503 487 SNPs with different population stratification severity: no population stratification (no PS), moderate population stratification (**moderate** PS) and strong population stratification (**strong** PS). The details of these datasets are presented in Table 2.1. Note that we simulate the dataset no PS only to prove the efficiency of the simulation procedure and to make sure that $\lambda = 1$. Hence, no PS dataset does not require correction for population stratification. In this study, we will examine the adjustment of the other simulation cases (**moderate** PS, **strong** PS).

We show in Figure 2.1 the obtained Q-Q plots for each simulated dataset case.

2.4.2 Preprocessing and quality control

We perform the following quality control steps on the studied data:

Minor allele frequency (MAF) We only keep SNPs with minor allele frequency with MAF>5% because the statistical power is extremely low for rare SNPs.

[1] `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010\T1\textendash05_phaseIII/`

[2] `https://github.com/asmanouira/GWAS-admixed-population-simulator`

Dataset	CEU case:control ratio	YRI case:control ratio	Disease loci location CHR: odds ratio
no PS	500:500	500:500	12: 1.5
moderate PS	450:550	550:450	19: 1
strong PS	400:600	600:400	21: 2 and 22: 2

Table 2.1: For each simulated dataset, the ratio of cases and controls is given for both subpopulations, the predefined disease loci are presented for all subpopulations in chromosomes 12, 19, 21 and 22

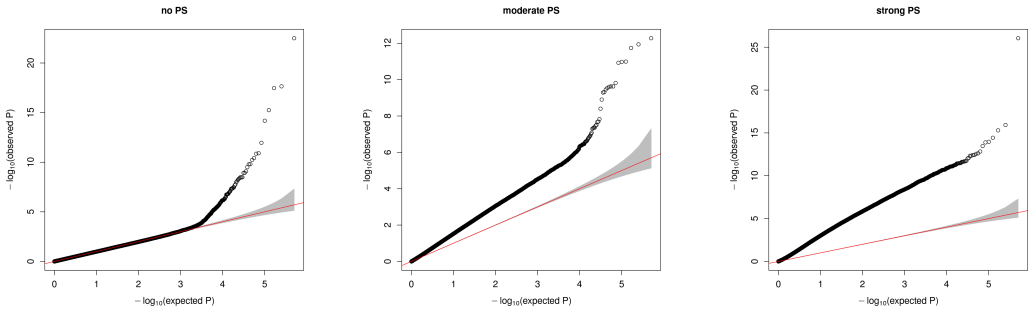


Figure 2.1: Q-Q plots obtained for three simulated datasets (no PS, **moderate** PS, **strong** PS) before population stratification adjustment

Hardy-Weinberg equilibrium (HWE) The main purpose of the HWE test is to identify poorly genotyped SNPs. We exclude SNPs with HWE-P-Value < 0.0001. Under Hardy-Weinberg assumptions, allele and genotype frequencies can be estimated through generations. Thus, the allele and the genotype frequencies are constant over generations once a population is in Hardy-Weinberg equilibrium. Its disequilibrium can be indicative of genotyping errors or population stratification.

Sex checks For ROOT data, we exclude samples with a male genotype as all participants involved in this study are known to be female.

Missing phenotypes Samples with missing phenotype (case or control status) are excluded from the study.

Genotype imputation The datasets presented in Section 2.4.1 contain missing SNPs values. Classic GWAS analysis can deal with non-imputed data as they scan the association between genotype and phenotype SNP per SNP. However, genotype imputation provides higher power and precision to association analysis by increasing the chances of detecting causal variants. In addition, low-dimension analyses such as PCA or machine learning algorithms

	Black	Black/White	Hausa	Ibo	Yoruba	Others	Unknown
African	0	0	11	140	1073	137	2
AfAm	0	0	0	0	0	0	2073
AfBB	312	18	0	0	0	0	0

Table 2.2: Samples number per subpopulation for ROOT dataset. AfAM is African American and AfBB is African Barbadian

USA	Denmark	Australia	Cameroon	Nigeria	Uganda
23 819	2 140	1 693	150	442	62

Table 2.3: Samples number per country for DRIVE dataset

require complete data. Therefore, imputation is a very important step before starting GWAS analysis, and it must be executed with a lot of care.

In practice, we perform imputation with IMPUTE2 [Howie *et al.*(2009)] software. The exploited reference dataset is the 1000 Genome Project(GP) Phase3 [Consortium(2015)], and it provides information about a huge number of SNPs for different ancestries. The method compares phased haplotypes to the reference haplotypes which contain no genotyped markers in the dataset. The term "phased" refers to the statistical estimation of haplotypes from the genotype data. The imputation is then performed according to the given score of probability of possible allele based on the haplotype frequencies.

LD pruning We use PLINK [Purcell *et al.*(2007)] to filter SNPs on strong LD. For the DRIVE dataset and the simulated datasets, the pruning was performed with an LD cutoff of $r^2 > 0.85$ and a sliding window size of 50Mbp. For the ROOT dataset, a cutoff of $r^2 > 0.1$ was needed to capture the population structure in a sliding 20Mbp window.

$\Longrightarrow$ After runing these preprocessing and quality control steps, 313 237 SNPs remain in DRIVE dataset, 262 454 SNPs in ROOT dataset. For simulated data, we obtain 304 605 SNPs, 305 100 SNPs and 304 536 SNPs respectively in no PS, **moderate** PS and **strong** PS datasets.

Population structure We present the outliers of population structure for the studied datasets by running PCA. Table 2.2 shows the ancestries in the ROOT data. Table 2.3 presents the countries of samples included in the DRIVE dataset.

Dataset	Number of included PCs for PCA-based methods
DRIVE	7
ROOT	3
Simulated (**moderate** PS)	2
Simulated (**strong** PS)	2

Table 2.4: For each dataset, the chosen number of included PCs is presented for each dataset

2.4.3 Implementation details

We present in this part the tools that we use in order to implement the presented methods for population stratification adjustment. We run PLINK to obtain the adjusted p-values after the genomic control correction and the first logistic regression PCA method, based on including the top PCs as covariates (**LogReg1**). The second logistic regression PCA method (**LogReg2**), based on phenotype adjustment by computing the residuals using top PCs, was implemented using `scikit-learn`, the computation of the eigenvectors and the obtained p-values was performed using PLINK. To test EIGENSTRAT, we use `EIGENSOFT`[3] that is developed by the authors of [Price *et al.*(2006)].

The choice of the number of included PCs is decided empirically. As presented in Section 2.2.2, we pick up the best number of PCs that adjust for population stratification and provide closest inflation factor λ to 1. We present in Table 2.4 the selected number of PCs for each studied dataset.

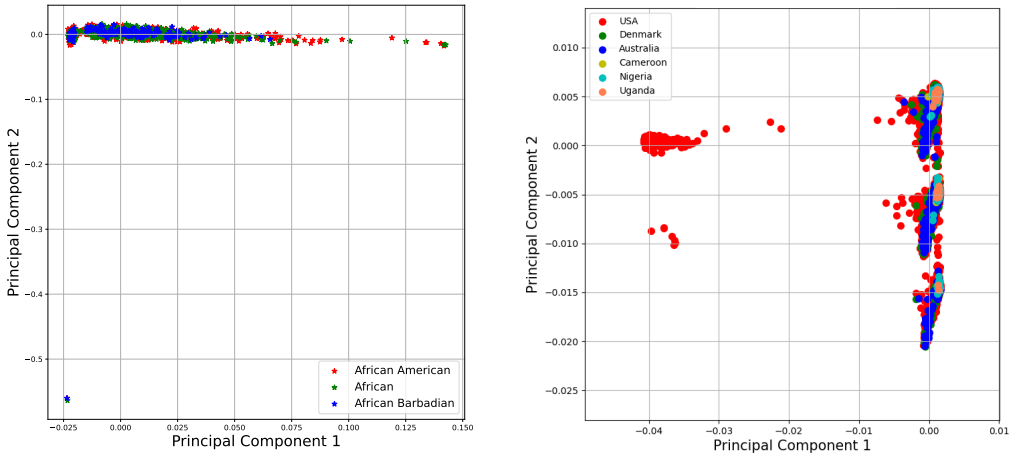
2.5 Results

2.5.1 LD pruning helps to capture the population structure in ROOT and DRIVE datasets Principal Components

In this section, we study the effect of LD pruning on capturing the populations of the studied datasets (see Table 2.3 and Table 2.2). As mentioned in Section 2.3, LD pruning is needed to capture population structure. The chosen parameters of pruning are detailed in Section 2.4.2. We show in Figure 2.2a the two first PCs representation before LD pruning for ROOT dataset, the population structure is not detected and diluted the genetic patterns that describe ancestry differences. In Figure 2.3, we highlight the importance of LD pruning in detecting the population structure. Figure 2.2b and Figure 2.4a illustrate respectively the obtained PCA outliers before and after LD pruning

[3] `https://github.com/DReichLab/EIG`

in DRIVE dataset. For both datasets, we observe that before LD pruning, PCA do not detect the population structure. After performing LD pruning, the outliers of each subpopulations are captured by PCA.



(a) ROOT dataset (b) DRIVE dataset

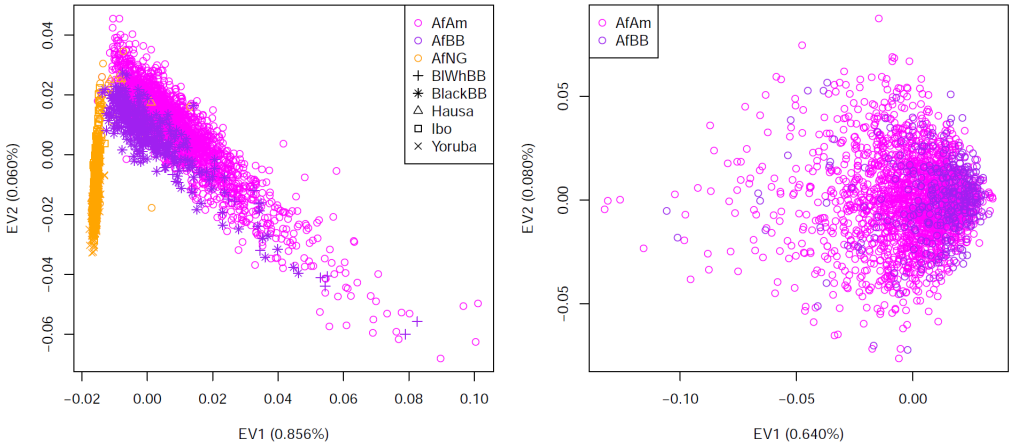
Figure 2.2: For ROOT and DRIVE datasets, PCA plots before LD pruning

2.5.2 Population stratification adjustment methods decrease the inflation factor

The inflation factor is one indicator of population stratification presence, especially when it exceeds 1 as we explained in Section 1.2.4. We show in Table 2.5 the inflation factor computed before and after correcting for population structure. For all studied datasets λ is higher than 1 before adjustment. Then, the table shows the effect of each tested method in correcting the population stratification by decreasing the inflation factor. In DRIVE dataset, we observe that λ decreases, but it remains higher than 1. This happens because this dataset represents a meta-analysis study where data was collected from different genotyping centers and different studies. These parameters participate as well in this rate of inflation factor. For ROOT dataset, all participants were genotyped in CIDR center.

2.5.3 Performance of adjustment methods in correcting for population stratification under simulated data

In this section, we analyze and compare the obtained results of each tested technique in adjusting for population structure in the simulated data under



(a) All subpopulations were included

(b) Only African American and African Barbadian subpopulations where included

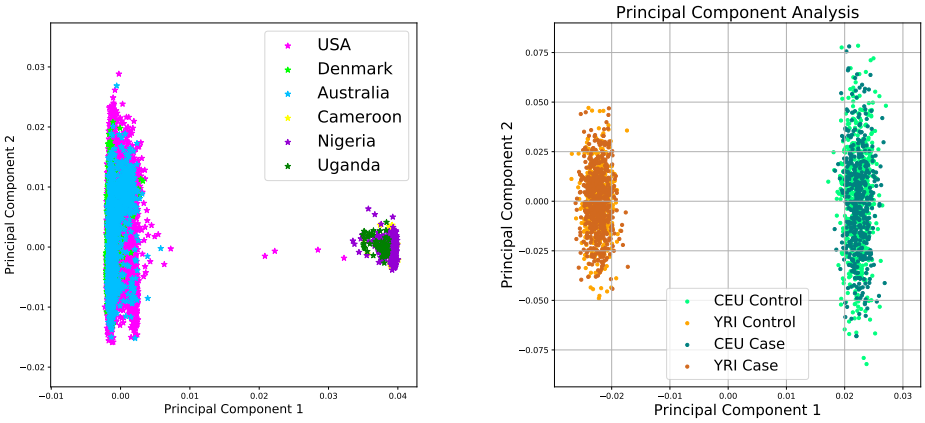
Figure 2.3: PCA for ROOT dataset after performing LD pruning of $r^2 > 0.1$, colors coding corresponds to subpopulations and symbols denotes races presented in Table 2.2. AfAm is African American, AfBB is African Barbadian, AfNG is African from Nigeria

moderate and **strong** population stratification.

First, to compare the efficiency of the methods presented in Section 2.2 at recovering predifined causal SNPs, we rely on the following performance metrics: false positive rate (FPR), false negative rate (FNR), precision, recall and accuracy.

For the **moderate** PS dataset, we observe in Table 2.8 that FastLMM, based on LMM, outperforms the other tested methods. FastLMM retrieves the simulated top causal SNPs with an accuracy, a recall and a precision of 100%, and an FPR and FNR of 0%. It is followed by the PCA-based models (EIGENSTRAT, LogReg1 and LogReg2) that present a fair amount of correction by decreasing remarkably the FPR (of 33.33% to 50%) with good accuracy, precision and recall in identifying predefined causal SNPs compared to the metrics obtained before adjustment. Lastly, GC shows the lowest performance. It is important to mention that some models caused an FNR that was not present before adjustment, specifically GC (25%), EIGENSTRAT and LogReg2 (both 12.5%). This can be explained as an overcorrection issue for some causal SNPs that happen when adjusting for population stratification.

Additionally, we compare in Table 2.6 the simulated odds ratios (OR) of 10 chosen SNPs with the estimated OR obtained by the tested adjustment models. The highlighted cells in green correspond to the lowest percentage of absolute change from true values of OR. The absolute change describes the



(a) PCA for DRIVE dataset, two subpopulations are captured after performing LD pruning of $r^2 > 0.85$: POP1 is composed of USA, Denmark and Australia and POP2 is composed of Cameroon, Nigeria and Uganda

(b) PCA for simulated dataset, two subpopulations are captured (CEU and YRI)

Figure 2.4: PCA plots on DRIVE on the left and simulated data on the right

actual increase or decrease from a true value of OR to a new value of OR. Hence, we confirm again that under **moderate** PS, FastLMM is the best and the less-biased model in estimating the true OR, followed by the PCA-based models mainly EIGENSTRAT and LogReg1 that show also low variation in the estimation of OR. However, GC gives the highest percentages of absolute change from true values of OR (highlighted in red in Table 2.6).

Second, for the **strong** PS dataset, Table 2.9 shows that EIGENSTRAT is the only model that succeed in discarding totally the FPR, but it produces an FNR of 33.33% due to an overcorrection effect. The model presents the best precision of 100%, a high accuracy of 88.88% and a recall of 66.66%. Also, FastLMM presents the best accuracy of 90% compared to other adjustment methods. However, LogReg2 and GC show the lowest performance. This confirms the observation in Table 2.5 showing that LogReg2 fails in adjusting properly for population structure confounder with a high inflation factor of $\lambda = 1.278$. Table 2.7 examines the estimated OR in regards to the true predefined OR. EIGENSTRAT yields the less-biased estimates under **strong** PS simulated data, followed by FastLMM that presents also low percentages of absolute change from true OR. However, we notice that LogReg2 has the highest percentages and fails approximately in estimating the true OR comparing to any other tested technique.

Dataset	Before adjustment	Population stratification adjustment methods				
		GC	EIGEN STRAT	LogReg1	LogReg2	FastLMM
DRIVE	1.153	1.125	1.124	1.126	1.069	1.123
ROOT	1.085	0.998	1.011	1.000	1.046	1.018
Simulated (**moderate** PS)	1.826	1.001	1.000	0.998	1.000	1.000
Simulated (**strong** PS)	4.780	1.000	1.000	0.999	1.278	1.000

Table 2.5: For each dataset, the inflation factor is given after population stratification adjustment obtained by the tested methods

This supports our observations about the low performance of LogReg2 mentioned before in the metrics evaluation.

Method	TP	FP	FN	TN	FPR	FNR	Precision	Recall	Accuracy
No adjustment	4	6	0	0	100%	0%	40%	100%	40%
GC	6	1	2	1	50%	25%	85.71%	75%	70%
EIGEN-STRAT	7	1	1	1	50%	12.5%	87.5%	87.5%	80%
LogReg1	7	1	0	2	33.33%	0%	87.5%	100%	90%
LogReg2	7	1	1	1	50%	12.5%	87.5%	87.5%	80%
FastLMM	7	0	0	3	0%	0%	100%	100%	100%

Table 2.8: Under **moderate** PS simulated data, the following metrics are given: **TR**: true positive, truly selected SNPs; **FP**: false positive, wrongly selected SNPs; **FN**: false negative, wrongly non-selected SNPs; **TN**: true negative, truly non-selected SNPs; **FPR**, false positive rate = $\frac{FP}{FP+TN}$; **FNR**: false negative rate = $\frac{FN}{TP+FN}$; **Precision** = $\frac{TP}{TP+FP}$; **Recall** = $\frac{TP}{TP+FN}$; and **Accuracy**= $\frac{TP+TN}{TP+FP+FN+TN}$

chr	SNP rsID	OR	OR estimated for each method (% Absolute change from true OR)				
			GC	EIGEN STRAT	LogReg1	LogReg2	Fast LMM
12	rs10846175	1.5	1.445(3.6%)	1.511(0.7%)	1.509(0.6%)	1.514(0.9%)	1.510(0.6%)
12	rs7976706	1.5	1.512(0.8%)	1.491(0.6%)	1.504(0.2%)	1.475(1.6%)	1.499(0.06%)
12	rs993123	1.5	1.478(1.4%)	1.488(0.8%)	1.487(0.8%)	1.489(0.7%)	1.486(0.9%)
19	rs344584	1	0.974(2.6%)	0.998(0.2%)	1.002(0.2%)	1.099(9.9%)	0.997(0.3%)
19	rs7257477	1	0.981(1.9%)	0.984(1.6%)	0.984(1.6%)	0.983(1.7%)	1.002(0.2%)
21	rs2833472	2	1.945(2.7%)	1.977(1.1%)	1.969(1.5%)	1.974(1.3%)	1.975(1.25%)
21	rs11910358	2	1.933(3.3%)	2.022(1.1%)	2.001(0.05%)	1.982(0.9%)	2.007(0.3%)
22	rs5996597	2	2.091(4.5%)	2.009(0.4%)	2.010(0.5%)	1.991(0.4%)	1.994(0.3%)
22	rs17004024	2	1.925(3.75%)	2.010(0.5%)	1.975(1.2%)	2.033(1.6%)	2.012(0.6%)
22	rs875643	2	1.977(1.1%)	1.995(0.2%)	2.006(0.3%)	2.018(0.9%)	2.001(0.05%)

Table 2.6: Under **moderate** PS simulated data, the estimated odds ratios (OR) for 10 predefined causal SNPs. Between parenthesis, the percentage of absolute change from true OR. In green, the less-biased method giving the lowest % absolute change from true OR is highlighted for each SNP. In red, the more-biased method giving the highest % absolute change from true OR is highlighted for each SNP.

chr	SNP rsID	OR	OR estimated for each method (% Absolute change from true OR)				
			GC	EIGEN STRAT	LogReg1	LogReg2	Fast LMM
12	rs10846175	1.5	1.447(3.5%)	1.489(0.7%)	1.466(2.2%)	1.436(4.2%)	1.498(0.1%)
12	rs7976706	1.5	1.526(1.7%)	1.497(0.2%)	1.488(0.8%)	1.478(3.3%)	1.506(0.4%)
12	rs993123	1.5	1.477(1.5%)	1.491(0.6%)	1.484(1%)	1.456(2.9%)	1.510 (0.6%)
19	rs344584	1	0.971(2.9%)	0.997(0.3%)	0.989(1.1%)	1.123(12.3%)	1.019(1.9%)
19	rs7257477	1	0.987(1.3%)	1.002(0.1%)	0.978(2.2%)	0.887(11.3%)	1.012(1.2%)
21	rs2833472	2	1.875(6.2%)	1.984(0.8%)	1.971(1.4%)	1.852(7.4%)	2.001(0.05%)
21	rs11910358	2	2.115(5.7%)	1.998(0.1%)	1.932(3.4%)	2.114(5.7%)	1.984(0.8%)
22	rs5996597	2	1.970 (1.5%)	1.987(0.6%)	1.978(1.1%)	1.942(2.9%)	2.010(0.5%)
22	rs17004024	2	1.994(0.3%)	2.015(0.7%)	1.984(0.8%)	1.899(5%)	1.994(0.3%)
22	rs875643	2	1.887(5.6%)	1.965(1.75%)	1.959(2%)	1.944(2.8%)	1.991(1.8%)

Table 2.7: Under **strong** PS simulated data, the estimated odds ratios (OR) for 10 predefined causal SNPs. Between parenthesis, the percentage of absolute change from true OR. In green, the less-biased method giving the lowest % absolute change from true OR is highlighted for each SNP. In red, the more-biased method giving the highest % absolute change from true OR is highlighted for each SNP.

Method	TP	FP	FN	TN	FPR	FNR	Precision	Recall	Accuracy
No adj	3	7	0	0	100%	0%	30%	100%	30%
GC	2	1	1	4	20%	33.33%	66.66%	66.66%	75%
EIGEN-STRAT	2	0	1	6	0%	33.33%	100%	66.66%	88.88%
LogReg1	7	1	0	2	14.28%	33.33%	66.66%	66.66%	80%
LogReg2	2	4	0	4	50%	0%	33%	100%	60%
FastLMM	3	1	0	6	14.28%	0%	75%	100%	90%

Table 2.9: Under **strong** PS simulated data, the following metrics are given: **TR**: true positive, truly selected SNPs; **FP**: false positive, wrongly selected SNPs; **FN**: false negative, wrongly non-selected SNPs; **TN**: true negative, truly non-selected SNPs; **FPR**, false positive rate = $\frac{FP}{FP+TN}$; **FNR**: false negative rate = $\frac{FN}{TP+FN}$; **Precision** = $\frac{TP}{TP+FP}$; **Recall** = $\frac{TP}{TP+FN}$; and **Accuracy**= $\frac{TP+TN}{TP+FP+FN+TN}$

2.5.4 Population stratification adjustment in real data

In this section, we examine the obtained Q-Q plots for all tested correction methods in real data and their corresponding inflation factors. However, we can not perform a metric evaluation as done for simulated data because we do not have a ground-truth or prior-knowledge about the true associations.

For DRIVE dataset, Table 2.5 shows that LogReg2 outperforms the other tested models. It produces the lowest inflation factor by reducing it from $\lambda = 1.153$ (before adjustment) to $\lambda = 1.069$. The other methods give very similar values after adjustment of $\lambda \simeq 1.12$. This can be visualized also in Figure 2.5 that presents the different Q-Q plots drawn before and after population stratification adjustments.

For ROOT dataset, all models help to reduce the inflation factor to be very close to 1, this interpretation is confirmed in the Q-Q plots shown in Figure 2.6.

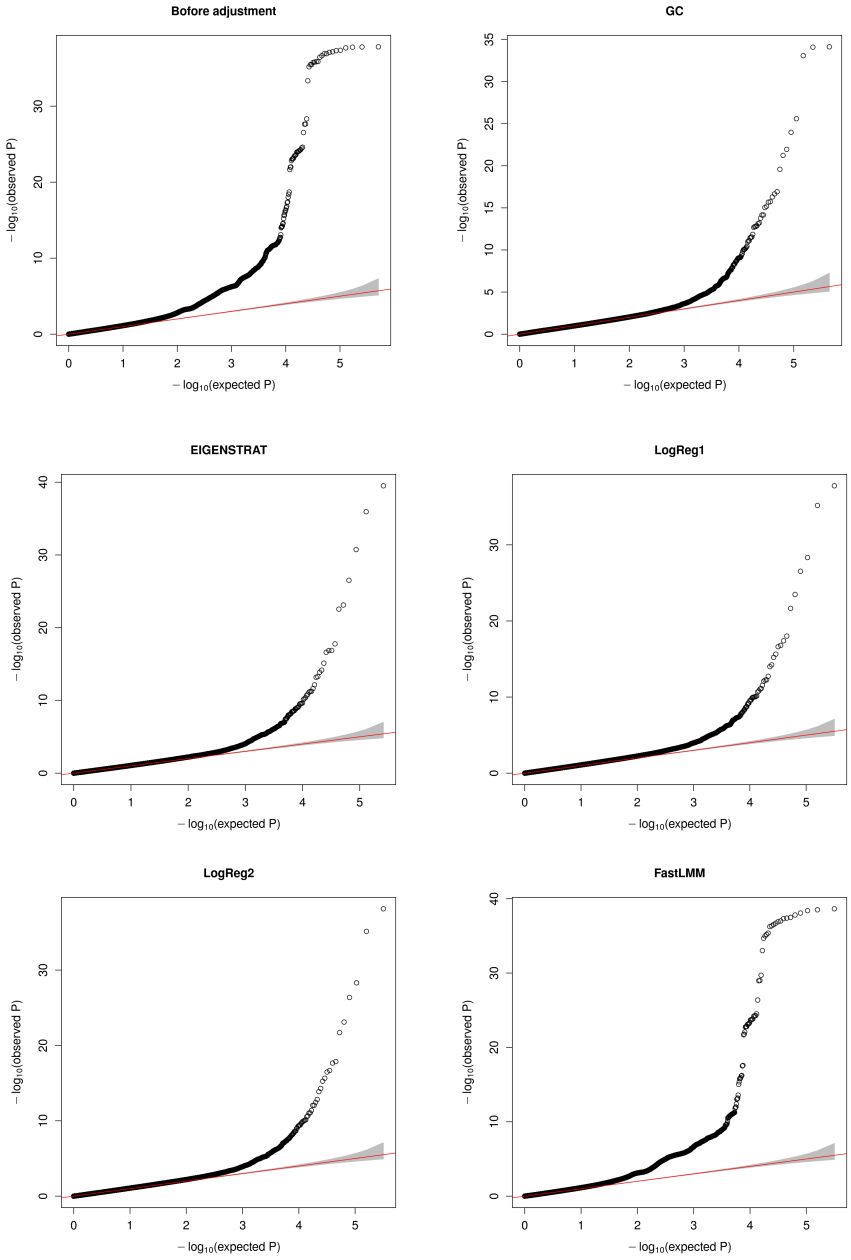


Figure 2.5: Q-Q plots obtained for DRIVE dataset before adjustment and after adjustment using the following methods: Genomic control(GC), EIGENSTRAT, Logistic Regression with top PCs as covariates (LogReg1), Logistic Regression for phenotype adjustment (LogReg2) and FastLMM.

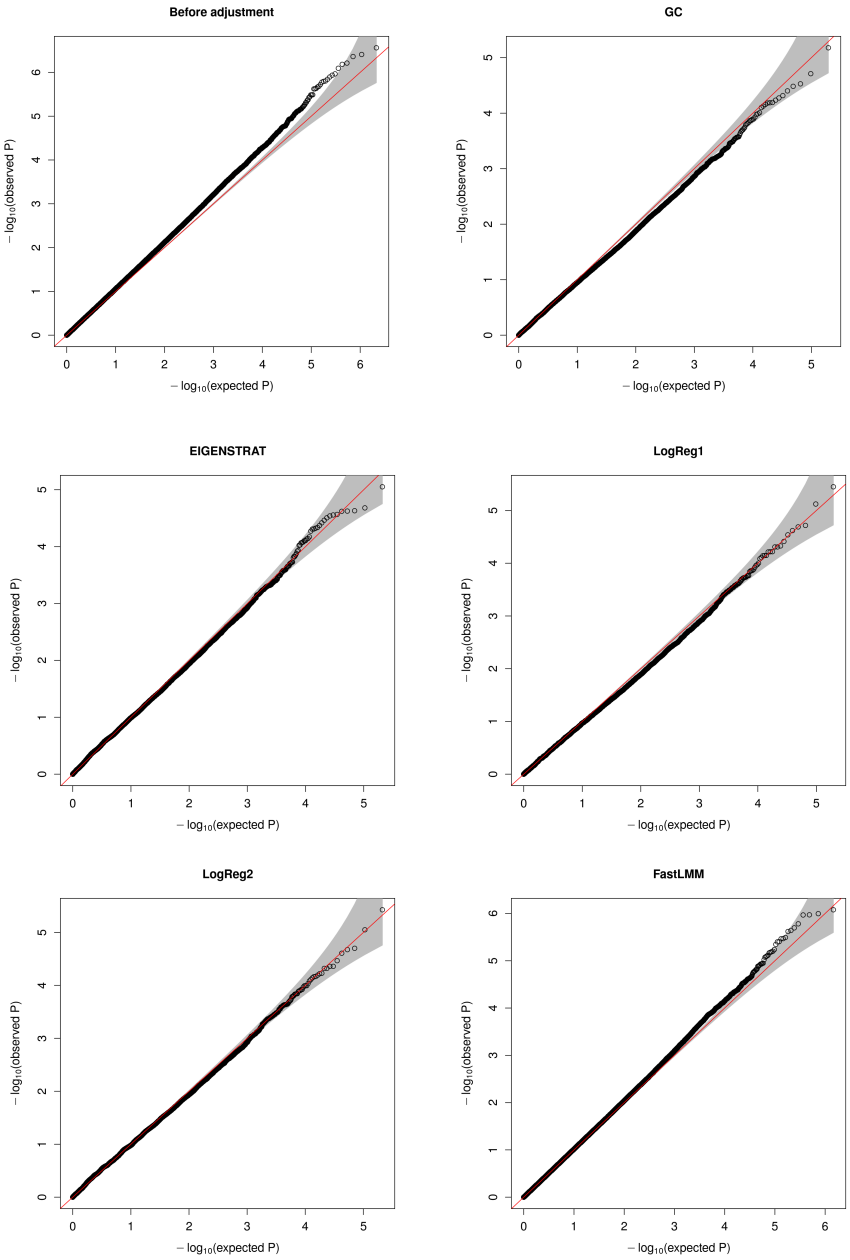


Figure 2.6: Q-Q plots obtained for ROOT dataset before adjustment and after adjustment for population stratification using the following methods: Genomic control(GC), EIGENSTRAT, Logistic Regression with top PCs as covariates (LogReg1), Logistic Regression for phenotype adjustment (LogReg2) and FastLMM.

2.6 Discussion and conclusion

In this chapter, we have presented a review of the main adjustment methods for population stratification confounder in case-control phenotype on GWAS data. We have applied particularly the following techniques: the genomic control (GC), three PCA-based models (EIGENSTRAT, LogReg1 and LogReg2) and FastLMM based on LMM. We have used different simulated data scenarios (**moderate** and **strong** population stratification) and two cancer real datasets (DRIVE and ROOT) to evaluate the tested methods. We have found that adjustment results differ from data to another, they depend also on the severity of population stratification. Globally, FastLMM has been the best method for population structure adjustment especially in both simulated data scenarios. However, it did not offer a good enough correction of the inflation factor λ in DRIVE dataset. PCA-based models, mainly EIGENSTRAT and LogReg1, offer also good adjustment performance. However, while examining results of LogReg2, our conclusions have been divided. On the one hand, the method has given the best correction on DRIVE dataset. On the other hand, in the **strong** PS simulated data, the performance of LogReg2 in retrieving the predefined causal SNPs and their corresponding odds ratios has been low compared to the other tested techniques.

We have observed also an overcorrection effect of some causal SNPs that occurs particularly after using some adjustments models. This has been detected by the presence of a considerable FNR after adjustment for population stratification that was not existent before correction.

Another critical point of these methods is that they make the application of feature selection models based on regularization terms more complex. For example, using mixed models increases the complexity of the problem for the next steps of this study. Regularization-based feature selection procedures become more complex to set up using mixed models. Also, EIGENSTRAT and LogReg1 models that consist in adding the top PCs as covariates in the model will not facilitate the feature selection task. For instance, if the feature selection model does not select all the considered PCs, then it will not consider the population stratification presence. Indeed, LogReg2 is one possible alternative, as it adjusts for the phenotype previously. The new phenotype is quantitative and represents no more a case-control analyze, it remains then possible to apply feature selection models with quadratic loss function. However, the observed reduced performance of LogReg2 in identifying causal SNPs in some data cases limits the confidence in its discoveries.

In addition, many studies prove that different genetic populations do not share always the same markers associated with disease or tumor growth [Medina-Gomez *et al.*(2015), Wu *et al.*(2013), Fu *et al.*(2011)]. Unfortunately, the presented adjustment methods perform a uniform correction and do not consider the possibility of the presence of populations-specific causal markers.

We aim in this thesis to address the population stratification problem in a more efficient way by providing a powerful framework that allows the application of machine learning models and the identification of populations-specific markers associated with a disease. This helps to avoid the discussed issues raised in traditional adjustments approaches. This study has allowed us to continue this work by proposing two novel methods for addressing population stratification issue efficiently. The developed models address also other GWAS problems such as the curse of dimensionality, the high computational complexity and the lack of stability. These methods are presented in Chapter 4 and Chapter 5.

Chapter 3

Multiscale genomic evaluation of the stability of the selection for Genome-Wide Association Studies

Abstract: *The stability of the feature selection refers to the robustness of the selected variables, it represents an important criterion in Genome-Wide Association Studies to trust the precision of the discovered features considered as associated with the phenotype. Thus, quantifying the stability of the selection is possible relying on a set of desirable properties. The state-of-the-art methods focus on measuring the stability at the selection level. We propose to study the stability at different genomic levels (Single Nucleotide Polymorphisms (SNPs) level, Linkage Disequilibrium (LD) blocks level and gene level) of several feature selection methods (single-marker analyses, Lasso, Elastic Net and stability selection models). We demonstrate that the stability of both feature selection models (i.e., Lasso and Elastic Net) increases remarkably at the LD-blocks and the gene level compared to the SNP level. Although we have found that single-marker analyses is the most stable technique, this method suffer from low statistical power in retrieving causal SNPs and, thus, misses many meaningful associations. As well, we show that stability selection remarkably improves the robustness of the tested methods (Lasso, Elastic Net).*

Résumé: *La stabilité de la sélection des variables fait référence à la robustesse des variables sélectionnées, elle représente un critère important dans Les études d'association pangénomiques pour faire confiance à la précision des variables découvertes considérées comme associées au phénotype. Ainsi, quantifier la stabilité de la sélection est possible en s'appuyant sur un ensemble de propriétés souhaitables. Les méthodes existantes se concentrent sur la mesure de la stabilité au niveau de la sélection. Nous proposons d'étudier la stabilité à différents niveaux génomiques (au niveau des polymorphismes nucléotidiques (SNPs), au niveau des blocs de déséquilibre de liaison (LD) et au niveau des gènes) de plusieurs méthodes de sélection de variables (analyses à marqueur unique, Lasso, Elastic Net et modèles de sélection de stabilité). Nous démontrons que la stabilité des deux modèles de sélection des variables (c'est-à-dire*

Lasso et Elastic Net) augmente remarquablement au niveau des blocs LD et au niveau de gène par rapport au niveau de SNP. Bien que nous ayons constaté que les analyses à marqueur unique sont la technique la plus stable, cette méthode souffre d'une faible puissance statistique dans la découverte des SNP causaux, et rate donc de nombreuses associations significatives. Aussi, nous montrons que la sélection de stabilité améliore remarquablement la robustesse des méthodes testées (Lasso, Elastic Net).

Contents

3.1 Introduction 50
3.2 Methods 51
3.2.1 Association analysis and feature selection models . . . 51
3.2.2 Measuring the stability at different genomic scales . . 52
3.2.3 Linkage Disequilibrium blocks clustering 54
3.2.4 FUMA for functional mapping and annotation of genes 55
3.2.5 Stability selection 55
3.2.6 Related work 58
3.3 Experiments 59
3.3.1 Data 59
3.3.2 Preprocessing 59
3.3.3 Implementation details 59
3.4 Results 60
3.4.1 Clustering the SNPs to LD-blocks and mapping the SNPs to genes 60
3.4.2 The stability of the selection in classical GWAS 60
3.4.3 Lasso and Elastic Net lead to better biological interpretation for biomarker discovery 61
3.4.4 Stability selection methods increase the stability index of Lasso 63
3.5 Discussion and conclusion 64

3.1 Introduction

Genome-Wide Association Studies (GWAS) represent a powerful approach in the identification of associated causal Single Nucleotide Polymorphisms (SNPs) with a phenotype of interest and genes involved in human diseases. In high-dimensional data, this task remains challenging due to the small number of participants compared to the number of SNPs. Numerous methods have been proposed to reduce the number of variables by keeping only a meaningful set of variants that explain the studied phenotype [Tibshirani(1996), Zou and Hastie(2005), Obozinski *et al.*(2006), Jacob *et al.*(2009)]. This helps to improve the prediction power and reducing possible over-fitting risk.

Besides, the selection of relevant regions in the genome leads to discovering candidate genes in relationship with some diseases or the growth of some tumors. The classical approach consists of using single-marker analyses. From a machine learning point of view, further methods using regularization terms have been developed. These approaches are based on a multivariate approach where the effect of variants is treated jointly. Some studies have shown that a major concern of feature selection is the lack of stability, defined as the variability of the selection for small perturbation in the input set [Haury *et al.*(2011), Nogueira and Brown(2016), Nogueira *et al.*(2018)]. Many ways of scores for measuring the stability of the selection have been proposed with different motivations relying on a set of desirable properties that a stability index must fullfill [Kalousis *et al.*(2007), Kuncheva(2008), Lustgarten *et al.*(2009), Somol and Novovicova(2010), Wald *et al.*(2013), Nogueira and Brown(2015), Nogueira and Brown(2016), Nogueira *et al.*(2018)] (See Section 1.4).

An additional concern in GWAS is Linkage Disequilibrium (LD) presented in Section 2.3, which corresponds to high correlation between SNPs and leads to strong statistical dependence between the predictors and a considerable loss of statistical power. One proposed solution in the literature is to merge strongly correlated SNPs into blocks and perform single-marker analyses or feature selection at the LD-block level to alleviate the effect of LD [Dehman *et al.*(2015), Ambroise *et al.*(2019)]. Indeed, few studies have proven that performing the selection at the LD-block level addresses the curse of dimensionality and improves remarkably the prediction power [Liu *et al.*(2012), Dehman *et al.*(2015)]. However, [Haury *et al.*(2011)] have shown that classical methods based on single-marker analyses remain more stable compared to multivariate feature selection methods. Thus, few articles [Meinshausen and Bühlmann(2009)] and [Shah and Samworth(2013)] have presented a way to improve the stability and to decrease type I error by introducing subsampling technique in two different ways. In this chapter, we present an empirical evaluation to compare several methods based on single-marker analyses and machine learning at different genomic scales: the SNP level, the LD-block

level and the gene level. We have found that the stability of the selection of Lasso and Elastic Net remarkably increased at the LD-block and gene levels compared to the SNP level. Hence, to improve the stability of the selection, we have applied two different stability selection models based on subsampling procedures. We have used Wellcome Trust Case Control Consortium 1 data (WTCCC1) [Consortium(2007)] to conduct our analysis on three different diseases: Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D).

3.2 Methods

3.2.1 Association analysis and feature selection models

A classic GWAS approach based on single-marker analyses is carried out by performing a simple statistical test for each SNP individually. This technique was detailed in Section 1.2.1.

From a Machine Learning perspective, the choice of an appropriate set of features is an important step in which the dimensionality of the space is reduced, in order to train a minimal subset of features that are relevant for building the predictive model. In a GWAS study, the aim is to select features which are fully associated with a given phenotype of interest. Several feature selection algorithms exist within the Machine Learning framework. We conduct our study using two regularization based models: Lasso (described in Section 1.3.2) and the Elastic Net (explained in Section 1.3.6), both of them ensure sparsity and association between the genotype presented by X and the phenotype of interest y, which is qualitative (binary) in our case. One can claims that basic methods based on linear models can be considered as elementary techniques, that do not ensure a good representation of GWAS problem. However, modeling nonlinear effects in GWAS is not straightforward. It is necessary to add biological interactions to model efficiently the nonlinearities. In fact, these interactions increase dramatically the complexity of the problem and the curse of dimensionality. Consequently, linear methods remain an interesting approach to model a GWAS problem efficiently, thanks to their satisfying interpretability.

We have already established in Section 1.3 the objective functions of these feature selection models using the generic loss. We rewrite them using the logistic loss where y denotes a case-control phenotype:

(1) The loss function of Lasso is given by:

$$\min_{\beta \in \mathbb{R}^p} -yX^\top \beta + \log\left(1 + \exp\left(X^\top \beta\right)\right) + \lambda \sum_{j=0}^{p} \mid \beta_j \mid$$

where the penalization parameter λ controls the strength of the penalty.

(2) The loss function of Elastic Net is defined as follows:

$$\min_{\beta \in \mathbb{R}^p} -yX^\top \beta + \log\left(1 + \exp\left(X^\top \beta\right)\right) + \lambda_1 \sum_{j=0}^{p} \beta_j^2 + \lambda_2 \sum_{j=0}^{p} \mid \beta_j \mid,$$

where λ_1 and λ_2 are the penalization terms that control the strength of both regularization terms.

3.2.2 Measuring the stability at different genomic scales

The stability of the selection index

We have presented in Section 1.4.2 the measurement of the stability of the selection. The choice of stability index is made following a set of desirable properties detailed in Section 1.4.3. The general steps to quantify the stability of the selection are listed below:

- Generate M randomized samples of the dataset by using resampling or bootstrapping.
- Perform a chosen feature selection model on these M generated samples to obtain the selected feature subset $\mathcal{Z} = \{s_1, \ldots, s_M\}$.
- Pick up a stability index $\hat{\Phi}(\mathcal{Z}) : \{0,1\}^{M \times p} \longrightarrow \mathbb{R}$ to compute the stability of the applied feature selection procedure.

[Nogueira and Brown(2015)] compared the state-of-the-art methods of stability measurement. Among these methods, two main indexes fulfill all the properties that a stability measure must respect. The first method is carried out with the Pearson similarity index and represents an extension of [Kuncheva(2008)]. Assume $\mathcal{Z} = \{s_1, \ldots, s_M\}$ is the set of M selected features where each s_u is a subset of the features. The total number of features is denoted by p and the number of features selected on the u^{th} feature set is given by k_u. A set of selected features s_u can be represented by an indicator vector $z_{u,.} \in \{0,1\}^p$, where $z_{u,j} = 1$ if feature j is selected and 0 otherwise. The Pearson correlation between two feature sets s_u and s_v is presented by the following equation:

$$\phi_{\text{Pearson}}\left(s_u, s_v\right) = \frac{\frac{1}{p}\sum_{j=1}^{p}\left(z_{u,j} - \bar{z}_{u,.}\right)\left(z_{v,j} - \bar{z}_{v,.}\right)}{\sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(z_{u,j} - \bar{z}_{u,.}\right)^2}\sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(z_{v,j} - \bar{z}_{v,.}\right)^2}},$$

where $\forall u \in \{1, \ldots, M\}, \bar{z}_{u,.} = \frac{1}{p}\sum_{j=1}^{p} z_{u,j} = \frac{k_u}{p}$

The stability of the selection based on Pearson correlation can be rewritten as follows:

$$\phi_{\text{Pearson}}\ (\mathbf{s}_u, \mathbf{s}_v) = \frac{r_{u,v} - \mathbb{E}_\nabla\left[r_{u,v}\right]}{p h_u h_v} = \frac{r_{u,v} - \frac{k_u k_v}{p}}{p h_u h_v},$$

where $\forall u \in \{1, \ldots, M\}, h_u = \sqrt{\frac{k_u}{p}\left(1 - \frac{k_u}{p}\right)}$ and $r_{u,v}$ denotes the number of features in common between the feature sets s_u and s_v. $\mathbb{E}_\nabla$ is an adjustment term equal to the expected value of $r_{u,v}$ when the feature selection model selects randomly k_u and k_j features from all features p.

When the number of selected features k is the same for all feature sets, assuming S an index of the variability in the choice of features, the Pearson correlation is established as follows:

$$\hat{\Phi}_{\text{Pearson}}(\mathcal{Z}) = 1 - \frac{S}{S_{\max}} = \frac{\frac{1}{p}\sum_{j=1}^{p} s_j^2}{\frac{k}{p}\left(1 - \frac{k}{p}\right)},$$

where $S_{\max}$ is the maximal value of S when the feature selection model selects k features per feature set. $s_j^2 = \frac{M}{M-1}\hat{q}_j\left(1 - \hat{q}_j\right)$ corresponds to the sample variance of selection of the j^{th} feature. $\hat{q}_j$ corresponds to the observed frequency of the selection of a feature j, also the sample mean of the variable $\mathcal{Z}_j$.

The second stability index was proposed by [Nogueira *et al.*(2018)], it is similar to the first method. When the number of selected features is the same for selected subsets, the only difference is that the selected number of features k_i is the same for all subsets, denoted by k, and computed as the average number of selected features across all selected sets.
Under the null model of feature selection procedure H_0 and for all features j, the expected value of the sample variance of Z_j is given by $\mathbb{E}\left[s_j^2 \mid H_0\right] = \frac{\bar{k}}{p}\left(1 - \frac{\bar{k}}{p}\right)$.

Then, the stability measurement was defined by [Nogueira *et al.*(2018)] as follows:

$$\hat{\Phi}_{\text{Nogueira}}(\mathcal{Z}) = 1 - \frac{\frac{1}{p}\sum_{j=1}^{p} s_j^2}{\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} s_j^2 \mid H_0\right]} = 1 - \frac{\frac{1}{p}\sum_{j=1}^{p} s_j^2}{\frac{\bar{k}}{p}\left(1 - \frac{\bar{k}}{p}\right)}$$

Empirically, both of the presented stability measurements give similar results, as they fulfill all desirable properties that an index must have (see Table C.1). We have decided to work with the Pearson correlation alternative, as it is preferable to consider the exact number k_i of selected features for each selected set, rather than the average number k of selected features across all selected sets.

The stability evaluation at different levels

The typical methodology of the stability estimation is performed in general at the same level as the feature selection procedure. In other words, we tend to compute the stability index at the level of SNPs if the feature selection algorithm select SNPs, and at the level of genes if it selects genes, and so on.

Thus, one interesting direction is to evaluate the stability of the selection of widely-used methods for GWAS at different scales, more precisely the SNP level, the LD-block level and the gene level. In this study, we conduct the four following tasks:

- Implementation of four different approaches: single-marker analyses, Lasso, Elastic Net, bootstrap-stabilized Lasso using two methods [Meinshausen and Bühlmann(2009), Shah and Samworth(2013)].
- Measurement of the stability of the selection at the SNP level for each applied model.
- Definition of LD-blocks of strongly correlated SNPs, to compute the stability at the LD-block level. We consider an LD-block to be selected if one SNP within it was selected by the feature selection model.
- Mapping of SNPs to genes to compute the stability at the gene level. A gene is considered as selected if the SNPs associated to that gene was selected.

3.2.3 Linkage Disequilibrium blocks clustering

As explained in Section 1.2.5, LD induces a strong correlation between nearby SNPs in the same chromosome. From a Machine Learning point of view, including correlated features in a model is similar to use redundant information. Thus, such features representation can be seen as inconsistent and reduces the performance of the model. One solution to resolve this issue and improve GWAS data representation is to group highly correlated SNPs in the same chromosome to biologically relevant blocks, named LD-blocks. [Ambroise *et al.*(2019)] have proposed a clustering approach to obtain these LD-blocks. The main idea of this algorithm is based on incorporating a constraint in the classical hierarchical agglomerative clustering where each SNP belongs to its own cluster and iteratively merges the two most similar clusters according to a distance function called a linkage criterion, this constraint relies on Ward's linkage given by d_{wl} as follows:

$$d_{\text{wl}}(A, B) = \frac{p_A \times p_B}{p_A + p_B} \delta\left(\mathbf{g}_A, \mathbf{g}_B\right)^2,$$

where p_A and p_B are the cardinals of the clusters A and B respectively, and $\mathbf{g}_A$ and $\mathbf{g}_B$ are the centers of the clusters, respectively. δ denotes the dissimilarity between $\mathbf{g}_A$ and $\mathbf{g}_B$.

It is possible therefore to perform feature selection at the LD-block level instead of the single SNP level. This can be conducted using the group Lasso for example that handles groups set from prior knowledge. We will handle this procedure to develop novel methods presented in Chapter 4 and Chapter 5.

3.2.4 FUMA for functional mapping and annotation of genes

After the selection of a subset of causal SNPs associated with the disease, one major question is to interpret the relevance of these results biologically. A possible solution is to perform functional mapping of these SNPs to genes. FUMA [Watanabe *et al.*(2017)] is a useful tool in order to map functionally annotated SNPs to genes according to the physical position in the genome, eQTL mapping and 3D chromatin interaction. FUMA uses information from multiple biological data to perform these mapping analysis, several controller parameters are to be set such as the physical window to map SNPs to genes (the default parameter is 10 kb). In this study, we perform functional mapping for all SNPs included in the datasets using FUMA. Hence, we compute the stability of the selection at the gene level. Note that one SNP could be mapped to several genes. But, the mapping to genes is sometimes not possible for some SNPs.

3.2.5 Stability selection

We present first the stability selection procedure developed by [Meinshausen and Bühlmann(2009)] where they propose improving the stability using a subsampling method. In this formulation, variable selection is performed repeatedly on subsamples. They demonstrate that the subsampling approach can be used to determine the amount of regularization needed to control the family-wise error type I rate. Stability selection is a feature selection based method, it can be combined with several existing methods and aims to improve their performance. The procedure relies on computing the stability path, which represents the probability of a feature to be selected across random subsamples, as a function of the regularization parameter.

We denote by I a random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, we call $\widehat{S}^{\lambda}(I)$ the set of features selected by the selection procedure of interest (for example, Lasso), with a hyperparameter λ, on this subsample of the data. For any feature $j \in \{1, \ldots, p\}$, we call $\widehat{\Pi}_j^{\lambda}$ the probability that feature j is selected on a random subsample of size $\lfloor n/2 \rfloor$ of the data. This probability is determined, given m such random subsamples $I_1, I_2, \ldots, I_m$, as the proportion

of those subsamples for which the feature selection procedure selects a feature j:

$$\widehat{\Pi}_j^{\lambda} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{j \in S^{\lambda}(I_j)}.$$

Finally, given a threshold $\frac{1}{2} < \pi_{\text{cutoff}} \leq 1$ (in this work, we used $\pi_{\text{cutoff}} = 0.75$), the stable set of selected features is determined as:

$$\widehat{S}^{\text{stable}} = \{j : \max_{\lambda \in \Lambda} \widehat{\Pi}_j^{\lambda} \geq \pi_{\text{cutoff}}\}.$$

To bound the number of expected false selected features, [Meinshausen and Bühlmann(2009)] present the following theorem (Theorem 1 in their paper):

Assuming the set of feature with non-zero coefficients by $S = \{j : \beta_j \neq 0\}$, and the set of features with zero coefficients by $N = \{j : \beta_j = 0\}$. Assuming that the distribution of $\left\{1_{\{j \in \hat{S}^{\lambda}\}}, j \in N\right\}$ is exchangeable for all $\lambda \in \mathbb{R}^+$. Also, assuming that the original procedure is not worse than random setting, i.e. for any $\lambda \in \mathbb{R}^+$:

$$\frac{\mathbb{E}\left(\left|S \cap \hat{S}^{\lambda}\right|\right)}{\mathbb{E}\left(\left|N \cap \hat{S}^{\lambda}\right|\right)} \geq \frac{|S|}{|N|}.$$

The number of falsely selected variables $V = \left|N \cap \hat{S}^{\text{stable}}\right|$ is then bounded by:

$$\mathbb{E}(V) \leq \frac{1}{2\pi_{\text{cutoff}} - 1} \frac{q_{\lambda}^2}{p},$$

where $q_{\lambda} = \mathbb{E}\left(|S_{\lambda}(I)|\right)$ denotes the expected number of selected variables The desired calibration is to obtain $\mathbb{E}(V) \leq \alpha$ with α small.

As an example, for a cutoff $\pi_{\text{cutoff}} = 0.75$, and $\alpha = 0.05$, λ is then chosen such as $q_{\lambda} < \sqrt{0.025p}$ in order to obtain a FWER $< \alpha$. Therefore, $\mathbb{E}(V)$ is controlled at the desired level following Theorem 1 if we select 93 features from $p = 350\,000$ features.

Another variant of stability selection, named complementary pairs stability selection (CPSS), was proposed by [Shah and Samworth(2013)]. It improves the applicability of the previous stability selection method. The subsamples are introduced in B complementary pairs $\{(I_{2m-1}, I_{2m}) : m = 1, \ldots, B\}$ where each I_m is a subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$.

Assuming a feature selection model $\hat{S}_n := \hat{S}_n(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are vector-valued data, the authors define the probability of the selection of a feature of index $j \in \{1, ..., p\}$ as follows:

$$p_{j,n} = \mathbb{P}\left(j \in \hat{S}_n\right) = \mathbb{E}\left(\mathbf{1}_{\left\{j \in \hat{S}_n\right\}}\right).$$

Here, $\mathbf{1}_{\left\{j \in \hat{S}_n\right\}}$ has a Bernoulli distribution with parameter $p_{j,n}$ and can be seen as an unbiased estimator of $p_{j,n}$. As mentioned before, the main goal of this method is to improve $p_{j,n}$ estimation by applying the subsampling procedure.

The authors introduce the CPSS version of the feature selection model $\hat{S}_n$ by $\hat{S}^{\text{CPSS}}_{n,\pi_{\text{cutoff}}} = \left\{j : \hat{\Pi}_B(j) \geqslant \pi_{\text{cutoff}}\right\}$. The function $\hat{\Pi}_B : \{1, \ldots, p\} \to \{0, 1/(2B), 1/B, \ldots, 1\}$ is defined by the following equation:

$$\hat{\Pi}_B(j) := \frac{1}{2B} \sum_{m=1}^{2B} \mathbf{1}_{\left\{j \in \hat{S}(I_j)\right\}}.$$

In addition, they define the simultaneous selection of $\hat{S}_n$ for both complementary pairs (I_{2m-1}, I_{2m}) with $\tilde{\Pi}_B$ as follows:

$$\tilde{\Pi}_B(j) := \frac{1}{B} \sum_{m=1}^{B} \mathbf{1}_{\left\{j \in \hat{S}(I_{2m-1})\right\}} \mathbf{1}_{\left\{j \in \hat{S}(I_{2m})\right\}}.$$

Finally, this variant of stability selection methods provides bounds for both:

1. The expected number of features integrated by this model (CPSS) with low selection probability features that are excluded.

2. The expected number of features integrated by this model with high selection probability that are excluded.

These bounds result in higher precision in the feature selection procedure by improving the error control. We present below Theorem 1 given by [Shah and Samworth(2013)] that allows to control these bounds:

For $\alpha \in [0, 1]$, we denote by $L_\alpha = \left\{j : p_{j,\lfloor n/2 \rfloor} \leqslant \alpha\right\}$ the set of features indexes with low selection probability obtained using the feature selection procedure $\hat{S}_{\lfloor n/2 \rfloor}$. We denote also by $H_\alpha = \left\{j : p_{j,\lfloor n/2 \rfloor} > \alpha\right\}$ the set of features indexes with high selection probability.

1. If $\pi_{\text{cutoff}} \in \left(\frac{1}{2}, 1\right]$, then:

$$\mathbb{E}\left|\hat{S}^{\text{CPSS}}_{n,\pi_{\text{cutoff}}} \cap L_\alpha\right| \leqslant \frac{\alpha}{2\pi_{\text{cutoff}} - 1} \mathbb{E}\left|\hat{S}_{\lfloor n/2 \rfloor} \cap L_\alpha\right|.$$

2. Assuming $\hat{N}^{\text{CPSS}}_{n,\pi_{\text{cutoff}}} = \{1, \ldots, p\} \backslash \hat{S}^{\text{CPSS}}_{n,\pi_{\text{cutoff}}}$ and $\hat{N}_n = \{1, \ldots, p\} \backslash \hat{S}_n$. If $\pi_{\text{cutoff}} \in \left[0, \frac{1}{2}\right)$, then:

$$\mathbb{E}\left|\hat{N}^{\text{CPSS}}_{n,\pi_{\text{cutoff}}} \cap H_\alpha\right| \leqslant \frac{1 - \alpha}{1 - 2\pi_{\text{cutoff}}} \mathbb{E}\left|\hat{N}_{\lfloor n/2 \rfloor} \cap H_\alpha\right|.$$

3.2.6 Related work

Stability selection was first presented by [Bach(2008)] where he has introduced Bolasso (Bootsrapped Lasso). The author has proved that using a small number of subsamples for Lasso and fixing the probability threshold $\pi_{\text{cutoff}} = 1$ provide a robust feature selection procedure. Thus, only consistentely selected features are kept by the model. However, [Meinshausen and Bühlmann(2009)] have claimed that Bolasso relies on choosing the regularization term λ across subsamples. If the value of λ is too large on more than 10% of all subsamples, the model discards relevant features.

Following the success of stability selection methods, further studies have been conducted to propose other alternatives. For instance, [Alexander and Lange(2011)] have proposed to apply stability selection procedure to GWAS applications. Their approach provides feature selection on groups of SNPs contained in gene regions instead of raw SNPs selection, in order to decrease the predictor correlation among markers, and increase the biological interpretability.

In addition, [Haury *et al.*(2012)] have proposed another stability selection method called TIGRESS. In this work, a combination of Least-angle regression (LARS) and stability selection was developed. Note that LARS is a feature selection model, where the coefficients are first initialized by zeros. The algorithm tends to find the features that are associated with the phenotype (the disease) by increasing the relevant coefficients. [Haury *et al.*(2012)] approach integrates a new scoring method for stability selection. This score is defined by the area under the stability curve. The advantage of the new measure compared to [Meinshausen and Bühlmann(2009)] is that it takes into consideration the full distribution of ranks of a variable in the feature selection procedure. They have proved experimentally that this method provides better performance in terms of stability. However, TIGRESS was tested in gene-expression data that has much fewer features (around 20 000), as compared to GWAS datasets (containing hundreds of thousands up to millions of features). Hence, the method does not fit computationally the GWAS scale.

Recently, [Sabourin *et al.*(2019)] have proposed ComPaSS-GWAS, their approach is based on complementary pairs stability selection of [Shah and Samworth(2013)]. It was applied to GWAS data for a quantitative phenotype. ComPaSS-GWAS splits randomly the samples in half repeatedly and compare the results of the selection between both splits using a traditional GWAS test. The significant SNPs that were selected across each random split are finally returned with a score of corroboration between 0 and 1. Nevertheless, the method was only conducted on single-marker analyses that do not suffer from the lack of stability.

3.3 Experiments

3.3.1 Data

Wellcome Trust Case Control Consortium 1 (WTCCC1)

We study three datasets from the WTCCC1 data from different diseases (Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D) that we have introduced in Section 1.5.2.

3.3.2 Preprocessing

We exclude poorly performing SNPs that do not pass the following quality control filters recommended in [Consortium(2007)]. We remove SNPs with a minor allele frequency lower than 5%, a p-value for Hardy-Weinberg Equilibrium in controls lower than 0.001 and a missing genotyping rate larger than 10%. In addition, samples that have an overall genotyping missing rate larger than 10% are also excluded. The SNPs of sex chromosomes are removed because they were not genotyped for all participants. The WTCCC1 data was already imputed using CHIAMO, but the remaining missing values are replaced with the major allele denoted by 0.

3.3.3 Implementation details

In this section, we detail the packages and tools that we use to implement the developed methods.

We start by generating bootstrapped samples from the data. Then, we perform classic GWAS based on single-marker analyses using PLINK [Purcell *et al.*(2007)] that runs a $1df\,\chi^2$ allelic test between each SNP individually and the phenotype on each bootstrapped sample.

We use `scikit-learn` package to implement Lasso and Elastic Net models. The Lasso model is evaluated for different values of the penalization parameter λ. In Elastic Net, we fix $\lambda_1 = 0.05$ and the model is evaluated for different values of λ_2 similarly to what is done for Lasso.

In order to improve the stability of both feature selection algorithms, we use the `stability-selection` package[1] that handles scikit-learn feature selection estimators and both stability selection methods presented in Section 3.2.5.

As mentioned before, the stability is evaluated at different genomic scales (SNP, LD-block and gene). We obtain the LD-blocks using the `R` package `adjclust` [Ambroise *et al.*(2019)]. Finally, the SNPs to genes mapping is determined with the web-based platform of FUMA [Watanabe *et al.*(2017)].

[1] https://github.com/scikit-learn-contrib/stability-selection

Dataset	Number of SNPs	Number of LD-blocks	Number of SNPs mapped to genes
RA	354 678	1 580	93 788
T1D	354 523	2 035	93 741
T2D	354 615	1 887	93 870

Table 3.1: For each dataset, the number of SNPs and their corresponding LD-blocks and genes obtained after the clustering and the positional mapping respectively

The stability index at each genomic scale (SNP, LD-block and gene) is computed using the Pearson Correlation Coefficient method described in Section 3.2.2.

The implemented codes are available online in the following github repository: https://github.com/asmanouira/multiscale-stability

3.4 Results

3.4.1 Clustering the SNPs to LD-blocks and mapping the SNPs to genes

In order to compute the stability of the selection at the different genomic scales (SNP, LD-block and gene levels), we determine first the LD-blocks and the genes corresponding to the SNPs from each dataset. Thus, we identify the LD-blocks and the genes that were selected across all bootstrapped samples. We present in Table 3.1 for each dataset, the number of obtained LD-blocks after performing the hierarchical agglomerative clustering of the SNPs, as well as the number of mapped SNPs to genes using FUMA. We observe particularly that the mapping of genes was not possible for all SNPs. Thus, the interpretation of the stability index at the gene level is different to the SNP level and the LD-block level. Indeed, many SNPs selected consistently across the samples bootstraps were not mapped to genes. In this case, it is then normal to have higher stability index at the LD groups level or/and the SNP level compared to the gene level.

3.4.2 The stability of the selection in classical GWAS

In this work, we choose $M = 10$ the number of bootstrapped samples generated randomly to compute the stability of the selection index. We then run PLINK repeatedly 10-times for the 10 bootstrapped samples. For single-marker analyses, the obtained p-values are an indicator of the statistical significance and a strong association with the disease. Thus, we consider that the

Dataset	Average number of selected SNPs	Average number of selected LD-blocks	Average number of mapped SNPs to genes
RA	103	19	2
T1D	227	13	3
T2D	10	10	2

Table 3.2: For single-marker analyses, the average number of selected SNPs, LD-blocks and genes across all bootstraps

Dataset	Stability index at SNP level	Stability index at LD-block level	Stability index at gene level
RA	0.829	0.833	0.810
T1D	0.851	0.770	0.753
T2D	0.860	0.859	0.870

Table 3.3: For single-marker analyses, the stability indexes at different genomic scales: SNP level, LD-block level and gene level

selected features correspond to the significant SNPs with p-values exceeding the Bonferroni threshold in each bootstrapped sample. Table 3.2 shows the average number of selected SNPs, LD-blocks of strongly correlated SNPs and genes along all the bootstraps. Table 3.3 demonstrates that the stability of the selection in traditional GWAS methods is robust for all genomic scales, as the variability of selected sets across samples is very small. However, these approaches remain limited in recovering regions of interest associated with the disease. Indeed, very few genes associated with the phenotype were discovered (see Table 3.2).

3.4.3 Lasso and Elastic Net lead to better biological interpretation for biomarker discovery

Here, we examine the results for RA dataset in detail. Note that the results and the observations in T1D and T2D datasets are similar to those in RA study. We discuss results obtained in T1D and T2D in Appendix C.

In this section, we evaluate the stability of the selection for machine learning methods, i.e., Lasso and Elastic Net. The empirical quantification of the stability for Lasso presented in Table 3.4 shows lower stability index values compared to classic GWAS methods, especially at the SNP-level. However, we observe that the stability values at the LD-block level and the gene level increase remarkably. Such an observation demonstrates that the robustness of the selection in the LD-block level and the gene level allows better performance using Lasso. From a biological point of view, selecting a SNP that belongs to an LD-block is similar to select all SNPs of that LD-block. In other words,

Dataset	Stability index at SNP level	Stability index at LD-block level	Stability index at gene level
RA	0.379	0.567	0.451
T1D	0.404	0.539	0.468
T2D	0.317	0.553	0.292

Table 3.4: For Lasso, the stability indexes at different genomic scales: SNP level, LD-block level and gene level

all SNPs in the same LD-block explain the same biological information. We underline also that feature selection models such as Lasso select more SNPs, LD-blocks and mainly genes compared to traditional GWAS methods (see Table 3.4). Consequently, Lasso recovers markers that classical single-marker analyses have missed. At the gene level, results remain unclear to compare with the SNP-level and the LD-block level because some selected SNPs were not mapped to genes as mentioned before in Section 3.4.1. For RA, Figure 3.1 illustrates the number of selected SNPs, LD-blocks and genes for different values of λ. The choice of the best λ is a compromise between the stability and the prediction error. Figure 3.2 shows that the best trade-off between the greatest stability for the smallest error is given by $\lambda = 0.013$, that results in selecting 231 SNPs, 138 LD-blocks and 22 genes. Figure 3.3 presents the stability index against the average prediction error, it illustrates the trade-off to consider between both metrics to choose the best regularization λ. For $\lambda = 0.013$, the average error is equal to 0.108 and results stability indexes of 0.331 at the SNP level, of 0.484 at the LD-block level and of 0.451 at the gene level. The empirical results of Lasso for the phenotypes T1D and T2D are presented and evaluated respectively in Appendix C.2 and Appendix C.3.

Next, we implement Elastic Net that is known to improve the stability of Lasso by adding an ℓ_2-norm penalty. Indeed, metrics in Table 3.5 highlight a notable gain of stability at all studied genomic scales, and in particular at the LD-block level. Similarly to Lasso, Elastic Net selects higher number of discovered genes compared to single-marker analyses. For Elastic Net, the choice of the best λ_2 is also a trade-off that offers better stability for lower error. Figure 3.6 demonstrates that the optimal value of $\lambda_2 = 0.028$, producing an error of 0.005 and resulting in stability values of 0.359 at the SNP level, of 0.585 at the LD-block level and of 0.466 at the gene level. Consequently, Figure 3.5 shows that for the chosen best $\lambda_2 = 0.028$, we select 1 001 SNPs, 800 LD-blocks and 158 genes. Figure 3.5 presents the obtained values of the stability index at the three genomic scales across the different values of lambda.

The results of Elastic Net for the phenotypes T1D and T2D are detailed respectively in Appendix C.4 and C.5.

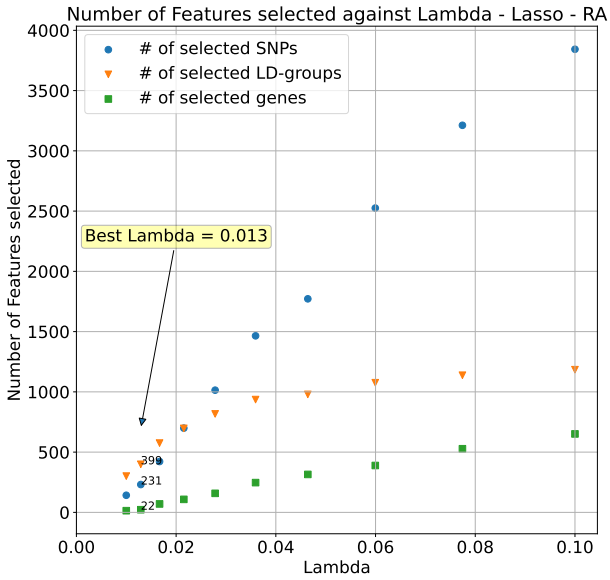


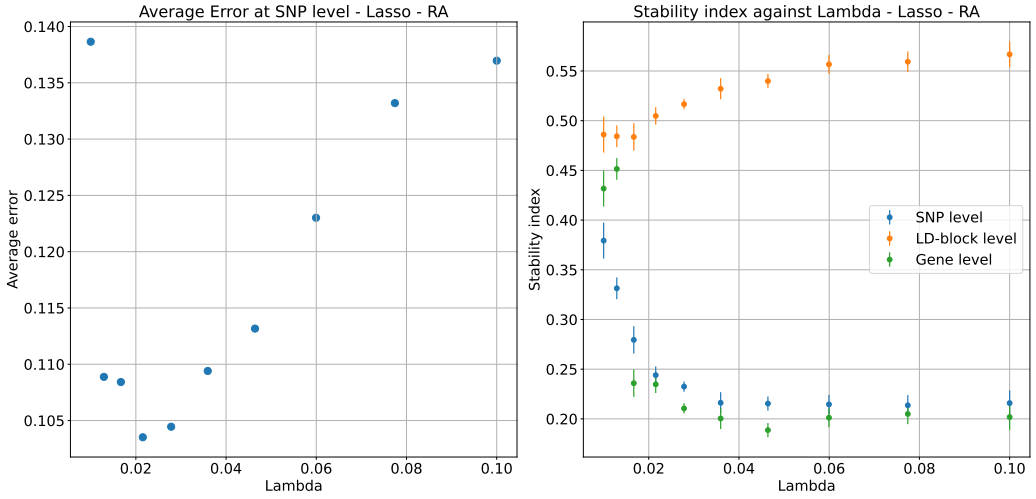
Figure 3.1: For Lasso, number of selected SNPs, LD-blocks and genes against values of lambda

Dataset	Stability index at SNP level	Stability index at LD-block level	Stability index at gene level
RA	0.431	0.602	0.496
T1D	0.425	0.590	0.521
T2D	0.405	0.612	0.466

Table 3.5: For Elastic Net, the stability index values at different genomic scales: SNP level, LD-block level and gene level

3.4.4 Stability selection methods increase the stability index of Lasso

In this section, we study the stability of both stability selection methods (presented in Section 3.2.5). First, we find that [Meinshausen and Bühlmann(2009)] approach improves the stability index of Lasso for all studied phenotypes at the different genomic scales, as shown in Table 3.6. The obtained stability values given in Table 3.7 using [Shah and Samworth(2013)] method are higher than those obtained with the first approach and any other tested feature selection model. However, [Shah and Samworth(2013)] method is very intensive computationally compared to the other methods and requires higher memory resources due to the CPSS procedure. It is important to mention that both stability selection approaches restrict the number of selected variables as compared to basic feature selection methods, which reduce the false positive rate.



(a) The average error against lambdas.

(b) The stability index at different genomic scales (SNP, LD-block and gene) against lambdas.

Figure 3.2: For Lasso, the average error and stability index for different values of lambdas

3.5 Discussion and conclusion

We have presented in this chapter an empirical evaluation approach to compare the stability of the selection measurement of common methods in GWAS: single-marker analyses and Lasso. The stability of the selection is an important criterion to evaluate the efficiency of a feature selection model in identifying causal variants. In fact, considering false candidate SNPs selected randomly produces high rate of false discoveries, and leads consequently to wrong biological interpretation. Thus, the robustness of the selection procedure is essential to obtain a meaningful selected set of features that will not be affected by the variability among samples of the input dataset. Our

Dataset	Stab at SNP (# of sel SNPs)	Stab at LD-block (# of sel LD-blocks)	Stab index at gene (# of mapped SNPs)
RA	0.493 (524)	0.602 (483)	0.482 (96)
T1D	0.581 (435)	0.650 (386)	0.561 (77)
T2D	0.554 (621)	0.623 (540)	0.500 (81)

Table 3.6: For Lasso, the stability index values obtained at different genomic scales (SNP level, LD-block level and gene level) after stability selection using [Meinshausen and Bühlmann(2009)] method. In brackets, the number of selected/mapped features is given for each studied level

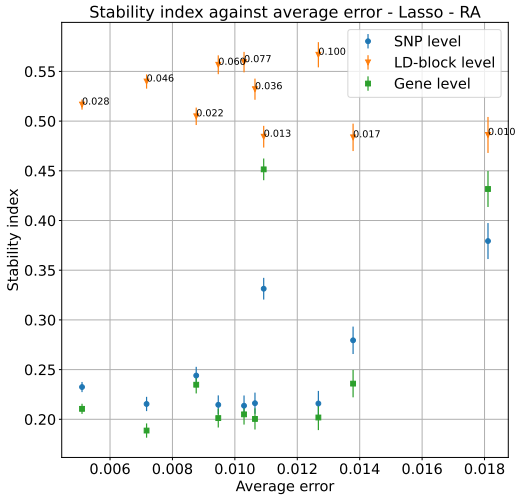


Figure 3.3: For Lasso, the stability index at different genomic scales (SNP, LD-block and gene levels) against the average error

Dataset	Stab at SNP (# of sel SNPs)	Stab at LD-block (# of sel LD-blocks)	Stab index at gene (# of mapped SNPs)
RA	0.521 (498)	0.614 (428)	0.479 (75)
T1D	0.593 (423)	0.704 (359)	0.587 (68)
T2D	0.620 (599)	0.664 (512)	0.521 (71)

Table 3.7: For Lasso, the stability index values obtained at different genomic scales(SNP level, LD-block level, gene level) after stability selection using [Shah and Samworth(2013)] method. In brackets is given the number of selected/mapped features at each studied level

analysis show that traditional GWAS based on single-marker analyses remain the leading approach to avoid false discoveries, as it gives the best stability measurements in all studied datasets. However, it is essential to realize that the classical GWAS test is limited to an univariate test that often detects very few associations with the phenotype.

We show in this work effective alternatives to improve the stability of feature selection methods by using Elastic Net or stability selection techniques. Indeed, these approaches improve significantly the robustness of the selection with close stability indexes of single-marker analyses and better biological information.

The novelty of this contribution is the quantification of the stability of the selection at various genomic scales, i.e., SNP level, LD-block level and gene level. Our study helps to explore the ability of feature selection models in targeting common regions in higher genomic scales. Hence, we have discovered one of the causes behind the lack of stability in regularization techniques, such

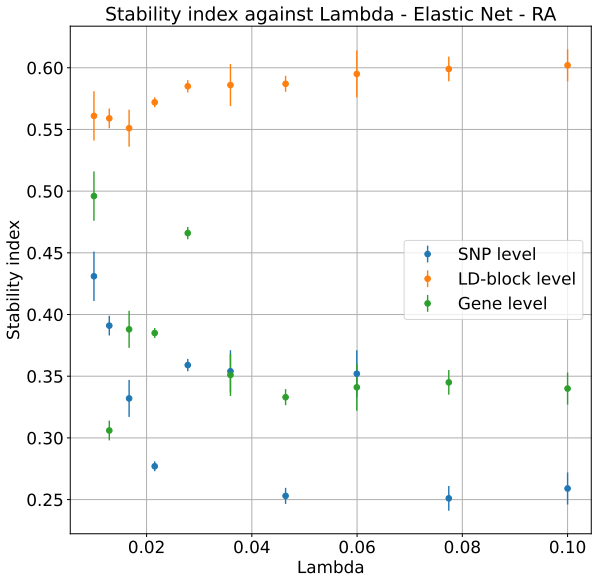


Figure 3.4: For Elastic Net, the average error and stability index for different values of lambda

as Lasso. In other words, some SNPs were not selected repeatedly along the bootstrapped samples, but they were in the same LD-block as other SNPs that were selected also by the same model. As a consequence, we obtain much lower values of stability index at the SNP level than at a higher level. Indeed, we observe that the stability of the selection increases at the LD-block and gene scales. From a biological point of view, selecting any SNP that belongs to the same LD-block results in the same interpretation. Hence, doing the selection process at the LD-block level results in higher stability without losing any biological information.

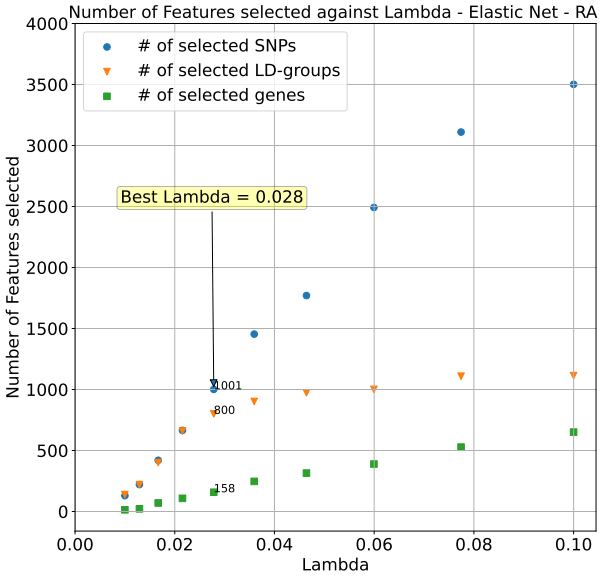


Figure 3.5: For Elastic Net, number of selected SNPs, LD-blocks and genes against the values of lambda

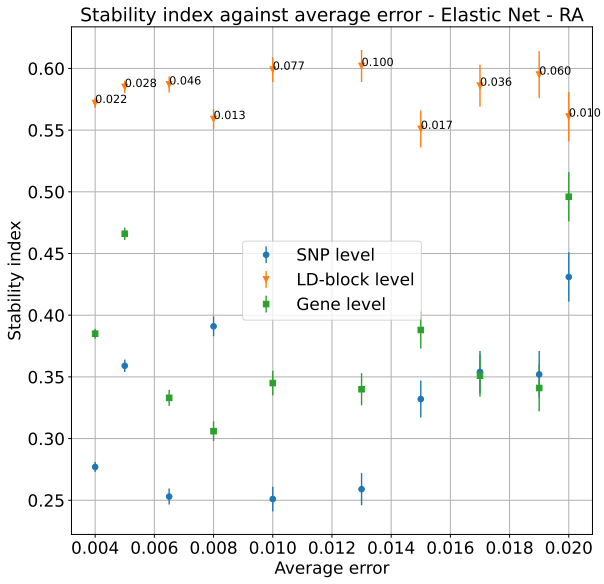


Figure 3.6: For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene levels) against the average error

Chapter 4

Multitask group Lasso for Genome-Wide Association Studies in diverse populations

---

Abstract: *Genome-Wide Association Studies, or GWAS, aim at finding Single Nucleotide Polymorphisms (SNPs) that are associated with a phenotype of interest. GWAS are known to suffer from the large dimensionality of the data with respect to the number of available samples. Other limiting factors include the dependency between SNPs, due to linkage disequilibrium (LD), and the need to account for population structure, that is to say, confounding due to genetic ancestry. We propose an efficient approach for the multivariate analysis of multi-population GWAS data based on a multitask group Lasso formulation. Each task corresponds to a subpopulation of the data, and each group to an LD-block. This formulation alleviates the curse of dimensionality, and makes it possible to identify disease LD-blocks shared across populations/tasks, as well as some that are specific to one population/task. In addition, we use stability selection to increase the robustness of our approach. Finally, gap safe screening rules speed up computations enough that our method can run at a genome-wide scale. To our knowledge, this is the first framework for GWAS on diverse populations combining feature selection at the LD-groups level, a multitask approach to address population structure, stability selection, and safe screening rules. We show that our approach outperforms state-of-the-art methods on both a simulated and a real-world cancer datasets.*

Résumé: *Les études d'association pangénomiques, ou GWAS, visent à trouver des polymorphismes nucléotidiques (SNPs) associés à un phénotype d'intérêt. Les GWAS sont connus pour souffrir de la grande dimensionnalité des données par rapport au nombre d'échantillons disponibles. D'autres facteurs limitants incluent la dépendance entre les SNP, à cause du déséquilibre de liaison (LD), et la nécessité de tenir compte de la structure de la population, c'est-à-dire de la confusion due à l'ascendance génétique. Nous proposons une approche efficace pour l'analyse multivariée des données GWAS multi-population basée sur une formulation multi-tâches group Lasso. Chaque tâche correspond à une sous-population des données, et chaque groupe à un bloc LD. Cette formulation atténue le fléau de la dimension et permet d'identifier les blocs LD de la maladie partagés entre les populations/tâches, ainsi que certains qui sont spécifiques à une population/tâche. De plus, nous utilisons la*

sélection de stabilité pour augmenter la robustesse de notre approche. Enfin, les approches gap safe screening rules accélèrent suffisamment les calculs pour que notre méthode puisse fonctionner à l'échelle du génome. À notre connaissance, notre méthode est la première approche proposée pour les GWAS sur les populations diverses combinant la sélection de variables au niveau des groupes LD, une approche multitâche pour traiter la structure de la population, la sélection de stabilité et les approches gap safe screening rules. Nous montrons que notre approche surpasse les méthodes existantes sur des ensembles de données simulées et réelles de cancer.

Contents

4.1 Introduction 72
4.2 Methods 73
4.2.1 Population stratification 73
4.2.2 Linkage disequilibrium groups 74
4.2.3 Multitask group Lasso 74
4.2.4 Stability selection 77
4.3 Experiments 77
4.3.1 Data 77
4.3.2 Preprocessing 78
4.3.3 Comparison partners 78
4.4 Results 79
4.4.1 MuGLasso draws on both LD-groups and the multitask approach to recover disease SNPs 79
4.4.2 MuGLasso provides the most stable selection 80
4.4.3 MuGLasso selects both task-specific and global LD-groups 81
4.5 Discussion and Conclusions 81

4.1 Introduction

Over the last 15 years, Genome-Wide Association Studies (GWAS) have become one of the most prevalent methods to identify regions of the genome associated with complex phenotypic traits, and in particular complex diseases in humans [Visscher *et al.*(2017)]. One of the major concerns in GWAS is population stratification, which arises when allele frequency differences between cases and controls are due to differences in genetic ancestry rather than to association with the phenotype. Many correction methods have been proposed to adjust the inflation of associations in diverse populations, including methods based on principal components analysis or on linear mixed models [Yiwei and Wei(2015)]. However, it is possible that these techniques lead to overcorrection, in particular by masking population-specific disease loci.

An additional issue in GWAS is Linkage Disequilibrium (LD), which manifests as correlation between adjacent Single Nucleotide Polymorphisms (SNPs), creating statistical dependence between those markers and reducing statistical power [Dehman *et al.*(2015)]. Combining strongly correlated SNPs into blocks, that is to say, groups of adjacent and correlated SNPs, and modeling the association signal over an entire region, is one way to address this limitation.

Classical approaches for GWAS are based on single-marker analyses, testing for association between each SNP and the phenotype independently. This may prevent the detection of effects that are due to SNPs acting additively, leading many authors to favor fitting a linear model to all SNPs jointly [Sebastian *et al.*(2014)]. Penalized regression approaches, such as the Lasso, which uses an ℓ_1-norm regularization to shrink some coefficients of the model to zero, effectively removing them from the model, are particularly suited to this task.

Additional regularizers can be used to enforce additional prior hypotheses on the coefficients of such a linear model. Among them, the group Lasso [Yuan and Lin(2006),Dehman *et al.*(2015)] ensures sparsity at the level of pre-defined groups of features, and the multitask Lasso [Obozinski *et al.*(2006), Kriti *et al.*(2010)] fits models on related tasks jointly, encouraging similar sparsity patterns across all tasks.

In this work, we propose to combine both approaches into a multitask group Lasso framework, in which groups correspond to pre-defined LD patterns, and each task corresponds to a subpopulation, therefore simultaneously addressing the limitations of single-marker analyses and the issues of both LD and population structure.

In addition, we draw on the stability selection framework [Meinshausen and Bühlmann(2009)] to improve the stability of the results, that is to say, their robustness to small perturbations in the input data, such as the removal of a few samples. Indeed, because the number of SNPs is typically much larger than that of samples, penalized regression approaches tend to select different

sets of SNPs when presented with slightly different subsets of the same data, which severely limits their interpretability.

Finally, we use the recently proposed gap safe screening rules proposed by E. Ndiaye et al. [Ndiaye *et al.*(2017)] to improve computational complexity, and scale our approach to about one million SNPs.

In what follows, we present our proposed approach in detail, place it in the context of existing work, and evaluate it on both a simulated data set and a real-world cancer GWAS data set.

4.2 Methods

Our proposed approach, MuGLasso, follows four steps, which we detail in this Section. First, we assign each sample to a genetic population, hence forming different but related tasks (Section 4.2.1). Second, we create LD-groups from correlations between SNPs, so as to perform feature selection at the level of groups rather than individual SNPs(Section 4.2.2). Third, we jointly fit one regularized model per task, using an $\ell_{2,1}$ penalty that enforces sparsity at the level of LD-groups (Section 4.2.3). Finally, we use stability selection to improve the robustness of the solution (Section 4.2.4).

4.2.1 Population stratification

Population structure, whereby the data is made of subsets of individuals that differ systematically both in genetic ancestry and in the phenotype under investigation, is a major confounding factor in GWAS. Indeed, it leads to detecting allele frequency differences in cases and controls that correspond to differences in ancestry, instead of a more direct association between genotype and phenotype. Several approaches have been developed to adjust for population structure.

Among them, a large number of methods rely on Principal Component Analysis (PCA) [Zeggini *et al.*(2008), Need *et al.*(2009), Price *et al.*(2006)], and consist of including top Principal Components (PCs) of the genotypes as covariates in regression models. In addition, linear mixed models [Yu *et al.*(2006)] can be used to model the phenotype as a combination of fixed and random effects, with the covariance of the latter being computed from a genetic similarity matrix. Although they often outperform PCA-based methods, the mixed model approaches tend to be more computationally demanding. Both approaches are similar in that regressing out principal components can be seen as approximation of a linear mixed model [Yiwei and Wei(2015)].

However, these techniques may lead to ignoring population-specific SNPs, which is why we propose a multitask approach that can identify disease loci that are either population-specific or shared between populations. We there-

fore form tasks by separating the data into subpopulations, identified as clusters (using k-means clustering) on the projection of the genotypes on their top PCs.

4.2.2 Linkage disequilibrium groups

Linkage disequilibrium (LD) is the non-random association of alleles of at least two loci [Slatkin(2008)]. LD can be leveraged to form groups of correlated SNPs. Grouping SNPs helps to alleviate the curse of dimensionality in GWAS by reducing the number of testing possibilities. This can be achieved by combining p-values within a group of correlated SNPs [Hu *et al.*(2016)], or through the use of penalized regression approaches that perform feature selection at the level of groups, rather than at the level of individual SNPs [Dehman *et al.*(2015)]. The latter has the advantage over individual statistical testing of modeling the additive effects of multiple genetic markers simultaneously.

Adjacency-constrained hierarchical clustering

In many species, including humans [Reich *et al.*(2001)], LD is known to be correlated to the physical distance between SNPs. Hence, genomes can be clustered in LD blocks of strongly correlated adjacent SNPs, called in this chapter LD-groups. Such LD-groups can be obtained using adjacency-constrained hierarchical agglomerative clustering [Ambroise *et al.*(2019)], in which only physically adjacent clusters can be merged. We detailed the clustering method in Section 3.2.3.

LD-groups across populations

Because LD patterns may be influenced by genetic ancestry [Boehnke(2000)], we perform LD-groups partitioning for each population separately. We then combine those LD-groups into common shared LD-groups. More specifically, the set of coordinates of the boundaries of the shared LD-groups is obtained as the union of the sets of coordinates of the boundaries of the LD-groups for each population. This procedure is described in Supplementary Figure D.1.

4.2.3 Multitask group Lasso

General framework and problem formulation

We use a penalized regression approach to fit a multivariate linear model between the phenotype and the SNPs, with a regularization term that ensures that (1) the solution is sparse at the level of LD-groups and (2) the regression coefficients are smoothed within groups and across tasks. Such an approach

provides shared LD-groups associated with the phenotype across all tasks, and allows for some LD-groups to be specific to each task.

Problem formulation Given a set of p SNPs measured for n samples, we split the n samples in T subpopulations/tasks, each of size n_t for $t = 1, \ldots, T$, and the p SNPs in G LD-groups, each of size p_g for $g = 1, \ldots, G$. For each population t, we denote by $\boldsymbol{x}_m^{(t)}$ the p-dimensional vectors of SNPs of the m-th sample in the population ($m = 1, \ldots, n_t$), and by $y_m^{(t)}$ its phenotype. We then formulate the following optimization problem:

$$\min_{B \in \mathbb{R}^{p \times T}} \frac{1}{n} \sum_{t=1}^{T} \sum_{m=1}^{n_t} \mathcal{L}\left(y_m^{(t)}, \sum_{j=1}^{p} \beta_j^{(t)} x_{mj}^{(t)}\right) + \lambda \sum_{g=1}^{G} \sqrt{p_g} \left\|B^{(g)}\right\|_F, \tag{4.1}$$

where $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^p$ is the vector of regression coefficients specific to task t : $\boldsymbol{\beta}^{(t)} = (B_{1t}, \ldots, B_{pt})$, $\mathcal{L}$ is the quadratic loss if the phenotype is quantitative ($y \in \mathbb{R}$) and the logistic loss if it is qualitative ($y \in \{0, 1\}$), $\|.\|_F$ denotes the Frobenius norm, and $B^{(g)}$ is a $p_g \times T$ matrix containing the regression coefficients, across all tasks, for the SNPs of group g. Hence the penalization term ties the regression coefficients across tasks and groups, and ensures sparsity at the group level. The penalization parameter $\lambda > 0$ controls the amount of sparsity. Note that to fit an intercept, it is sufficient to add a feature that is equal to 1 to each sample.

Related work

$\ell_{2,1}$-norm regularization Our approach is closely related to the group Lasso [Yuan and Lin(2006)] and multitask Lasso [Obozinski *et al.*(2006)], which both make use of an $\ell_{2,1}$-norm regularization. More precisely, the group Lasso corresponds to a special case of Equation (4.1), with a single task ($T = 1$), resulting in sparsity at the group levels. Using a group Lasso where the groups are defined based on LD blocks has been successfully applied to GWAS on up to 20 000 SNPs [Dehman *et al.*(2015)]. The multitask Lasso corresponds to a special case of Equation (4.1), with each group containing exactly one SNP. This formulation ties sparsity patterns across tasks and has been applied before to multi-population GWAS, although only a few thousand SNPs [Kriti *et al.*(2010)].

The multitask group Lasso we propose can also be reformulated as an $\ell_{2,1}$-norm regularization problem, through the creation of a new dataset $(\widetilde{X}, \widetilde{\boldsymbol{y}})$ where $\widetilde{X} \in \mathbb{R}^{n \times pT}$ is a block-diagonal matrix such that each of the T diagonal blocks corresponds to the SNP matrix $X^{(t)} \in \mathbb{R}^{n_t \times p}$ for task t, and $\widetilde{\boldsymbol{y}}$ is a n-dimensional vector obtained by stacking the phenotype vectors for each task.

Equation (4.1) can then be rewritten as:

$$\min_{\boldsymbol{b}\in\mathbb{R}^{pT}} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\widetilde{y}_i, \sum_{k=1}^{pT} b_k \widetilde{x}_{ik}\right) + \lambda \sum_{g=1}^{G}\sqrt{p_g}\left\|\boldsymbol{b}^{(g)}\right\|_2, \tag{4.2}$$

with $\boldsymbol{b}^{(g)} \in \mathbb{R}^{p_g T}$ the regression coefficients corresponding to all SNPs of group (g) for all tasks. In essence, this is a group Lasso with G groups each containing T copies (one per task) of the p_g features of SNP group g. Thus $B_{jt} = \boldsymbol{b}_{p(t-1)+j}$.

Other multitask group Lassos Other authors have proposed variations on the idea of a multitask group Lasso before. Several publications [Wang *et al.*(2012), Lin *et al.*(2014)] add a second regularization term to our formulation, increasing within-group or across-task sparsity. Unfortunately, this dramatically increases computational time, and indeed none of these publications analyze genome-wide data sets. In addition, because interpretation will be done at the group level rather than at the SNP level, within-group sparsity is not necessarily desirable in this context.

Several authors have built on these propositions and add a third regularization term, either enforcing group-independent task sparsity [Xiaoli *et al.*(2017)] or overall sparsity (with an ℓ_1-norm over all coefficients) [Li *et al.*(2020)]. Again, the addition of these regularizers severely hinders the applicability of these methods at a genome-wide scale due to computational limitations.

Hence none of these methods is readily applicable to our setting. In addition, their stability has never been evaluated, even though it is an important criterion for the reliability and interpretability of the results (see Section 4.2.4).

Gap safe screening rules

To speed up the computation of the solution of Equation (4.2), we call upon gap safe screening rules [Ndiaye *et al.*(2017)], which are used to efficiently identify features for which the regression coefficients will be zero and hence ignore them when solving the problem. Such screening rules have been proposed for a large number of popular regularized regressions [Ndiaye *et al.*(2017)], including $\ell_{2,1}$-norm regularizations. In particular, Equation (4.2) can be solved using the `Gap_Safe_Rule` package[1]. We briefly summarize the idea underlying gap safe screening rules in Appendix D.2.3.

[1] `https://github.com/EugeneNdiaye/Gap_Safe_Rules`

4.2.4 Stability selection

Unfortunately, in GWAS, penalized regression approaches often lack stability, that is to say, robustness to slight variations in the input dataset [Alexander and Lange(2011)]. However, stability increases both the reliability of the results and the interpretability. To address this limitation, *stability selection* [Meinshausen and Bühlmann(2009), Alexander and Lange(2011)] consists of performing feature selection repeatedly on subsamples of the data and only retains the features most often selected. More specifically, given a subsample $I \subset \{1, \dots, n\}$ of size $\lfloor n/2 \rfloor$ of the data, we call $\widehat{S}^{\lambda}(I)$ the set of features selected by the selection procedure of interest (for example, a Lasso), with hyperparameter λ, on this subsample of the data. For any feature $j \in \{1, \dots, p\}$, we call $\widehat{\Pi}_j^{\lambda}$ the probability that feature j is selected on a random subsample of size $\lfloor n/2 \rfloor$ of the data. This probability is determined, given m such random subsamples $I_1, I_2, \dots, I_m$, as the proportion of those subsamples for which the feature selection procedure selects feature j: $\widehat{\Pi}_j^{\lambda} = \frac{1}{m} \sum_{k=1}^{m} \mathbf{1}_{j \in S^{\lambda}(I_k)}$. Finally, given a a threshold $\frac{1}{2} < \pi_{\text{cutoff}} \leq 1$ (in this work, we used $\pi_{\text{cutoff}} = 0.75$), the stable set of selected features is determined as $\widehat{S}^{\text{stable}} = \{j : \max_{\lambda \in \Lambda} \widehat{\Pi}_j^{\lambda} \geq \pi_{\text{cutoff}}\}$.

We presented in detail stability selection methods in Section 3.2.5.

4.3 Experiments

4.3.1 Data

Simulated data Using GWAsimulator [Li and Li(2008)], we simulated GWAS data with realistic LD patterns from two populations (CEU : Utah residents with Northern and Western European ancestry and YRI: Yoruba in Ibadan, Nigeria) of the HapMap 3 data. We induced population structure by varying the case:control ratio within each subpopulation (CEU 1 100:900 and YRI 900:1 100), as well as by simulating population-specific disease loci. We simulated a total of 149 970 disease SNPs, 2 999 (resp 4 999) of which are specific to the CEU (resp. YRI) population (see Appendix D.1.1). The data contains 4 000 samples and 1 400 000 SNPs.

DRIVE Breast Cancer OncoArray The DRIVE OncoArray dataset (dbGap study accession phs001265/GRU) contains 28 281 individuals that were genotyped for 582 620 SNPs. 13 846 samples are cases and 14 435 are controls. More details are available in Appendix D.1.2.

4.3.2 Preprocessing

Quality control and imputation We removed SNPs with a minor allele frequency lower than 5%, a p-value for Hardy-Weinberg Equilibrium in controls lower than 10%, or a missing genotyping rate larger than 10%. We removed duplicate SNPs and excluded samples with more than 10% of SNPs missing. We imputed missing genotypes in DRIVE using IMPUTE2 [Howie *et al.*(2009)].

LD pruning We performed LD pruning using PLINK [Purcell *et al.*(2007)] with a LD cutoff of $r^2 > 0.85$ and a window size of 50Mb, both to reduce the number of SNPs and to better capture population structure using PCA [Abdellaoui *et al.*(2013)]. 1 000 000 SNPs remain in the simulated data and 313 237 in DRIVE.

PCA and population structure We used PLINK [Purcell *et al.*(2007)] to compute principal components of the genotypes. We thus identify two populations in the simulated data, matching the CEU and YRI populations (see Supplementary Figure D.3a). In DRIVE, we identify two populations (see Supplementary Figure D.3b), which we call POP1 (samples from the USA, Australia and Denmark) and POP2 (samples from the USA, Cameroon, Nigeria and Uganda).

LD-groups choice We obtain LD-groups for each of the PCA-based populations using adjclust [Ambroise *et al.*(2019)] and obtain shared LD-groups as described in Section 4.2.2. Table 4.1 shows the number of LD-groups obtained for each subpopulation and the final number of shared groups.

Data	Subpopulations	# of LD-groups	# of shared LD-groups
Simulated data	CEU YRI	25 281 15 636	35 792
DRIVE real data	POP1 POP2	8 152 5 032	17 782

Table 4.1: For each subpopulation of both datasets (simulated and real), LD-groups number is given and the shared LD-groups number after combination

4.3.3 Comparison partners

As a baseline, we use PLINK [Purcell *et al.*(2007)] to perform tests of association between each SNP and the phenotype, either using the top PCs as covariates (**Adjusted GWAS**), or treating each population separately (**Stratified**

GWAS). We also compute a PCA-adjusted phenotype as the residuals of a regression between the top PCs and the phenotype. To evaluate the effects of grouping correlated SNPs and separating the populations in tasks, we compare MuGLasso to a Lasso (single task, no groups) on each population (**Stratified Lasso**) or on the adjusted phenotype (**Adjusted Lasso**), as well as a group Lasso (single task) on each population (**Stratified group Lasso**) or on the adjusted phenotype (**Adjusted group Lasso**).

For computational efficiency, we use bigLasso [Yaohui and Patrick(2017)] for the Lasso, and `Gap_Safe_Rule` [Ndiaye *et al.*(2017)] for the group Lasso. For all methods, we set the regularization hyperparameter by cross-validation.

To compare these methods, we report runtime, ability to recover true causal SNPs (in the case of simulated data), and stability of the selection. To measure selection stability, we repeat the feature selection procedure on 10 subsamples of the data, and report the average Pearson's correlation between all pairs of indicator vectors representing the selected features for each subsample (see Appendix D.2.4 for details).

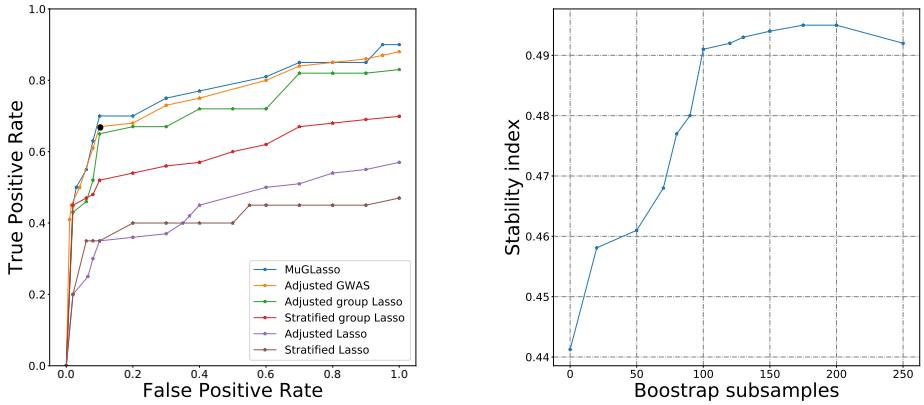
4.4 Results

4.4.1 MuGLasso draws on both LD-groups and the multitask approach to recover disease SNPs

On the simulated data, we observe (Figure 4.1a) that MuGLasso is better than any other method at recovering the true disease SNPs. Performing feature selection at the level of LD-groups, rather than individual SNPs, improves performance. Indeed, the group Lassos and MuGLasso outperform the SNP-level Lassos. In addition, treating all samples simultaneously (as in MuGLasso or the adjusted approaches) also improves performance. This confirms our hypothesis that grouping features and using all samples simultaneously both alleviate the curse of dimensionality.

On DRIVE, MuGLasso recovers 1 051 SNPs in addition to all SNPs from the adjusted GWAS. They point to 32 risk genes that cannot be identified by the classical GWAS; half of those have been identified in meta-GWAS that included our samples, and another 7 have been associated with breast cancer risk or growth in other studies (see Supplementary Table D.3).

However, this increased ability to recover relevant SNPs comes with an increase in computational time (see Supplementary Figure D.4 on simulated data and Figure 4.2a on DRIVE). However, the implementation is efficient enough to allow computations on 10^6 SNPs, even with the added cost of repeated subsampling to increase stability.



(a) Recovery of disease loci

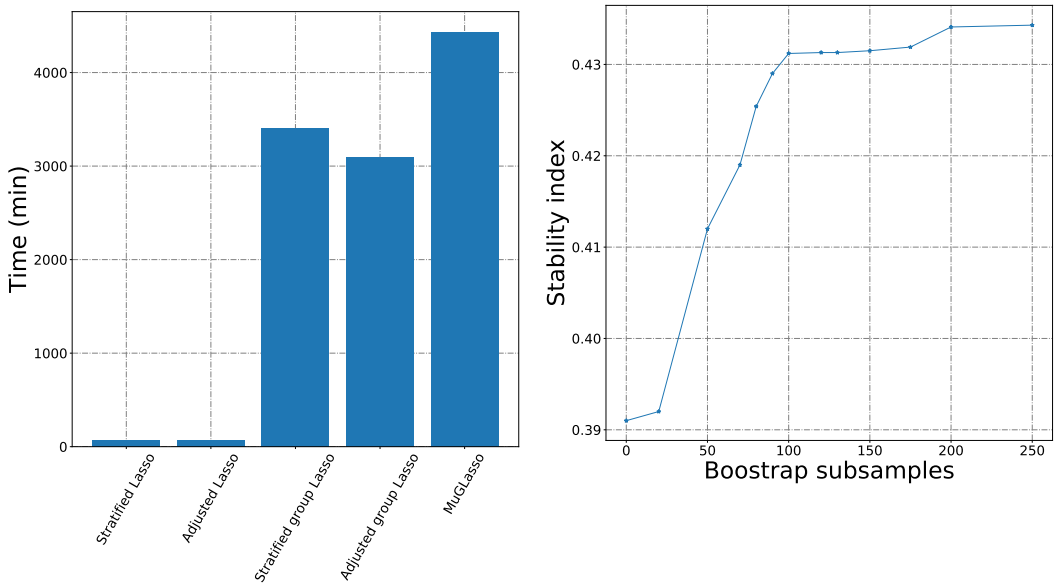
(b) Stability of MuGLasso

Figure 4.1: On simulated data, ability of different methods to retrieve causal disease SNPs as a ROC plot (4.1a), and stability index of MuGLasso as a function of the number of bootstrap samples (4.1b). On the ROC plot, the black dot indicates the performance of the stratified GWAS at the Bonferonni-corrected significance threshold.

4.4.2 MuGLasso provides the most stable selection

Figures 4.1b (simulated data) and 4.2b (DRIVE) show the stability index of MuGLasso as a function of the number of subsamples. Increasing the number of subsamples increases the stability of the selection. We use 100 bootstrap samples in all subsequent experiments as it appears to be an acceptable trade-off between runtime and stability.

Tables 4.2 and 4.3 show the stability index of the different methods, on simulated data and DRIVE, respectively. We ran the adjusted GWAS once on the entire data set, as would usually be done, and therefore cannot report its stability. Our results again illustrate that stability selection does increase the stability of Lasso methods. We confirm this by running MuGLasso without stability selection as well as Adjusted group Lasso with stability selection on top. In both cases, the stability index increases when stability selection is used. In addition, we report the total number of selected SNPs and LD-groups. For methods that select individual SNPs, we obtain the number of selected LD-groups by considering that each selected SNP selects its entire LD-group. Our results illustrate that the improved stability of MuGLasso does not come at the expense of selecting more features. On the contrary, stability selection provides fewer SNPs/LD-groups with better stability.



(a) Runtimes of the Lasso approaches

(b) Stability of MuGLasso

Figure 4.2: On DRIVE, runtimes of the different Lasso approaches (4.2a) and stability index of MuGLasso as a function of the number of bootstrap samples (4.2b).

4.4.3 MuGLasso selects both task-specific and global LD-groups

For both datasets, the LD-groups selected by MuGLasso are a mixture between population-specific LD-groups (identified as those with near-zero regression coefficients for one task) and LD-groups that are shared between both populations. Table 4.4 shows the number of LD-groups/SNPs in each of these categories for MuGLasso. By contrast, the adjusted approaches do not provide population-specific LD-groups or SNPs.

Finally, we report on Figure 4.3 the precision and recall of MuGLasso and the stratified approaches on the population-specific SNPs. MuGLasso outperforms all other approaches in both precision and recall.

4.5 Discussion and Conclusions

We presented MuGLasso, an efficient approach for detecting disease loci in GWAS data from diverse populations. Our approach is based on a multi-task framework, where input tasks correspond to subpopulations, and feature selection is performed at the level of LD-groups. Assigning samples from

Methods	# of selected LD-groups	# of selected SNPs	Stability index	Selection level
MuGLasso	5 623	155 312	0.4912	LD-groups
MuGLasso without stab sel	6 124	161 221	0.4412	LD-groups
Adjusted group Lasso + stab sel	6 054	162 104	0.4134	LD-groups
Adjusted group Lasso	6 347	167 204	0.3714	LD-groups
Stratified group Lasso	4 836	154 732	0.3398	LD-groups
Adjusted Lasso	5 379	158 856	0.2368	Single-SNP
Stratified Lasso	5 704	168 158	0.1742	Single-SNP
Adjusted GWAS	5 063	141 340	-	Single-SNP

Table 4.2: Stability index and number of selected features for different methods, on simulated data

PCA-identified populations to different tasks addresses the issue of population stratification, and retains the flexibility of identifying population-specific disease loci. Treating all samples together, by contrast with stratified approaches, alleviates the curse of dimensionality. Ensuring sparsity at the level of LD-groups addresses the high correlation between nearby SNPs and also alleviates the curse of dimensionality. Although more time-consuming than a classical GWAS, our implementation is computationally efficient enough to scale to the analysis of entire GWAS data sets of about one million SNPs.

On simulated data, MuGLasso outperforms state-of-the-art approaches in its ability to recover disease loci. This also holds for population-specific SNPs; hence performance is not driven solely by the ability to recover disease loci that are common to all populations. In addition, MuGLasso is the most stable of all evaluated method, which increases interpretability.

Finally, although we presented MuGLasso in the context of admixed populations, our tool could be used in other multitask settings. In particular, tasks can stem from related phenotypes [Wang *et al.*(2012)] or from different studies pertaining to the same trait, in a meta-analysis approach [Lin *et al.*(2014)]. Groups could also be defined according to different prior biological knowledge, for example based on functional units such as genes, in the spirit of gene-set analyses of GWAS data. In addition, although we only presented results on case-control studies with two populations, the method directly applies to quantitative phenotypes and any number of tasks.

An important outcome of our study is that, although we have not included in MuGLasso a regularization term that would enforce sparsity at the level of tasks as in [Li *et al.*(2020)], we still obtain task-specific groups. Including such an additional term in Equation (4.1) would perhaps improve the already

Methods	# of selected LD-groups	# of selected SNPs	Stability index	Selection level
MuGLasso	62	1 357	0.4312	LD-groups
MuGLasso without stab sel	72	1 524	0.3911	LD-groups
Adjusted group Lasso + stab sel	59	1 293	0.3234	LD-groups
Adjusted group Lasso	68	1 466	0.2613	LD-groups
Stratified group Lasso	58	1 119	0.2498	LD-groups
Adjusted Lasso	41	874	0.2068	Single-SNP
Stratified Lasso	38	789	0.1581	Single-SNP
Adjusted GWAS	16	306	-	Single-SNP

Table 4.3: Stability index and number of selected features for different methods, on DRIVE

Data	Population	# of selected LD-groups (and SNPs)
Simulated data	CEU	95 (2 418 SNPs)
	YRI	103 (3 081 SNPs)
	shared (CEU and YRI)	5 227 (149 813 SNPs)
DRIVE	POP1	6 (148 SNPs)
	POP2	2 (43 SNPs)
	shared (POP1 and POP2)	54 (1166 SNPs)

Table 4.4: For MuGLasso, number of selected LD-groups/SNPs, across and per population

state-of-the-art task-specific precision and recall of MuGLasso, but this would unfortunately come at the expense of a notable increase in computational time, if only because of the cross-validation needed to set the value of a second hyperparameter.

An in-depth biological analysis of the loci identified by MuGLasso on DRIVE would illustrate the biological relevance of our method, but is out of the scope of this methodological approach.

In the future, we are looking forward to making use of the post-inference selection framework for group-sparse linear models [Fan *et al.*(2016)] to provide p-values for the selected loci. As of now, it is unclear how to apply these ideas to case-control studies in a computationally efficient manner.

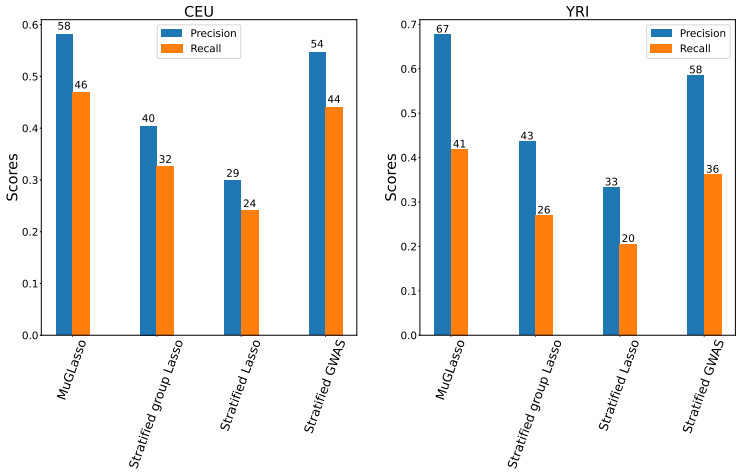


Figure 4.3: For simulated data, precision and recall of MuGLasso and the stratified approaches on the populations-specific SNPs

Code

Code is available at https://github.com/asmanouira/MuGLasso_GWAS.

Chapter 5

Sparse multitask group Lasso for Genome-Wide Association Studies in diverse populations

---

Abstract: *Among the challenges in Genome-Wide Association Studies is the population stratification that refers to the presence of differences in allele frequencies between subpopulations within samples, due to different ancestry. Moreover, diseases can have differences in prevalence across populations, thus, risk variants can differ from one genetic ancestry to another. We propose an extended approach of MuGLasso, called SMuGLasso, taking in account the presence of population-specific linkage disequilibrium groups (LD-groups). To do so, we add an additional ℓ_1-norm regularization to select causal markers in a precise and refined approach. Alike MuGLasso, the groups in SMuGLasso correspond to Single Nucleotide Polymorphisms (SNPs) in strong LD and the tasks correspond to ancestral subpopulations. We include the stability selection procedure to boost the robustness of our algorithm. We also handle the computational complexity of the method by using gap safe screening rules. We conduct our analysis in a real case-control breast cancer dataset, a real plant dataset presenting a quantitative phenotype and a simulated data.*

Résumé: *Parmi les défis des études d'association pangénomiques, il y a la structure de population qui fait référence à la présence de différences dans les fréquences alléliques entre les sous-populations au sein des échantillons, à cause d'une ascendance différente. De plus, les maladies peuvent avoir des différences de prévalence entre les populations, ainsi, les variantes de risque peuvent différer d'une ascendance génétique à l'autre. Nous proposons une approche étendue de MuGLasso, appelée SMuGLasso, prenant en compte la présence de groups LD de populations spécifiques. Dans ce but, nous ajoutons une régularisation supplémentaire de norme ℓ_1 pour sélectionner les marqueurs causaux dans une approche précise et raffinée. Comme pour MuGLasso, les groupes de SMuGLasso correspondent aux polymorphismes nucléotidiques (SNPs) en forte déséquilibre de liaison (LD) et les tâches correspondent à des sous-populations ancestrales. Nous intégrons la procédure de sélection de stabilité pour renforcer la robustesse de notre algorithme. Nous gérons également la complexité de calcul de la méthode en utilisant les approches gap safe screening rules. Nous menons notre analyse sur un jeu de*

données réel d'un phénotype qualitative qui est le cancer du sein, un jeu de données végétales présentant un phénotype quantitatif et un jeu de données simulées.

Contents

5.1 Introduction 88
5.2 Methods 89
5.2.1 Population structure 89
5.2.2 Linkage disequilibrium groups clustering 90
5.3 Sparse multitask group Lasso 90
5.3.1 Notations 90
5.3.2 Related work 90
5.3.3 General framework and problem formulation 91
5.3.4 Gap safe screening rules 92
5.4 Experiments 92
5.4.1 Data 92
5.4.2 Preprocessing 93
5.4.3 Comparison patterns 96
5.5 Results 96
5.5.1 SMuGLasso and MuGLasso rely on both LD-groups and the multitask approach to recover disease SNPs . 96
5.5.2 SMuGLasso and MuGLasso outperform the other methods in terms of stability 98
5.5.3 The selection of both task-specific and shared LD-groups 98
5.6 Discussion and conclusion 99

5.1 Introduction

Feature selection models have become a popular approach in GWAS to discover the genetic causes of many complex diseases such as cancer. A common GWAS analysis relies on analyzing genotype data, typically presented by SNPs to find association with a studied disease or a related quantitative trait. However, many factors can influence the power of identifying causal markers such as curse of dimensionality, population stratification, linkage disequilibrium and lack of stability. Thus, feature selection application needs particular attention to avoid wrong discoveries. The major challenge is to maximize the robustness of the selection in discovering regions of interest and discarding false positives.

Most of existing feature selection methods consider that causal SNPs are shared across diverse populations. Nonetheless, many studies have reported that some populations present different genes in relationship with the development of some diseases [Medina-Gomez *et al.*(2015)]. Indeed, diseases can have differences in prevalence across populations, thus, risk variants can differ from one genetic ancestry to another [Rosenberg *et al.*(2010)]. [Tishkoff *et al.*(2006)] have mentioned that Africans and Europeans do not share the same genes associated with lactase-persistence phenotype. Also, [Zubair *et al.*(2016)] have conducted a study to retrieve causal SNPs related to lipid traits in diverse populations. They have discovered novel SNPs mapping three genes in African American population that were not identified in either East Asian or European populations.

We have presented in Chapter 4 a novel framework for feature selection, called the multitask group lasso (MuGLasso), in which the groups correspond to SNPs in strong LD and the tasks correspond to ancestral subpopulations. We have shown the effectiveness of the model and its stability in retrieving causal SNPs related with breast cancer or/and its tumor growth. From a biological point of view, most of our gene findings support previous discoveries in other studies. Moreover, MuGLasso obtains task-specific LD-groups in addition to the shared ones across tasks, even without including a regularization term enforcing sparsity at the level of tasks. However, in the current model design, retrieving task-specific LD-groups needs additional post processing steps to identify LD-groups with close-to-zero regression coefficients for one task. Hence, adding a second regularization term to carry out populations-specific sparsity may improve the performance of the selection at this level.

In this chapter, we present the Sparse Multitask group Lasso (SMuGLasso), an extended approach of MuGLasso. Our goal is to improve population-specific selection of LD-groups, by combining the $\ell_{1,2}$-norm penalty of MuGLasso with an additional ℓ_1-norm at the level of LD-groups. We compare risk genes findings of SMuGLasso to MuGLasso on simulated data and DRIVE breast cancer data. Moreover, we study the performance of the

algorithm for Arabidopsis thaliana in quantitative phenotype. Finally, we compare the stability of the selection of SMuGLasso, MuGLasso and other existing methods in identifying causal LD-groups/SNPs.

5.2 Methods

In this section, we present SMuGLasso which consists of four steps:

1. Alike MuGLasso, we assign each sample to a genetic population. Thus, each population is assigned to an input task in the multitask framework.

2. Identically, we form LD-groups of strongly correlated SNPs to perform feature selection at the group level using biological prior knowledge.

3. We fit a linear model with a regularization composed of two penalty terms: (1) a MuGLasso term that consists of an $\ell_{1,2}$-norm which enforces sparsity at the level of LD-group across all tasks/populations, and (2) an ℓ_1-norm which ensures sparsity at LD-groups level for specific populations.

4. We include the stability selection procedure to boost the robustness of our algorithm.

Unlike MuGLasso, this setting does not require additional steps after training to determine populations-specific LD-groups. SMuGLasso provides indeed a more precise and refined approach in the selection for population-specific causal markers.

5.2.1 Population structure

Diverse and admixed populations studies are a double-edged sword in GWAS. On the one hand, they offer a good solution to increase the number of samples unlike homogeneous studies where the number of samples is in most data very restricted. Indeed, genotyping hundreds of thousands of participants from different ancestries to study a phenotype of interest helps to alleviate the curse of dimensionality. On the other hand, such analyses require close attention to the confounder raised by population stratification, that is, when association is detected on the population structure rather than on the phenotype of interest.

We have presented in Chapter 2 a full review of existing population stratification adjustment methods. Our results have shown that these techniques can lead sometimes to overcorrection of some causal SNPs. Thus, we have observed a false negative rate appearing after correction in simulated data. Moreover, adjustment techniques do not consider the presence of some population-specific markers related to disease.

Consequently, there is a need in the GWAS field to provide efficient frameworks that profit from the number of samples advantage of diverse studies, while addressing the population stratification issue and considering the existence of population-specific causal LD-groups.

We follow the same procedure as for MuGLasso to identify the subpopulations that we infer as tasks. We use PCA with k-means clustering to define which sample belongs to which subpopulation. We have detailed the procedure in Section 4.2.1

5.2.2 Linkage disequilibrium groups clustering

We have shown in Chapter 3 and Chapter 4 in two different studies that the selection at LD-groups level has improved remarkably the stability (i.e., the variability to small changes in the input samples) compared to the selection of individual SNPs. Indeed, grouping SNPs together decreases the number of choices of selection for a regularization-based model. Thus, selection at the LD-groups level alleviates the curse of dimensionality in GWAS data.

We use adjacency-constrained hierarchical clustering algorithm to form the LD-groups assigned to SMuGLasso (see Section 4.2.2).

5.3 Sparse multitask group Lasso

5.3.1 Notations

In this chapter, we use the same notations as MuGLasso, given in Chapter 4 for the problem formulation.

Given a set of p SNPs measured on n samples, we split the n samples in T subpopulations/tasks, each of size n_t for $t = 1, \ldots, T$, and the p SNPs in G LD-groups, each of size p_g for $g = 1, \ldots, G$. For each population t, we denote by $\boldsymbol{x}_m^{(t)}$ the p-dimensional vector of SNPs of the m-th sample in the population ($m = 1, \ldots, n_t$), and by $y_m^{(t)}$ its phenotype.

5.3.2 Related work

We have presented in Section 4.2.3 several studies related to multitask variants composed of either two or three regularization terms. We have reported that these models do not scale to high-dimensional data, and therefore none of them has been applicable to our setting.

To efficiently choose the additional population-specific regularization term of SMuGLasso, we have investigated thoroughly the applicability of these methods. We have found that the proposed sparsity enforcing penalties are inappropriate to the problem we consider. Our goal is to implement a regularization term that enforces the sparsity for specific populations at the level of

the LD-group. In this section, we examine particularly the method proposed by [Li *et al.*(2020)] that have suggested implementing three regularization-based multitask models. Their optimization problem is reformulated by the following equation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p\times k}} \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{n_t}\left\|y_m^{(t)}-\sum_{j=1}^{p}\boldsymbol{\beta}_j^{(t)}x_{mj}^{(t)}\right\|_2^2 + \lambda_1\underbrace{\sum_{j=1}^{p}\|\boldsymbol{\beta}_j\|_2}_{\mathcal{R}_1(\boldsymbol{\beta})} + \lambda_2\underbrace{\sum_{t=1}^{T}\sum_{g=1}^{G}\sqrt{p_g}\left\|\boldsymbol{\beta}_g^{(t)}\right\|_2}_{\mathcal{R}_2(\boldsymbol{\beta})} + \lambda_3\underbrace{\sum_{t=1}^{T}\left\|\boldsymbol{\beta}^{(t)}\right\|_1}_{\mathcal{R}_3(\boldsymbol{\beta})}. \tag{5.1}$$

In this setting, [Li *et al.*(2020)] enforce population-specific groups sparsity using the term $\mathcal{R}_2(\boldsymbol{\beta})$ in order to select some groups only for some tasks/subpopulations. However, using only this regularization term, the optimization problem is separated over tasks. In other words, the selection is done separately for each single task. Thus, to ensure that tasks are fitted simultaneously, the authors add the term $\mathcal{R}_1(\boldsymbol{\beta})$ which corresponds to multitask regularization at the single-SNP level across the T tasks. Finally, they also aim to enforce sparsity within groups with an ℓ_1-norm over all SNPs, using a third regularization term (defined by $\mathcal{R}_3(\boldsymbol{\beta})$). It corresponds to the second regularization term of the sparse group lasso presented in Section 1.3.4.

We aim in this study to improve the selection for populations-specific LD-groups. We have tested the regularization terms proposed by [Li *et al.*(2020)]. First, adding $\mathcal{R}_2(\boldsymbol{\beta})$ to MuGLasso did not maintain the multitasking over the tasks T. Also, implementing $\mathcal{R}_3(\boldsymbol{\beta})$ combined with MuGLasso hinders the interpretation of the selected features. Thus, selecting SNPs within groups for specific populations make it hard to decide the number of SNPs within an LD-group g that must be 0 to consider the group as not selected for a specific task t. Finally, the addition of two penalties to MuGLasso increases dramatically the computational limitations at a GWAS scale.

Consequently, we present in Section 5.3.3 our solution to formulate the problem.

5.3.3 General framework and problem formulation

The optimization problem of SMuGLasso is written as follows:

$$\min_{B\in\mathbb{R}^{p\times T}} \frac{1}{n}\sum_{t=1}^{T}\sum_{m=1}^{n_t}\mathcal{L}\left(y_m^{(t)},\sum_{j=1}^{p}\boldsymbol{\beta}_j^{(t)}x_{mj}^{(t)}\right) + \lambda_1\sum_{g=1}^{G}\sqrt{p_g}\left\|B^{(g)}\right\|_F + \lambda_2\sum_{g=1}^{G}\sqrt{p_g}\left\|B^{(g)}\right\|_1, \tag{5.2}$$

where $\boldsymbol{\beta}_j^{(t)} \in \mathbb{R}$ is the regression coefficient of the j-th SNP for the task t, $\mathcal{L}$ is the quadratic loss if the phenotype is quantitative ($y \in \mathbb{R}$) and the logistic loss if it is qualitative ($y \in \{0, 1\}$), $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|_1$ denotes the ℓ_1-norm. $B^{(g)}$ is a $p_g \times T$ matrix containing the regression coefficients, across all tasks, for the SNPs of group g.

SMuGLasso can be reformulated through the creation of a new dataset $(\widetilde{X}, \widetilde{\boldsymbol{y}})$ where $\widetilde{X} \in \mathbb{R}^{n \times pT}$ is a block-diagonal matrix such that each of the T diagonal blocks corresponds to the SNP matrix $X^{(t)} \in \mathbb{R}^{n_t \times p}$ for a task t, and $\widetilde{\boldsymbol{y}}$ is a n-dimensional vector obtained by stacking the phenotype vectors for each task. The model can then be rewritten as:

$$\min_{\boldsymbol{b} \in \mathbb{R}^{pT}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\widetilde{y}_i, \sum_{k=1}^{pT} b_k \widetilde{x}_{ik}\right) + \lambda_1 \sum_{g=1}^{G} \sqrt{p_g} \left\|\boldsymbol{b}^{(g)}\right\|_2 + \lambda_2 \sum_{g=1}^{G} \sqrt{p_g} \left\|\boldsymbol{b}^{(g)}\right\|_1,$$

where $\boldsymbol{b}^{(g)} \in \mathbb{R}^{p_g T}$ is the vector of regression coefficients corresponding to all SNPs of group g for all tasks. The penalization parameters λ_1 and λ_2 control the strength of both regularization terms.

5.3.4 Gap safe screening rules

Gap safe screening rules [Ndiaye *et al.*(2017)] is a technique that offers a notable speed-up by discarding irrelevant features prior to starting a sparse optimizer. The method can be inserted to any iterative solver. Here, we solve the optimization problem formulated in Section 5.3.3 using the coordinate descent that is commonly used in feature selection models and can easily ignore useless coefficients. We have detailed the fundamental of these rules in Appendix D.2.3.

5.4 Experiments

5.4.1 Data

Simulated data Using GWAsimulator (presented in Section 1.5.4), we simulate GWAS data following LD patterns of two populations (CEU: Utah residents with Northern and Western European ancestry and YRI: Yoruba in Ibadan, Nigeria) from the HapMap 3 data (see Appendix A.1.2). We generate different numbers of samples through subpopulations to mimic the structure of real data where samples through subpopulations are not necessarily equally distributed. We also produce the population stratification confounder by varying the case:control ratio within each subpopulation (CEU 1 300:1 700 and YRI 400:600). We predefine a total of 200 disease SNPs as shown in Table 5.1, in which 50 SNPs (respectively 50 SNPs) are specific to the CEU (respectively YRI). We decide to locate the predefined disease loci randomly

Populations	Number of SNPs
Specific-CEU	50
Specific-YRI	50
Shared (CEU+YRI)	100
Total	200

Table 5.1: For simulated data, number of predefined causal SNPs

and without loss of generality through chromosome 12, 19, 21 and 22 (See Table 5.2). In total, the data is composed of 4 000 samples and 50 000 SNPs.

DRIVE Breast Cancer OncoArray The DRIVE OncoArray dataset contains 28 281 individuals that were genotyped for 582 620 SNPs. 13 846 samples are cases and 14 435 are controls. We have detailed the description of the data in Section 1.5.1. Additional information about data access and ethical approval are presented in Appendix D.1.2.

Chromosome	Subpopulations	
	CEU	YRI
12	4 000 - 4 050	4 000 - 4 050
19	1 000 - 1 050	1 000 - 1 050
21	$\emptyset$	10 000 - 10 050
22	1 000 - 1 050	$\emptyset$

Table 5.2: For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 12, 19, 21 and 22

Arabidopsis thaliana We perform a quantitative analysis using Arabidopsis thaliana dataset (presented in Section 1.5.3). We study DTF3 phentoype that corresponds to the flowering time per days. The dataset contains 923 samples and 6 973 565 SNPs.

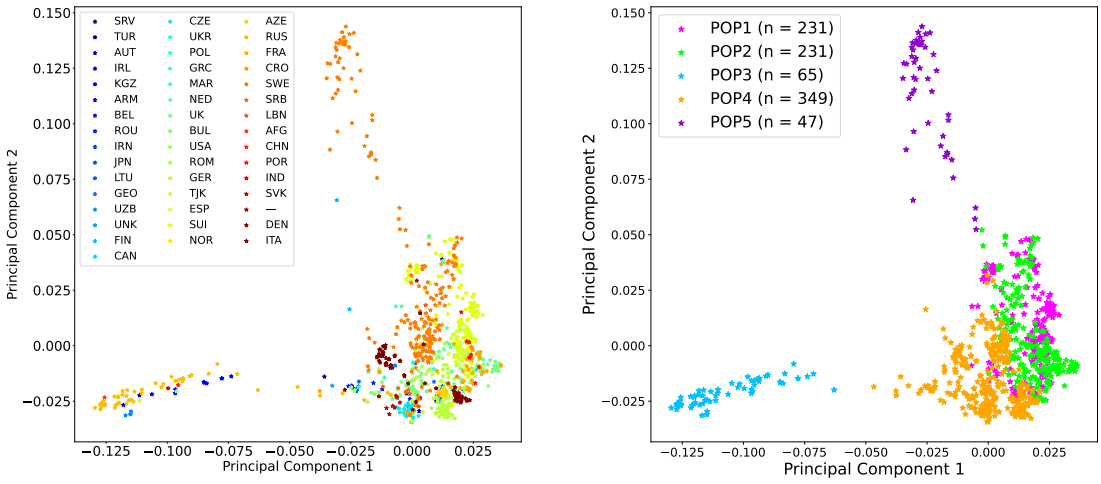
5.4.2 Preprocessing

Quality control and imputation For simulated dataset and DRIVE breast cancer, we follow the same quality control procedure as presented in Section 4.3.2. For Arabidopsis thaliana, we perform the quality control steps recommended by [Grimm *et al.*(2017)]. The phenotype was Box-Cox transformed [Box and Cox(1964)] to improve the measurements normality. We remove SNPs with a minor allele frequency lower than 5%.

LD pruning We perform LD pruning using PLINK [Purcell *et al.*(2007)] with an LD cutoff of $r^2 > 0.85$ and a sliding window of 50Mb for simulated data and DRIVE. For Arabidopsis thaliana, we use an LD cutoff of $r^2 > 0.75$ and a window size of 50Mb. After preprocessing steps, we obtain 50 000 SNPs in simulated data, 312 237 SNPs in DRIVE and 564 291 SNPs in Arabidopsis thaliana.

Population structure We use PLINK [Purcell *et al.*(2007)] to compute principal components of the genotype matrix. In the simulated data, we find two populations, corresponding to the CEU and YRI populations identically as simulated data used in Chapter 4 (see Figure D.3a). In DRIVE, we identify two populations (see Supplementary Figure D.3b) that we have called in Chapter 4 POP1 (samples from the USA, Australia and Denmark) and POP2 (samples from the USA, Cameroon, Nigeria and Uganda).

In the Arabidopsis thaliana dataset, from 46 samples countries, we retrieve 5 populations using k-means clustering of the top 4 principal components (see Figure 5.1 and Figure 5.2). In Appendix E.1, we examine population stratification adjustment methods following the study presented in Chapter 2. Here, the phenotype being continuous (DTF3), we perform linear regression for PCA-based models instead of logistic regression (when the phenotype is qualitative).



(a) PCA plot for 46 countries

(b) k-means clustering of PCs

Figure 5.1: PCA plots in Arabidopsis thaliana, we identify 5 subpopulations from 46 countries

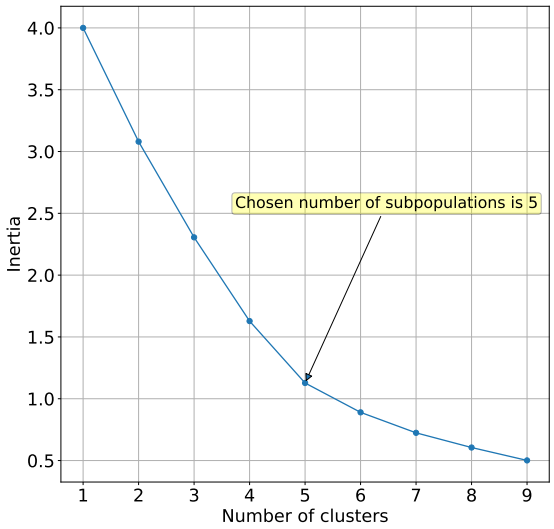


Figure 5.2: K-means clustering for Arabidopsis thaliana

Data	Subpopulations	# LD-groups	# shared LD-groups
Simulated data	CEU YRI	1 407 995	1 566
DRIVE real data	POP1 POP2	8 152 5 032	17 782
Athaliana data	POP1 POP2 POP3 POP4 POP5	1 846 1 950 2 002 1 728 1 834	7 080

Table 5.3: For each subpopulation of the studied datasets (simulated, DRIVE and Arapidopsis thaliana) LD-groups number is given and the shared LD-groups number after combination across subpopulations

LD-groups choice For simulated and DRIVE data, we determine the LD-groups for each subpopulation and each chromosome using `adjclust` [Ambroise *et al.*(2019)]. We thus obtain shared LD-groups across subpopulations as explained in Section 4.2.2. However, for Arabidopsis thaliana, adjclust did not scale computationally to the huge number of SNPs in the five chromosomes. Thus, we first split each chromosome to independent blocks of LD using `snpldsplit` [Privé(2021)] function from `bigsnpr R` package [Privé *et al.*(2018)]. We then form the LD-groups by applying adjclust on the obtained chunks of independent LD blocks. Table 5.3 shows the number of

LD-groups obtained for each subpopulation and the final number of shared groups.

5.4.3 Comparison patterns

We compare SMuGLasso with MuGLasso and the comparison patterns presented in Section 4.3.3 which are Adjusted GWAS, Stratified GWAS, Stratified Lasso, Adjusted Lasso, Stratified group Lasso and Adjusted group Lasso.

5.5 Results

5.5.1 SMuGLasso and MuGLasso rely on both LD-groups and the multitask approach to recover disease SNPs

On simulated data, we observe that SMuGLasso and MuGLasso outperform the other methods at recovering the predefined disease loci (See Figure 5.3). In addition, we confirm again that performing feature selection at the level of LD-groups provides better performance compared to the selection of single SNPs. Indeed, grouping SNPs helps to alleviate the curse of dimensionality and improve the identification of causal markers. Table 5.7 and Table 5.8

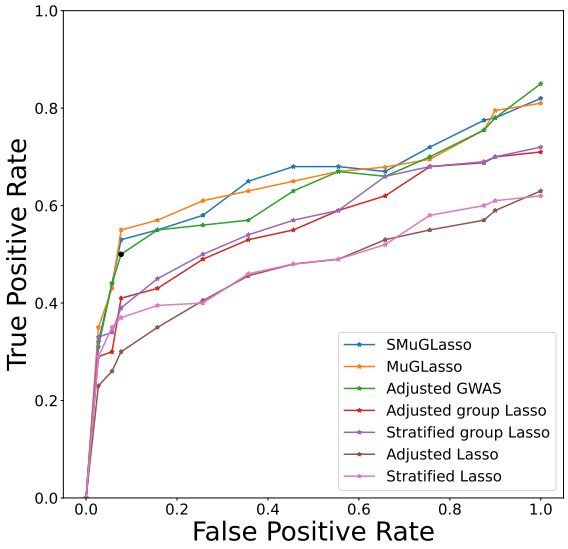
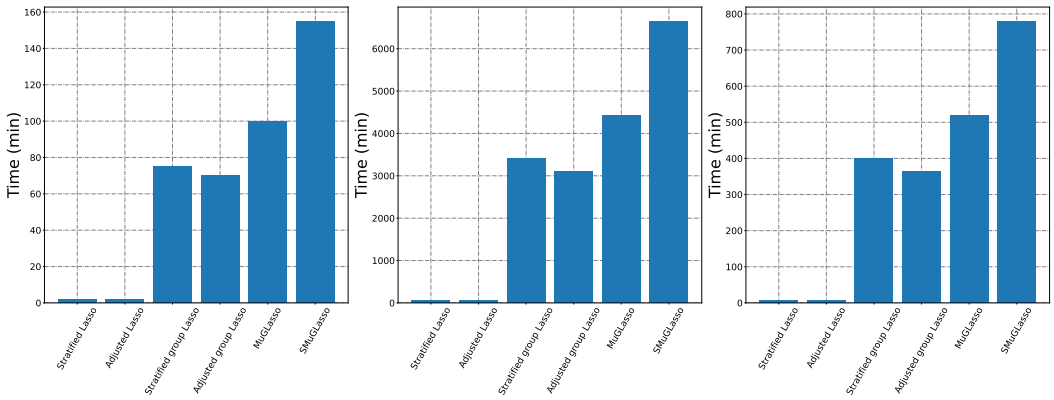


Figure 5.3: On simulated data, ability of different methods to retrieve causal disease SNPs as a ROC plot

detail respectively for SMuGLasso and MuGLasso, the number of selected LD-groups and SNPs across and per subpopulation for each dataset. Compared

to MuGLasso, we notice that SMuGLasso ensures more sparsity for shared selection across all subpopulations thanks to its additional ℓ_1-norm penalty. SMuGLasso provides a more precise selection for populations-specific level. Indeed, SMuGLasso recover successfully causal LD-groups/SNPs that MuGLasso have missed in simulated data.

On DRIVE, SMuGLasso recovers 1 279 SNPs including the 306 SNPs discovered by the adjusted GWAS. We detail in Table E.2 of Appendix E the breast cancer risk loci detected by SMuGLasso and MuGLasso on DRIVE. SMuGlasso successfully recover the 9 risk genes identified by classical GWAS. The model also identifies 18 new risk genes (also discovered by MuGLasso). From a total of 27 genes recovered by SMuGLasso, 17 have been identified in meta-GWAS data containing our samples. Also, 7 other genes have been proved to be associated with breast cancer risk. MuGLasso retrieves 5 additional risk genes that are not discovered by SMuGLasso, yet their association with breast cancer risk or growth was not proven in other studies. Also, we give in Table E.3 (Appendix E) the genes associated with DTF3 phenotype for Arabidopsis thaliana. We find that both SMuGLasso and MuGLasso recover the 7 genes selected by Adjusted GWAS. SMuGLasso recovers a total of 48 genes including 8 genes that are populations-specific findings. MuGLasso finds 7 additional genes that were not selected by SMuGLasso. MuGLasso recovers only 4 populations-specific genes from a total of its 55 discovered genes. We note that SMuGLasso is more intensive computationally compared



(a) Simulated data (b) DRIVE data (c) Arabidopsis thaliana data

Figure 5.4: Runtimes of Lasso approaches for simulated, DRIVE and Arabidopsis thaliana datasets

to MuGLasso and any other tested method (see Figure 5.4). This computational cost is caused by the additional populations-specific regularization term. However, the implementation is efficient enough to scale to high-dimensional GWAS data.

5.5.2 SMuGLasso and MuGLasso outperform the other methods in terms of stability

Methods	# selected LD-groups	# selected SNPs	Stability index	Selection level
SMuGLasso	8	290	0.5811	LD-groups
SMuGLasso without stab sel	9	328	0.5045	LD-groups
MuGLasso	10	363	0.7015	LD-groups
MuGLasso without stab sel	11	402	0.6124	LD-groups
Adjusted group Lasso + stab sel	11	374	0.5929	LD-groups
Adjusted group Lasso	12	392	0.5340	LD-groups
Stratified group Lasso	13	452	0.4491	LD-groups
Adjusted Lasso	12	422	0.4053	Single-SNP
Stratified Lasso	13	441	0.3140	Single-SNP
Adjusted GWAS	3	109	-	Single-SNP

Table 5.4: Stability index and number of selected features for different methods, on simulated data

Similarly to MuGLasso, we use 100 subsamples to perform stability selection [Meinshausen and Bühlmann(2009)]. Indeed, the obtained metrics in Tables 5.4, Table 5.5 and Table 5.6 show that stability selection increases the robustness of SMuGLasso, MuGLasso and Adjusted group Lasso for the three datasets. Also, we give the number of selected SNPs and LD-groups for each method. For methods providing the selection at single-SNP level, once a SNP is selected we consider that the entire LD-group is selected. MuGLasso remains the model that gives the best stability values on all datasets, followed by SMuGLasso that outperforms the other applied feature selection methods. Note that SMuGLasso produces less selected SNPs and LD-blocks compared to MuGLasso. Indeed, enforcing an additional penalty yields sparser model.

5.5.3 The selection of both task-specific and shared LD-groups

SMuGLasso ensures the selection of both shared (accross tasks) and task-specific LD-groups. As mentioned in Chapter 4, MuGLasso can also provide such a selection at the cost of a post-processing step, which consists in removing the groups with near-zero regression coefficients for a specific task. We present in Table 5.7 and Table 5.8 the number of both shared and populations-specific LD-groups (and SNPs) obtained respectively by SMuGLasso and Mu-

Methods	# selected LD-groups	# selected SNPs	Stability index	Selection level
SMuGLasso	58	1 279	0.3881	LD-groups
SMuGLasso without stab sel	60	1 354	0.3325	LD-groups
MuGLasso	62	1 357	0.4312	LD-groups
MuGLasso without stab sel	72	1 524	0.3911	LD-groups
Adjusted group Lasso + stab sel	59	1 293	0.3234	LD-groups
Adjusted group Lasso	68	1 466	0.2613	LD-groups
Stratified group Lasso	58	1 119	0.2498	LD-groups
Adjusted Lasso	41	874	0.2068	Single-SNP
Stratified Lasso	38	789	0.1581	Single-SNP
Adjusted GWAS	16	306	-	Single-SNP

Table 5.5: Stability index and number of selected features for different methods, on DRIVE

GLasso in simulated, DRIVE and Arabidopsis thaliana datasets. Feature selection in stratified models is determined separately for each task. Thus, the populations-specific LD-groups in stratified models correspond to LD-groups that were only selected in one population. However, the adjusted methods for population stratification (Adjusted group Lasso, Adjusted Lasso and Adjusted GWAS) do not allow the selection of population-specific LD-groups. As illustrated in Figure 5.5, SMuGLasso contributes to better recall performance for populations-specific SNPs. Thus, SMuGLasso reduces dramatically the number of falsely selected SNPs thanks to its additional ℓ_1-norm regularization.

5.6 Discussion and conclusion

We have presented in this chapter SMuGLasso, an extended approach of MuGLasso. The proposed model is based on a multitask framework in which the tasks are genetic populations and features are clustered in groups. The selection is performed at the scale of LD-groups. The populations are identified using PCA and k-means to assign each sample to a subpopulation. This setting alleviates the curse of dimensionality and addresses population stratification in diverse populations. SMuGLasso includes an additional regularization term compared to MuGLasso which penalizes the LD-groups for task-specific. Thus, our model provides indeed a more precise recovery of risk regions related to the phenotype at population-specific level.

We have shown in simulated data that SMuGLasso outperforms MuGLasso and the other implemented methods in retrieving population-specific true

Methods	# selected LD-groups	# selected SNPs	Stability index	Selection level
SMuGLasso	80	6 367	0.4315	LD-groups
SMuGLasso without stab sel	87	7 220	0.3883	LD-groups
MuGLasso	104	8 254	0.5733	LD-groups
MuGLasso without stab sel	149	10 935	0.5040	LD-groups
Adjusted group Lasso + stab sel	90	6 944	0.4489	LD-groups
Adjusted group Lasso	114	8 358	0.3654	LD-groups
Stratified group Lasso	133	10 135	0.3147	LD-groups
Adjusted Lasso	112	9 258	0.2600	Single-SNP
Stratified Lasso	135	9 897	0.2140	Single-SNP
Adjusted GWAS	7	31	-	Single-SNP

Table 5.6: Stability index and number of selected features for different methods, on Arabidopsis thaliana

Data	Population	# selected LD-groups (and SNPs)
Simulated data	CEU	2 (104 SNPs)
	YRI	3 (64 SNPs)
	shared (CEU and YRI)	3 (122 SNPs)
DRIVE	POP1	5 (155 SNPs)
	POP2	1 (21 SNPs)
	shared (POP1 and POP2)	52 (1 103 SNPs)
Athaliana	POP1	3 (247 SNPs)
	POP2	5 (381 SNPs)
	POP3	1 (81 SNPs)
	POP4	3 (232 SNPs)
	POP5	1 (72 SNPs)
	shared (5 populations)	67 (5 354 SNPs)

Table 5.7: For SMuGLasso, number of selected LD-groups/SNPs, across and per population

Data	Population	# selected LD-groups (and SNPs)
Simulated data	CEU	2 (88 SNPs)
	YRI	1 (14 SNPs)
	shared (CEU and YRI)	6 (261 SNPs)
DRIVE	POP1	6 (148 SNPs)
	POP2	2 (43 SNPs)
	shared (POP1 and POP2)	54 (1166 SNPs)
Athaliana	POP1	2 (164 SNPs)
	POP2	4 (303 SNPs)
	POP3	$\emptyset$
	POP4	3 (232 SNPs)
	POP5	$\emptyset$
	shared (5 populations)	95 (7 555 SNPs)

Table 5.8: For MuGLasso, number of selected LD-groups/SNPs, across and per population

disease loci. SMuGLasso reduces possible false discoveries that could occur in MuGLasso and the other feature selection methods. Although results demonstrate that MuGLasso is the most stable model, SMuGLasso gives also very close stability indexes in all tested datasets with the lower number of selected LD-groups/SNPs compared to MuGLasso and other regularization based models. Furthermore, stability selection technique has been proved to be efficient to improve the stability measurements of SMuGLasso.

In this study, thanks to gap safe screening rules we have addressed the computational complexity that occurs by including an additional regularizer to MuGLasso. Our model has been efficiently implemented in qualitative and quantitative phenotypes.

We have finally presented the discovered genes by SMuGLasso and MuGLasso in both studied real data for further biological interpretation. In the future, we aim to conduct pathway analysis to study the mechanisms underlying the studied phenotypes from the recovered risk genes.

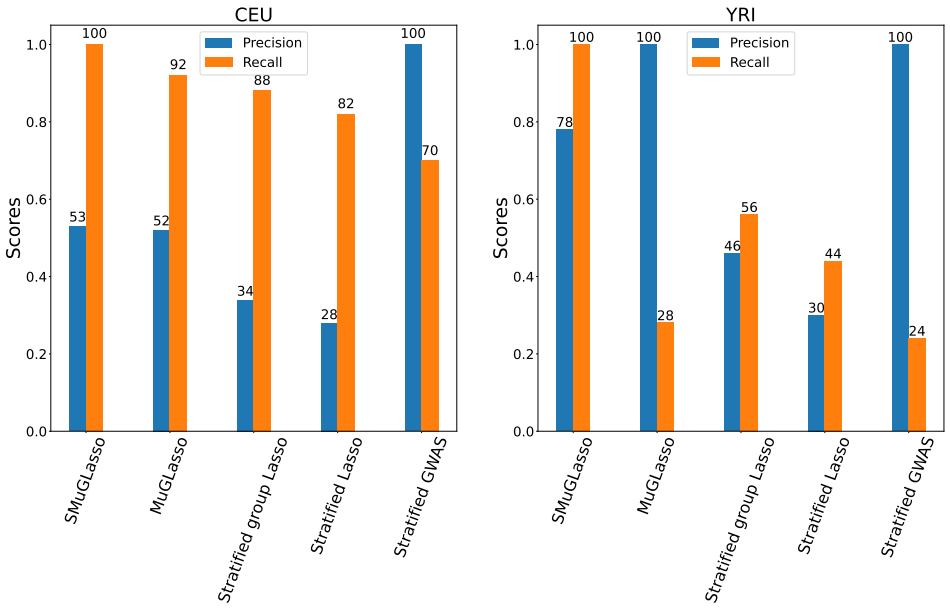


Figure 5.5: For simulated data, precision and recall of MuGLasso and the stratified approaches on the populations-specific SNPs

Chapter 6

Conclusions and perspectives

Contents

6.1 Introduction 104
6.2 Chapters summary 104
6.3 Future of GWAS 106
6.4 Final thoughts 109

6.1 Introduction

Since the accomplishment of the human genome project, the emergence of Genome-Wide Association Studies (GWAS) have provided valuable insights in explaining the influence of the genetic variation in disease development. Despite its success, there are many challenges that slow down the understanding of complex diseases mechanisms. GWAS suffer from the curse of dimensionality that leads to low statistical power, as well as, population stratification that produces ambiguous results. In addition, the strong correlation between the SNPs, due to LD, complicates the feature selection procedure. Another major issue is the lack of stability of regularization based methods, that is to say, robustness to slight variations in the input dataset. Thus, an unstable model leads to false biological interpretation due to wrong discoveries. Hence, the stability is an important indicator to trust feature selection discoveries. In this thesis, we aim to evaluate and improve the stability of the feature selection while addressing many limitations in GWAS. We develop novel machine learning models which deal with these challenges. The following sections provide in a first part a summary of the chapters presented in this dissertation. Then, a second part details speculation about the future of GWAS by raising some questions about the limitations in this research area, as well as, proposing some possible future directions to continue this work. Lastly, we conclude with some final thoughts.

6.2 Chapters summary

- In **Chapter 2** we have compared methods correcting for population stratification in case-control studies. To do so, we have used the genomic control technique, three PCA-based models and linear mixed models. We have conducted our analysis on simulated data and two breast cancer datasets. Also, we have studied empirically the performance of each method with data simulated using two different scenarios displaying either moderate or strong population stratification. Results have shown in most cases that linear mixed model FastLMM outperforms other methods followed with the PCA-based model EIGENSTRAT. However, these techniques do not consider the existence of populations-specific causal variants. Indeed, these methods provide a uniform correction for population stratification for all subpopulations. This study led us to propose novel frameworks that address population stratification efficiently and make it possible to discover populations-specific variants.

- In **Chapter 3** we have evaluated empirically the stability of the feature selection of several GWAS methods. The stability has been evaluated at different genomic scales (SNP level, LD-blocks level and gene level).

To do so, we have tested three different feature selection frameworks: the classical univariate statistical test, Lasso, Elastic Net. Moreover, we have examined stability selection techniques based on subsampling on random samples subsets. Our study shows that stability selection improves remarkably the robustness of any tested method. We have conducted our analysis using WTCCC1 data for three different diseases. Furthermore, we have shown that the stability of both feature selection models (i.e., Lasso and Elastic Net) increases remarkably at the LD-blocks and the gene level compared to the SNP level. Although we have found that classical GWAS technique based on single-marker analyses outperforms feature selection methods, in terms of stability, this method suffer from low statistical power in retrieving causal SNPs and thus misses many meaningful associations.

To our knowledge, this is the first study that quantifies the stability of feature selection models at different genomic scales. Finally, as we have found that grouping SNPs in strong LD produces better stability, we have decided to account for LD-blocks (or groups) in the methods we have developed in the following chapters.

- In **Chapter 4** we have presented the multitask group Lasso (MuGLasso), a novel feature selection model in diverse populations. In our approach, the tasks correspond to genetic populations and the groups correspond to groups of SNPs in LD. We have used PCA and k-means clustering to identify which sample belongs to which task. The LD-groups have been determined using adjacency-constrained hierarchical agglomerative clustering. The model relies on an $\ell_{1,2}$-norm regularization term. Despite its complex architecture, MuGLasso is efficient enough to scale high-dimensional GWAS data thanks to gap safe screening rules. We have also incorporated the stability selection procedure to improve the robustness of the model. Hence, our performance metrics show that MuGLasso outperforms any other tested technique in case-control simulated and breast cancer data. Furthermore, MuGLasso has been able to identify both shared and task-specific causal SNPs. The task-specific discoveries have been identified using further post-processing step (when some LD-groups yield values close-to-zero in only specific population). Our model efficiently addresses the population stratification issue thanks to multitasking. It has also reduced the curse of dimensionality severity thanks to SNPs grouping. MuGLasso has alleviated the computational complexity thanks to gape safe screening rules, as well as the lack of robustness thanks to stability selection. Finally, we have presented MuGLasso risk gene discoveries that are related to breast cancer risk or its tumor growth. Most of our findings were also identified either in other studies or in meta-data including our

samples.

- In **Chapter 5** we have presented an extended approach of MuGLasso, the sparse multitask group Lasso (SMuGLasso). The model includes an additional ℓ_1-norm regularization that penalizes populations-specific LD-groups. Our goal was to provide a more precise method to select task-specific LD-groups. Although MuGLasso provides the possibility to obtain populations-specific genes. The selection of populations-specific LD-groups is identified as those with near-zero regression coefficients for one task. However, the choice of the threshold of coefficients to consider an LD-group as selected for a particular population remains critical. We have shown in this chapter that SMuGLasso retrieves successfully causal LD-groups (and SNPs) better than MuGLasso and any other tested method. Furthermore, SMuGLasso reduces considerably false discoveries in simulated data. Indeed, adding an additional ℓ_1-norm regularizer results in a sparser model. On the DRIVE dataset, SMuGLasso has identified most of the genes that were found in MuGlasso. Yet, SMuGLasso has discarded some genes that were retrieved by MuGLasso. Interestingly, we did not find any study linking the genes discarded by SMuGLasso with breast cancer in the literature, indicating that those genes are most likely false discoveries. Thus, the model identify fewer LD-groups but still offer stability performances very close to MuGLasso stability performances. Finally, while SMuGLasso is more computationally intensive than MuGLasso, it has been efficiently implemented for GWAS data.

6.3 Future of GWAS

What about the predictive power of GWAS findings in clinical applications?

The predictive power of GWAS is limited in clinical application because of the low proportion of heritability explained and the small number of participants. Indeed, many studies have shown that the identified SNPs produce low performance in discriminating samples according to a phenotype of interest in most complex diseases [Janssens and van Duijn(2008), Loos and Janssens(2017)].

The goal of using GWAS discoveries in clinical applications in order to prevent and treat diseases remains very challenging at the moment. Even in diseases for which genetic variation is known to explain most of the heritability, prediction of the disease status is not successfully achievable because of the small number of participants. Also, false positive findings produced by GWAS and feature selection methods are another reason of the poor predictive power.

In this thesis, we have mainly worked on improving the robustness of feature selection models to reduce false positive findings and to boost the stability of the selection.

What about the detection of the epistasis in humans?

Epistasis is defined as the interaction between genetic loci. In that case, the effect of one locus on the phenotype depends on one or more other loci.

In this thesis, the influence of the interaction between SNPs were not studied. First, including epistasis would intensify the curse of dimensionality and would decrease the statistical power to identify causal SNPs. In addition, both developed models (MuGLasso and SMuGLasso) would not support computationally the huge number of features that would need to be added to the model if we were to include epistasis.

An important point of discussion to raise here is, does GWAS have been really successful in detecting epistasis in human?

In fact, there is limited evidence in the literature showing that epistasis explains a large percentage of complex disease heritability. It has been constantly reported that epistasis contributes to the missing heritability in complex disease [Okser *et al.*(2013), Wei *et al.*(2014), Cortes *et al.*(2015)]. Indeed, non-linear SNPs interactions have hardly identified genetic variants with significant effect on the phenotype. Furthermore, it remains challenging to identify gene-gene interactions by using GWAS and post-GWAS methods in humans because of the lack of statistical power. Also, modeling epistasis interaction is a very complex task, it remains unclear how to perform several common GWAS analysis such as adjusting for population stratification in diverse studies or accounting for LD.

In any case, very large number of samples and wider computing resources are needed to detect significant gene-gene interactions.

To summarize, the exploration of epistasis is a very motivational research problem that future work in GWAS will hopefully resolve, but it needs special attention. The usage of machine learning models may boost the identification of epistasis interactions in humans. Also, few algorithms have been proposed recently, integrating deep learning based methods for epistasis detection [Li *et al.*(2018), Wang *et al.*(2019), Fergus *et al.*(2020)].

What about post-selection inference in GWAS?

Feature selection models do not rank the selected SNPs by statistical significance with p-values. Thus, using all SNPs mapped to a gene, including those with low association power, can mask the association signal of a gene. To address the lack of interpretability in regularization based methods, post-selection inference [Zhang and Zhang(2014)] plays a pivotal role to produce valid p-values of identified features with Lasso methods. [Slim *et al.*(2022)]

have proposed a post-selection inference method for GWAS application using kernels to model epistasis interactions between nearby SNPs. Indeed, the method has been successfully implemented in quantitative BMI phenotype. However, applying these methods for qualitative phenotypes is not feasible because these tasks are computationally very intensive. In the future, integrating post-inference selection in multitask frameworks (such as MuGLasso and SMuGLasso proposed in this thesis) is a very interesting research direction of major interest. For now, it remains difficult to implement this framework in a computationally efficient manner especially for case-control studies.

What about biological networks exploration in GWAS?

Biological networks model a complete representation of the interactions between appropriate biological elements in a graph, where the nodes correspond to SNPs or genes and the edges correspond to association of SNPs or genes with the phenotype of interest. Biological networks were proved successful in explaining complex disease mechanisms [Climente-González *et al.*(2021)]. Indeed, adding prior knowledge about biological mechanisms in feature selection methods provides a realistic design of the problem. In this setting, feature selection models provide connectivity and association constraints, in addition to regularization terms in order to design biological interactions [Azencott *et al.*(2013)].

Another attractive direction to continue this work is to incorporate biological networks regularizers in multitask models such as MuGLasso and SMuGLasso.

What about the application of deep learning approaches in GWAS?

Deep learning is a technique widely applied in many fields areas including bio-informatics applications, such as bioimage analysis and computer vision of cellular and tissular phenotypes. One can assume that deep learning could facilitate inclusion of nonlinear transformations in GWAS by extracting relevant features from high-dimensional data. This would be an interesting addition to traditional machine learning models that are based on regularization terms that predict a linear combination of weights by assuming a linear relationship between variants and a phenotype of interest. However, the high-dimensional characteristic of GWAS data makes the task of exploring models like neural networks very complex. The major weakness of deep learning application to GWAS is the lack of interpretability in underlying genetic effects from SNPs, as well as, the computational issues. Few nonlinear models have been recently proposed by integrating multi-omics data along with GWAS data [Xu *et al.*(2021), Fang *et al.*(2022)]. Indeed, such analysis can boost the biological interpretation that GWAS data can not handle alone in nonlinear models. Thus, coupling different genetic data helps to identify high-confidence risk

genes. For now, it remains unclear how to build an efficient framework for identifying causal variants associated with diseases, or predicting phenotypes from genetic variants using only GWAS data.

6.4 Final thoughts

Despite the limitations mentioned above in GWAS field to fully explain the genetic background of complex diseases, many motivational studies have shown the efficiency of GWAS to discover genes associated to diseases (see Figure 6.1). Indeed, over the 15 past years, many contributions have been proposed making many advances in the field. In this thesis, we have proposed novel methods that outperform existing models in terms of stability of the selection and in identifying risk genes associated with the studied phenotypes. We have addressed many GWAS issues such as linkage disequilibrium, population stratification, lack of statistical power, curse of dimensionality and computational limitations. As explained before, many interesting future directions are possible to further improve the methods developed in this thesis. In the future, we hope that the developments of post-inference and biological networks methods will contribute to improve the confidence in feature selection models discoveries.

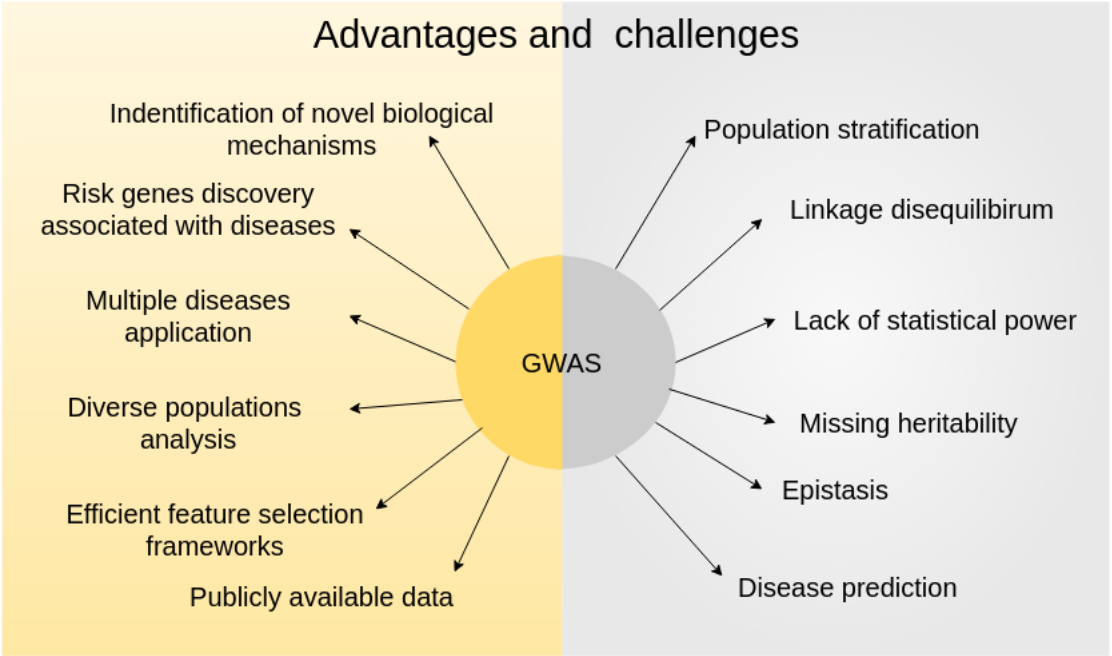


Figure 6.1: GWAS advantages and challenges

Appendix A

GWAS data

A.1 GWAS data

A.1.1 1000 Genome Project

In 2008, the International Genome Sample Resource has created 1000 Genome Project [Consortium(2015)] to develop a catalogue of common human genetic variation, using control samples from people who declared to be healthy. The datasets are publicly available, many released were provided. Mainly, the 1000 Genomes phase 3 release offer genotype data for 26 populations divided into 5 main populations as shown in Table A.1.

A.1.2 International HapMap Project

The International HapMap Consortium has launched the International HapMap Project in 2001 to develop HapMap Project of the human genome [Consortium(2003)]. It describes the common patterns of human DNA sequence variation. All HapMap data are freely available to the public through the database dbSNP. Three releases were produced: Phase I in 2005, Phase II in 2007 and Phase 3 in 2010.

Genotype samples were provided from the following 11 populations: **ASW** African ancestry in Southwest (USA), **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese in Beijing (China), **CHD** Chinese in Metropolitan Denver (Colorado), **GIH** Gujarati Indians in Houston (Texas), **JPT** Japanese in Tokyo (Japan), **LWK** Luhya in Webuye (Kenya), **MXL** Mexican ancestry in Los Angeles (California), **MKK** Maasai in Kinyawa (Kenya), **TSI** Toscani in Italia and **YRI** Yoruba in Ibadan (Nigeria).

Table A.2 gives the number of samples for each population that were genotyped in release 3 from samples that existed already in release I and II, as well as the number of SNPs that passed quality control.

HapMap is one of the important tools in GWAS for researchers to conduct multiple studies to find genes that affect health, disease, and response to drugs and environmental factors [Altshuler *et al.*(2010)]. However, no phenotype information is available for the HapMap samples, all samples are considered as controls. The tool was efficiently employed for phenotype simulation, it

Population ID	Population name	Superpopulation code	Superpopulation name	Superpopulation display order	Number of samples
CHS	Southern Han Chinese	EAS	East Asian Ancestry	3	171
FIN	Finnish	EUR	European Ancestry	4	105
KHV	Kinh Vietnamese	EAS	East Asian Ancestry	3	124
ACB	African Caribbean	AFR	African Ancestry	1	123
PUR	Puerto Rican	AMR	American Ancestry	2	150
BEB	Bengali	SAS	South Asian Ancestry	5	144
ASW	African Ancestry SW	AFR	African Ancestry	1	113
CHB	Han Chinese	EAS	East Asian Ancestry	3	112
GWD	Gambian Mandinka	AFR	African Ancestry	1	280
MSL	Mende	AFR	African Ancestry	1	129
YRI	Yoruba	AFR	African Ancestry	1	187
ESN	Esan	AFR	African Ancestry	1	173
LWK	Luhya	AFR	African Ancestry	1	117
IBS	Iberian	EUR	European Ancestry	4	162
JPT	Japanese	EAS	East Asian Ancestry	3	105
MXL	Mexican Ancestry	AMR	American Ancestry	2	107
CDX	Dai Chinese	EAS	East Asian Ancestry	3	109
CLM	Colombian	AMR	American Ancestry	2	148
TSI	Toscani	EUR	European Ancestry	4	112
PEL	Peruvian	AMR	American Ancestry	2	130
PJL	Punjabi	SAS	South Asian Ancestry	5	158
CEU	CEPH	EUR	European Ancestry	4	184
GIH	Gujarati	SAS	South Asian Ancestry	5	113
STU	Tamil	SAS	South Asian Ancestry	5	128
ITU	Telugu	SAS	South Asian Ancestry	5	118
GBR	British	EUR	European Ancestry	4	107

Table A.1: Population samples classification in the 1000 Genome Project data

Population	Number of individuals with HapMap 3 (Number of individuals total)	Number of SNPs (after QC)
ASW	71 (of 90)	1 632 186
CEU	162 (of 180)	1 634 020
CHB	82 (of 92)	1 637 672
CHD	70 (of 90)	1 619 203
GIH	83 (of 90)	1 631 060
JPT	82 (of 89)	1 637 610
LWK	83 (of 90)	1 631 688
MXL	71 (of 90)	1 614 892
MKK	171 (of 180)	1 621 427
TSI	77 (of 90)	1 629 957
YRI	163 (of 180)	1 634 666
Total	1 115 (of 1 261)	1 525 445

Table A.2: Population samples and SNPs for genotyping in HapMap 3 release

was used to generate realistic simulated data by following the LD patterns of populations provided in the study.

Appendix B

PLINK files format

- **PED file:** Samples data: each row corresponds to a participant. The first six columns refer to:
  - Family ID
  - Sample ID
  - Paternal ID
  - Maternal ID
  - Sex (1=male; 2=female; other=unknown)
  - Affection (1=control; 2=case; -9 or 0=missing)

  The following columns contain bi-allelic SNPs information. So, each SNP is presented by two columns. To summarize, the number of columns is:
  $6 + 2 \times number\ of\ SNPs$
- **FAM file:** This file contains the first six fields in a PED file presented above.
- **BED file:** Binary file contains the genotype information.
- **MAP file:** Markers data: each row represents a SNP. The fields in a MAP file are:
  - Chromosome
  - Marker ID
  - Genetic distance
  - Physical position
- **BIM file:** Similar to MAP file with two extras columns of allele names.

Appendix C

Lasso and Elastic Net stability evaluation for the phenotypes T1D and T2D

C.1 State-of-the-art stability of the selection methods

We present in Table C.1 the properties of stability measurements proposed in the literature. For each of the 17 index, and for each of the 5 properties, we show which measurement satisfies which property.

Reference	Name	Fully defined	Monotonicity	Bounds	Maximum	Correction
[Dunne *et al.*(2002)]	Hamming	X	X	X	X	
[Kalousis *et al.*(2005)]	Jaccard	X	X	X	X	
[Yu *et al.*(2008b)]	Dice	X	X	X	X	
[Zucknick *et al.*(2008)]	Ochiai	X	X	X	X	
[Consortium(2006)]	POG	X	X	X	X	
[Kuncheva(2008)]	Kuncheva		X	X	X	X
[Lustgarten *et al.*(2009)]	Lustgarten	X	X	X		X
[Wald *et al.*(2013)]	Wald	X	X			X
[Zhang *et al.*(2009)]	nPOG	X	X		X	X
[Goh and Wong(2016)]	Goh	X		X		
[Davis *et al.*(2006)]	Davis	X		X		
[Krizek *et al.*(2007)]	Krizek				X	
[Guzman-Martinez *et al.*(2011)]	Guzman			X	X	X
[Somol and Novovicova(2010)]	CW_{rel}	X	X	X		
[Lausser *et al.*(2013)]	Lausser		X	X	X	
[Nogueira and Brown(2015)]	Pearson	X	X	X	X	X
[Nogueira *et al.*(2018)]	Nogueira	X	X	X	X	X

Table C.1: Satisfied properties for each stability measurement [Nogueira *et al.*(2018)]

C.2 Results of Lasso for T1D phenotype

For T1D disease, Figure C.3 shows the stability index values against the average error values obtained by Lasso for different values of lambda. We choose $\lambda = 0.017$ as the best parameter that gives the trade-off compromised between the stability and the error average. For $\lambda = 0.017$, we obtain the following stability indexes: 0.272 at the SNP level, 0.412 at the LD-block level and 0.243 at the gene level. For the same value of λ, Figure C.1 highlights the number of selected features at different genomic scales: 367 SNPs, 258 LD-blocks and 60 genes. Finally, Figure C.2 gives the obtained values of stability and error across different values of λ.

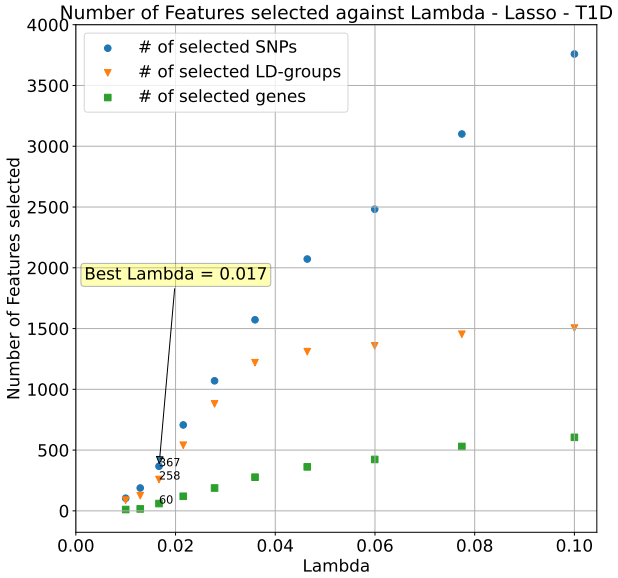
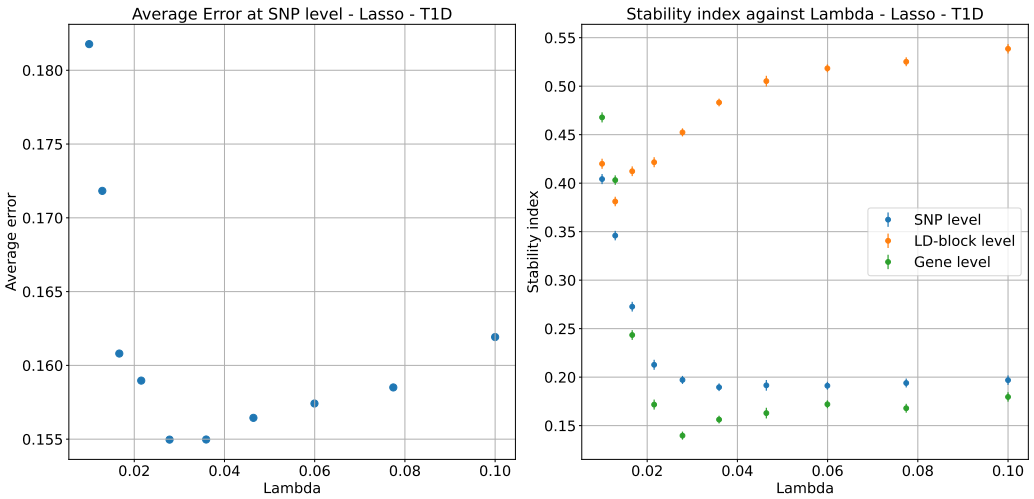


Figure C.1: For Lasso: number of selected SNPs, LD-blocks and genes against lambdas in T1D phenotype



(a) The average error against lambdas.

(b) The stability index at different genomic scales (SNP, LD-block and gene) against lambdas.

Figure C.2: For Lasso, the average error and stability index for different values of lambdas in T1D phenotype

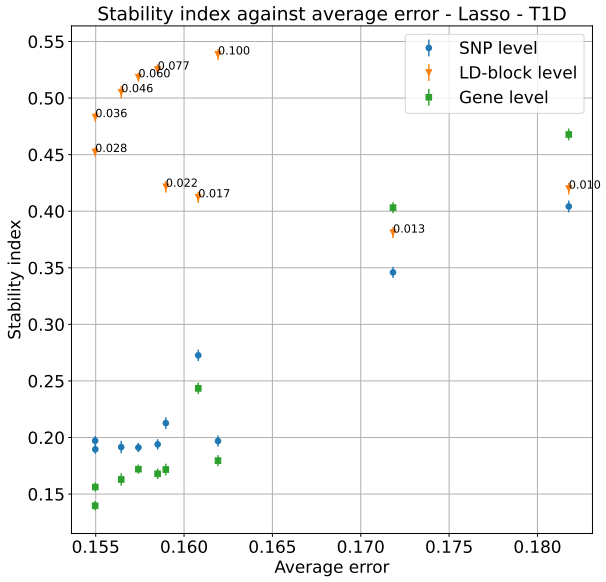


Figure C.3: For Lasso, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T1D phenotype

C.3 Results of Lasso for T2D phenotype

Figure C.5 shows the stability values obtained at the different genomic scales against the average error for different values of the penalization parameter λ of

Lasso studied in T2D phenotype. We choose $\lambda = 0.059$ as the best parameter compromising the stability and the error, it gives an error of 0.237 and stability values of 0.197 at the SNP level, of 0.552 at the LD-block level and of 0.213 at the gene level. Figure C.4 highlights that for chosen $\lambda = 0.059$, we obtain 3 099 SNPs, 1335 LD-blocks and 520 genes. Finally, Figure C.5 illustrates the measurements of the stability and the error across different values of λ.

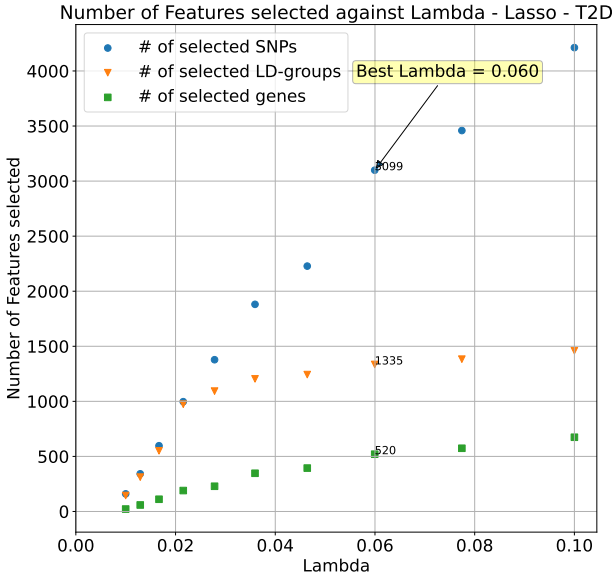
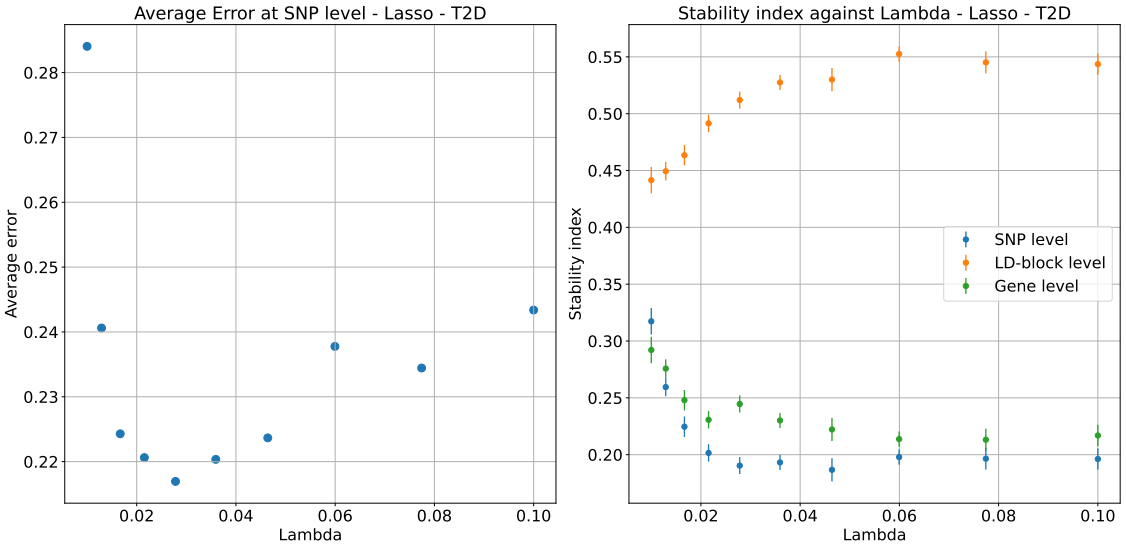


Figure C.4: For Lasso: number of selected SNPs, LD-blocks and genes against lambdas in T2D phenotype



(a) The average error against lambdas.

(b) The stability index at different genomic scales (SNP, LD-block and gene) against lambdas.

Figure C.5: For Lasso, the average error and stability index for different values of lambdas in T2D phenotype

C.4 Results of Elastic Net for T1D phenotype

According to Figure C.8, we choose here the best lambda to be $\lambda = 0.077$ that gives an error of 0.03 and stability indexes of 0.226 at the SNP level, of 0.557 at the LD-block level and of 0.27 at the gene level. Figure C.6 shows that for $\lambda = 0.077$ we obtain 2 140 SNPs, 987 LD-blocks and 398 genes. Also, Figure C.7 presents all the stability values obtained at different levels for different values of λ.

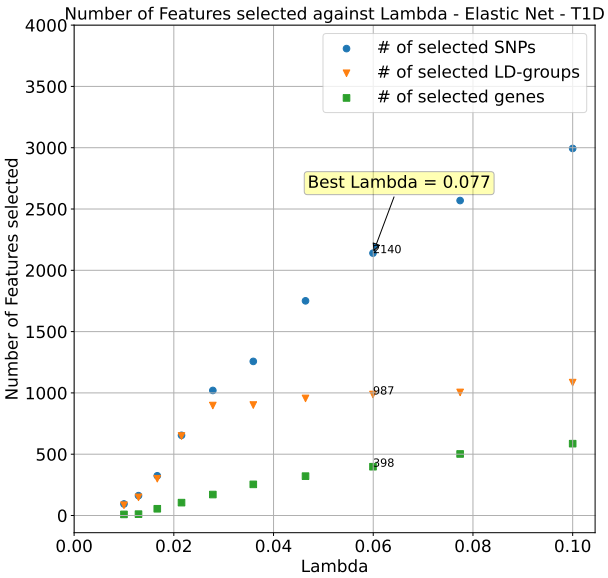


Figure C.6: For Elastic Net, number of selected SNPs, LD-blocks and genes against lambdas in T1D phenotype

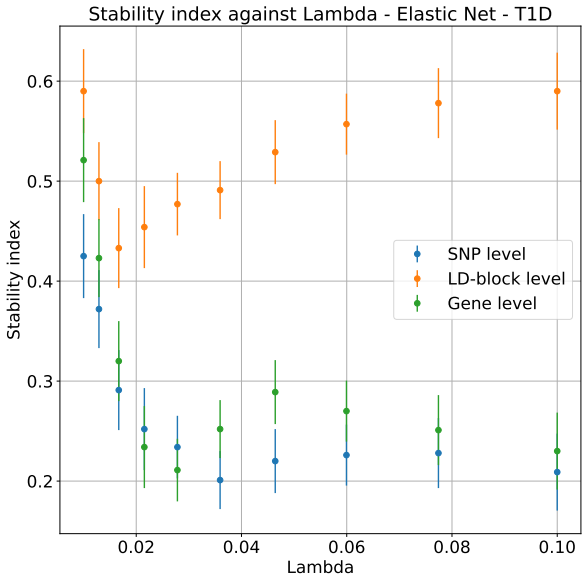


Figure C.7: For Elastic Net, the stability index is given for different values of lambda in T1D phenotype

C.5 Results of Elastic Net for T2D phenotype

We observe that all the stability indexes increase for T2D phenotype using Elastic Net compared to Lasso. Figure C.11 shows that the best stability values ensuring a low prediction error is obtained with $\lambda = 0.022$, it produces

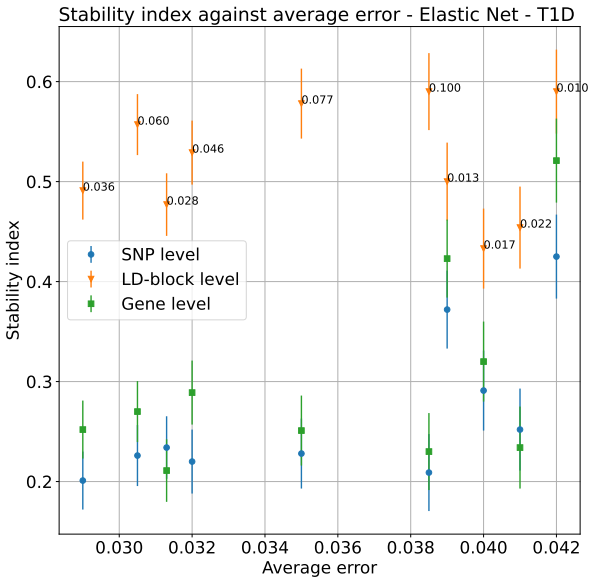


Figure C.8: For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T1D phenotype

an error of 0.006 and stability indexes of 0.29 at the SNP scale, of 0.566 at the LD-block scale and 0.384 at the gene scale. As shown in Figure C.10, for $\lambda = 0.022$, we obtain 941 selected SNPs, 824 selected LD-blocks and 178 selected genes. Figure C.9 presents the obtained stability values across the tested values of lambda.

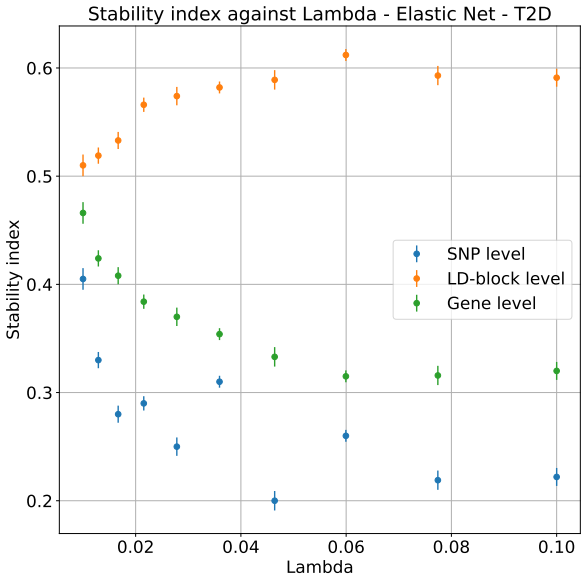


Figure C.9: For Elastic Net, the stability index is given for different values of lambda in T2D phenotype

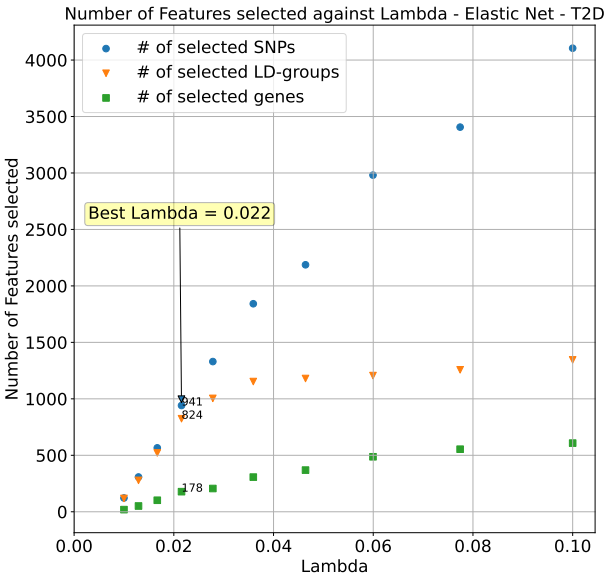


Figure C.10: For Elastic Net, number of selected SNPs, LD-blocks and genes against lambdas in T2D phenotype

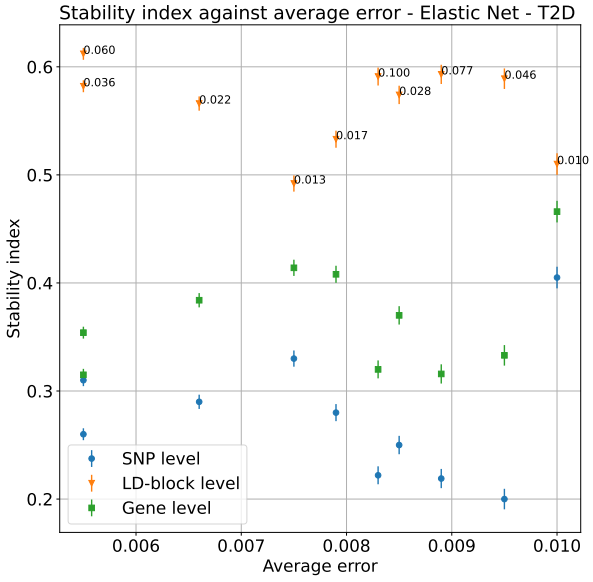


Figure C.11: For Elastic Net, the stability index at different genomic scales (SNP, LD-block and gene) against the average error in T2D phenotype

Appendix D

Multitask group lasso (MuGLasso) supplementary materials

D.1 Data availability

D.1.1 Simulated data

Code to reproduce our simulations is available on https://github.com/asmanouira/MuGLasso_GWAS

Table D.1 shows the location of the predefined disease loci, for each population. Table D.2 shows the number of predefined disease loci, both common to both population and specific to each population.

Chromosome	Subpopulations	
	CEU	YRI
2	1 000 - 50 000	1 000 - 50 000
12	10 - 37 000	10 - 40 000
19	1 000 - 50 000	1 000 - 50 000
21	10 - 10 000	10 - 7 000
22	-	10 - 2 000

Table D.1: For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 2, 12, 19, 21 and 22

Populations	Number of SNPs
Specific-CEU	2 999
Specific-YRI	4 989
Shared (CEU+YRI)	141 982
Total	149 970

Table D.2: For simulated data, number of predefined causal SNPs

D.1.2 DRIVE

Data access The dataset "General Research Use" in DRIVE Breast Cancer OncoArray Genotypes is available from the dbGaP controlled-access portal, under Study Accession phs001265.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1). Researchers can gain access the data by applying to the data access committee, see https://dbgap.ncbi.nlm.nih.gov.

Ethics approval The dataset was obtained from NIH after ethical review of project #17707, titled "Network-guided multi-locus biomarker discovery", and used under approval of this request (#67806-4).

Acknowledgments OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and theNational Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls: the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREOGAN), Copenhagen GeneralPopulation Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation intoCancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), Nashville-Breast Health Study (NBHS), Nurses Health Study (NHS), Nurses Health Study 2 (NHS2), Polish Breast CancerStudy (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiologyand Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort (SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women's Health Initiative (WHI).

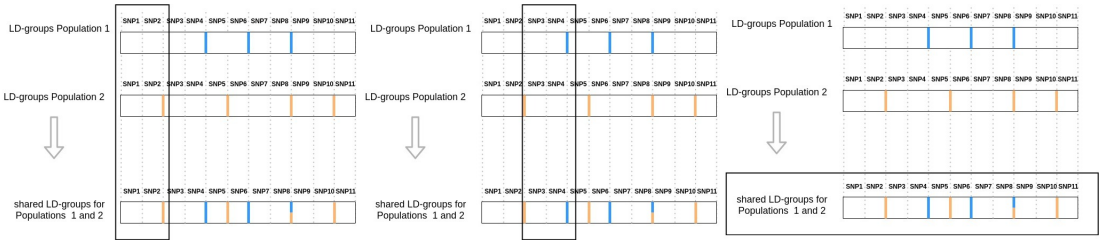


Figure D.1: Choice of shared LD-groups choice after adjacency-constrained hierarchical clustering for each population

D.2 Supplementary Methods

D.2.1 LD groups across populations

Figure D.1 illustrates the process by which we obtain LD-groups across populations, from LD-groups obtained on each population separately using adjacency-constrained hierarchical clustering (see Section 4.2.2)

D.2.2 Multitask group lasso

Figure D.2 illustrates the architecture of the multitask group Lasso described in Section 4.2.3.

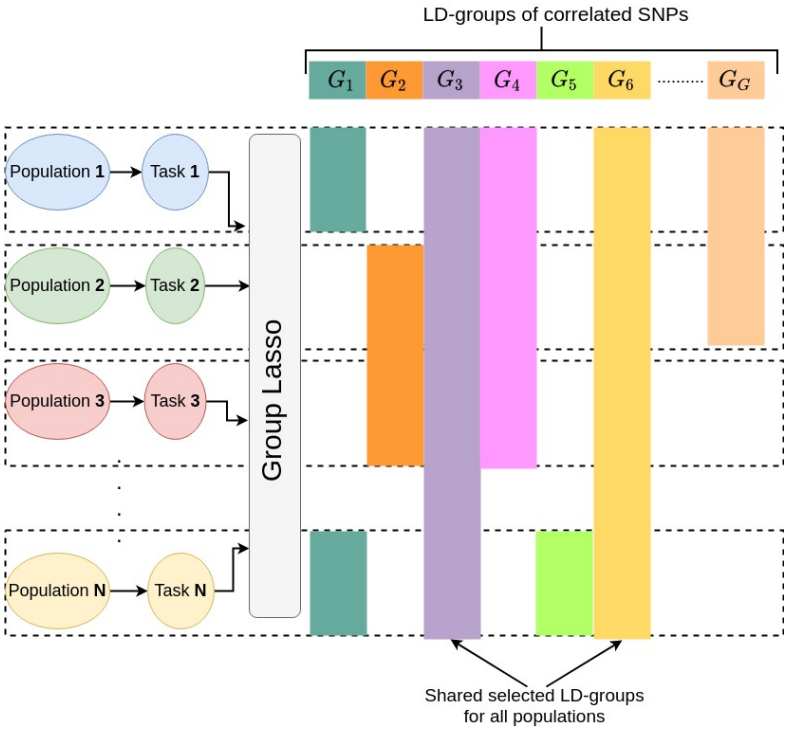


Figure D.2: Multitask group Lasso architecture

D.2.3 Gap safe screening rules

Let $X \in \mathbb{R}^{n \times d}$ be a design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ the corresponding vector of outcomes, which can be binary or real-valued. We consider the following optimization problem:

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} P_\lambda(\boldsymbol{\beta}) := \sum_{i=1}^{n} f_i\left(X_{i.}^\top \boldsymbol{\beta}\right) + \lambda\, \Omega(\boldsymbol{\beta}), \tag{D.1}$$

where all $f_i : \mathbb{R} \to \mathbb{R}$ are convex and differentiable functions with $1/\gamma-$Lipschitz gradient, and $\Omega : \mathbb{R}^d \to \mathbb{R}_+$ is a norm that is group-decomposable, i.e., the set of d features is partitioned in G groups of sizes $d_1, d_2, \ldots, d_G$, and

$$\Omega(\boldsymbol{\beta}) = \sum_{g=1}^{G} \Omega_g\left(\boldsymbol{\beta}^{(g)}\right),$$

where each Ω_g is a norm on $\mathbb{R}^{d_g}$ and, as previously, $\boldsymbol{\beta}^{(g)}$ corresponds to the coefficients of $\boldsymbol{\beta}$ restricted to the features in group g. As before, the λ parameter is a non-negative constant controlling the trade-off between the data fitting term and the regularization term.

Equation (4.2) is a special case of Equation (D.1) because the squared loss and the logistic loss are convex and differentiable.

Safe screening rules make it possible to solve such problems more efficiently by discarding features whose coefficients are guaranteed to be zero at the optimum, prior to using a solver. They usual rely on the dual formulation of Equation (D.1):

$$\widehat{\boldsymbol{\theta}}^{(\lambda)} = \arg\max_{\boldsymbol{\theta} \in \Delta_X} D_\lambda(\boldsymbol{\theta}) := -\sum_{i=1}^{n} f_i^*(-\lambda\theta_i), \tag{D.2}$$

where $f_i^* : \mathbb{R} \to \mathbb{R}$ is the Fenchel-Legendre transform of f_i, defined by $f_i^*(u) = \sup_{z \in \mathbb{R}} \langle z, u\rangle - f_i(z)$ and $\Delta_X \subset \mathbb{R}^n$ is defined by $\Delta_X = \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \forall g = 1, \ldots, G, \Omega_g^D(X^{(g)\top}\boldsymbol{\theta}) \leq 1\right\}$, where $\Omega_g^D : \mathbb{R}^{p_g} \to \mathbb{R}$ is the conjugate norm of Ω_g, defined by $\Omega_g^D(\boldsymbol{u}) = \max_{\boldsymbol{z} in \mathbb{R}^{p_g} : \Omega_g(\boldsymbol{z}) \leq 1} \langle \boldsymbol{z}, \boldsymbol{u}\rangle$, and $X^{(g)} \in \mathbb{R}^{n \times p_g}$ is the design matrix X restricted to the features/columns in group g.

In our setting,

- $\Omega_g^D(\boldsymbol{u}) = \left\|\boldsymbol{\beta}^{(g)}\right\|_2$ and $\Omega^D(\boldsymbol{u}) = \max_{g=1,\ldots,G} \frac{1}{w_g} \left\|\boldsymbol{u}^{(g)}\right\|_2$.
- If one uses the squared loss, that is to say, $f_i(z) = \frac{1}{2}(y_i - z)^2$, then $f_i^*(z) = \frac{1}{2}z^2 + y_i z$ and the Lipschitz constant is $\gamma = 1$.

- If one uses the logistic loss, that is to say, $\boldsymbol{y} \in \{0,1\}^n$ and $f_i(z) = -y_i z + \log(1+\exp(z))$, then

$$f_i^*(z) = \begin{cases} (z+y_i)\log(z+y_i) + (1-(z+y_i))\log(1-(z+y_i)) & \text{if } 0 \leq (z+y_i) \leq 1 \\ +\infty & \text{otherwise,} \end{cases}$$

  and the Lipschitz constant is $\gamma = 4$.

The general idea of safe screening rules, introduced by [Laurent El Ghaoui(2010)], is to find a region $\mathcal{R} \subset \mathbb{R}^n$ such that if $\widehat{\boldsymbol{\theta}}^{(\lambda)} \in \mathcal{R}$, for any $g \in \{1,\ldots,G\}$,

$$\Omega_g^D\left(X^{(g)\top}\widehat{\boldsymbol{\theta}}^{(\lambda)}\right) < 1 \Rightarrow \widehat{\boldsymbol{\beta}}^{(\lambda)} = 0.$$

Gap safe screening rules [Ndiaye *et al.*(2017)] exploit the duality gap $(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))$ to obtain the radius of the safe region $\mathcal{R}$. More specifically, Ndiaye et al. show that $\forall \boldsymbol{\beta} \in \mathbb{R}^p, \forall \boldsymbol{\theta} \in \Delta_X$,

$$\left\|\widehat{\boldsymbol{\theta}}^{(\lambda)} - \boldsymbol{\theta}\right\|_2 \leq \sqrt{\frac{2(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))}{\gamma\lambda^2}},$$

which leads them to define, for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Delta_X$, the ball centered in $\boldsymbol{\theta}$ and of radius $\sqrt{\frac{2P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta})}{\gamma\lambda^2}}$ as a safe region, that is to say a region that is guaranteed to contain $\widehat{\boldsymbol{\theta}}^{(\lambda)}$.

D.2.4 Measuring selection stability

To measure the stability of a feature selection property, we use the sample's Pearson coefficient [Nogueira and Brown(2016)]. This stability index is closely related to that proposed by Kuncheva [Kuncheva(2008)] and is appropriate for the comparison of feature sets of different sizes. This index relies on repeating the feature selection procedure M time (in this work, $M = 10$) and evaluating the overlap if the M resulting feature sets.

Each of the M sets of selected features can be represented by an indicator vector $\boldsymbol{s} \in \{0,1\}^p$, where $s_j = 1$ if feature j is selected and 0 otherwise. The stability index between two feature sets $\mathcal{S}$ and $\mathcal{S}'$, represented by their indicator vectors $\boldsymbol{s}$ and $\boldsymbol{s}'$, is computed as the Pearsons's correlation between these two vectors:

$$\phi(\mathcal{S},\mathcal{S}') = \frac{\sum_{j=1}^p (s_j - \bar{\boldsymbol{s}})(s'_j - \bar{\boldsymbol{s}'})}{\sqrt{\sum_{j=1}^p (s_j - \bar{\boldsymbol{s}})^2}\sqrt{\sum_{j=1}^p (s'_j - \bar{\boldsymbol{s}'})^2}}, \tag{D.3}$$

where $\bar{\boldsymbol{s}} = \frac{1}{p}\sum_{j=1}^p s_j$ and $\bar{\boldsymbol{s}'} = \frac{1}{p}\sum_{j=1}^p s'_j$.

Note that, because $\sum_{j=1}^{p} s_j = |\mathcal{S}|$, $\sum_{j=1}^{p} s_j s'_j = |\mathcal{S} \cap \mathcal{S}'|$, and $s_j^2 = s_j$, we can rewrite Equation (D.3) as

$$\phi(\mathcal{S}, \mathcal{S}') = \frac{|\mathcal{S} \cap \mathcal{S}'| - \frac{1}{p} |\mathcal{S}| \, |\mathcal{S}'|}{\sqrt{|\mathcal{S}| \left(1 - \frac{|\mathcal{S}|}{p}\right)} \sqrt{|\mathcal{S}'| \left(1 - \frac{|\mathcal{S}'|}{p}\right)}},$$

hence interpreting this index as the size of the intersection of the two sets, corrected by chance, that is to say, ensuring that the expected value of the index is 0 when the two selections are random.

The stability index between M sets of selected features is computed as the average pairwise stability index between all possible pairs of sets of selected features:

$$\phi(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M) = \frac{M(M-1)}{2} \sum_{k=1}^{M} \sum_{l=k+1}^{M} \phi(\mathcal{S}_k, \mathcal{S}_l). \tag{D.4}$$

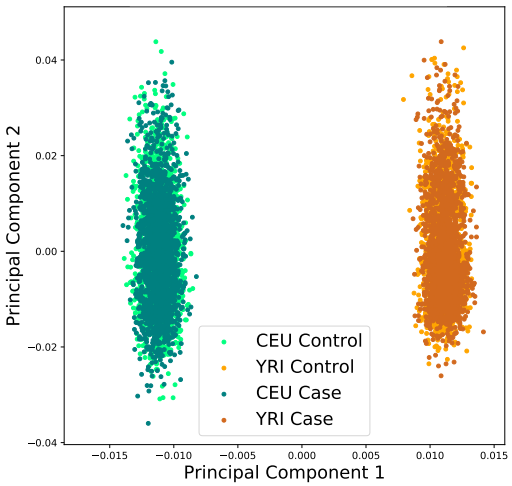
D.3 Supplementary Results

D.3.1 PCA of the genotypes

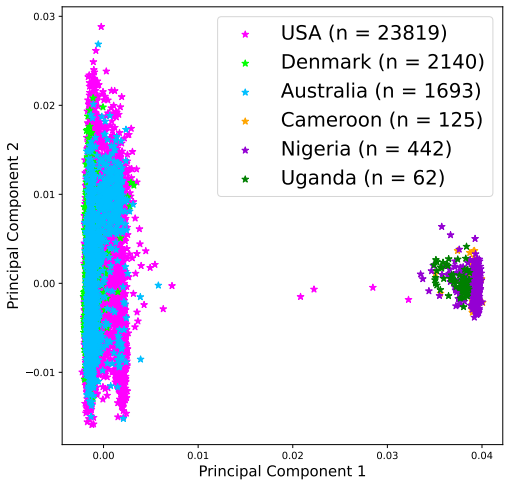
Figure D.3 shows the genotypes of the simulated data (Figure D.3a) and the DRIVE data (Figure D.3b) projected on the two first principal components of the data.

D.3.2 Runtimes

Figure D.4 shows the runtimes of the different Lasso methods on simulated data.



(a) Population structure in simulated data



(b) Population structure in the DRIVE data

Figure D.3: PCA for simulated and real datasets

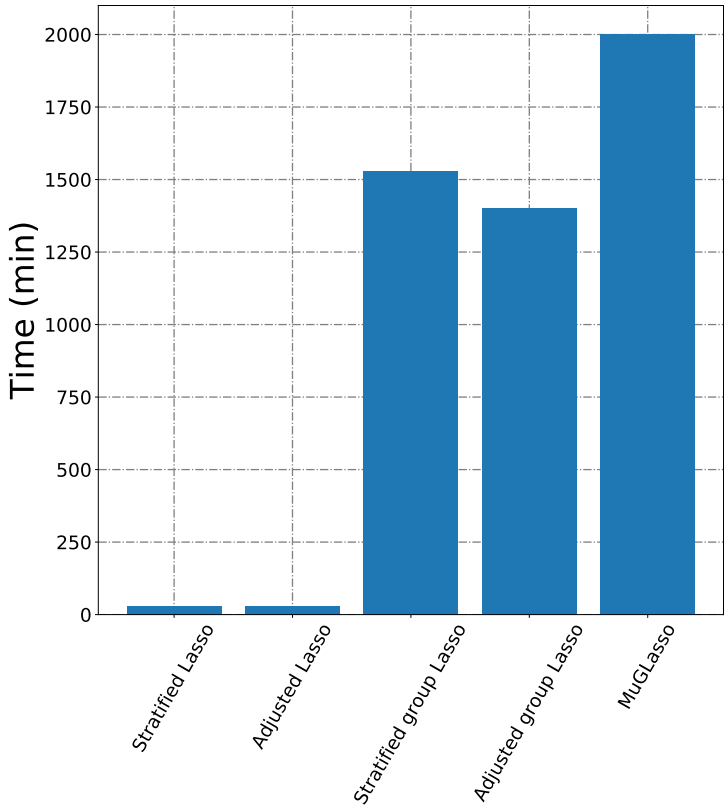


Figure D.4: Runtimes of the different Lasso approaches

D.3.3 Breast cancer risk loci detected by MuGLasso on DRIVE

On the DRIVE dataset, MuGLasso selected 1 357 SNPs, forming 62 LD groups. Those SNPs include all the 306 SNPs that are significant in the adjusted GWAS approach. We used FUMA [Watanabe *et al.*(2017)] to analyze the remaining 1 051 SNPs, and found that 57% of these SNPs are within 10kb of protein coding genes. Hence MuGLasso identifies a total of 32 genes (listed in in Table D.3), in addition to the 9 genes (*ITPR1*, *MRPS30*, *MAP3K1*, *SETD9*, *MIER3*, *EBF1*, *FGFR2*, *TOX3* and *MKL1*) identified by the adjusted GWAS.

Out of these 32 genes, 17 were previously identified in breast cancer meta-analyses which data include our 28 281 samples from the General Research Use dataset of the DRIVE Breast Cancer OncoArray Genotypes (see Table D.3). More specifically, these studies respectively used 10 707 ER-negative breast cancer cases 76 649 controls [Garcia-Closas *et al.*(2013)] 45 290 cases and 41 880 controls of European ancestry [Michailidou *et al.*(2013)], 62 623 breast cancer cases and 61 696 controls [Michailidou *et al.*(2015)], 122 977 cases and 105 974 controls of European ancestry together with 14 068 cases and 13 104 controls of East Asian ancestry [Michailidou *et al.*(2017)], and 210 088 controls (9 494 of which are BRCA1 mutation carriers) and 30 882 cases (21 468 ER-negative cases and 9 414 BRCA1 mutation carriers), all of European origin [Milne *et al.*(2017)].

This suggests that MuGLasso was able to rescue loci that are significant in a better-powered study (that is to say, a study with a larger number of samples).

In addition, we were able to find in the literature prior evidence of relationship with breast cancer risk or tumor growth for 7 additional genes, suggesting biological relevance of the MuGLasso findings.

Further analyses would be required to really get to the biological interpretation of these results. In particular, we restricted ourselves to mapping SNPs to genes based on a 10kb window, where other authors rather use 50kb, and FUMA provides many additional possibilities using known eQTLs and chromatin interactions across all tissues or for relevant tissues. In addition, pathway enrichment analyses could also be very relevant. One could also compare the selected SNPs to those significant in large meta-analyses such as [Milne *et al.*(2017), Michailidou *et al.*(2017)] in a more systematic manner to investigate how much power is gained by using MuGLasso on a subset of these GWAS data sets. Finally, we have analyzed jointly all selected SNPs and have not distinguished between those that are specific to one of the two populations and those that are common to both.

Genes found in meta-GWAS including the samples used in this work	
Gene symbols	Evidence
ASTN2	[Michailidou *et al.*(2017)]
CCDC170	[Garcia-Closas *et al.*(2013), Michailidou *et al.*(2013), Michailidou *et al.*(2015)] [Michailidou *et al.*(2017), Milne *et al.*(2017)]
CDYL2	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)]
DIRC3	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)] [Milne *et al.*(2017)]
ELL	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)] [Milne *et al.*(2017)]
ESR1	[Garcia-Closas *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)] [Milne *et al.*(2017)]
FTO	[Garcia-Closas *et al.*(2013), Michailidou *et al.*(2013), Michailidou *et al.*(2015)] [Michailidou *et al.*(2017), Milne *et al.*(2017)]
GRHL1	[Michailidou *et al.*(2017)]
KCNU1	[Michailidou *et al.*(2015), Michailidou *et al.*(2017)]
NEK10	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)] [Milne *et al.*(2017)]
PAX9	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)]
PTHLH	[Garcia-Closas *et al.*(2013), Michailidou *et al.*(2013), Michailidou *et al.*(2015)] [Michailidou *et al.*(2017), Milne *et al.*(2017)]
SSBP4	[Michailidou *et al.*(2017)]
TGFBR2	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)]
TNRC6B	[Michailidou *et al.*(2017)]
ZMIZ1	[Michailidou *et al.*(2013), Michailidou *et al.*(2015), Michailidou *et al.*(2017)]
ZNF365	[Michailidou *et al.*(2017), Milne *et al.*(2017)]
Genes found to be associated with breast cancer risk or tumor growth in the literature	
Gene symbols	Evidence
ADSL	oncogenic driver in triple negative breast cancer [Zurlo *et al.*(2019)]
CACNA1I	underexpressed in breast cancer [Phan *et al.*(2017)]
CCDC91	likely target gene of breast cancer risk variants [Ferreira *et al.*(2019)]
NUP205	forms a complex with NUP93 which regulates breast tumor growth [Bersini *et al.*(2020)]
POP1	expression correlates with prognosis in breast cancer [Liu *et al.*(2021)]
PPFIBP1	promotes cell motility and migration in breast cancer [Chiaretti *et al.*(2016)]
SGSM3	associated with breast cancer in a Chinese population [Tan *et al.*(2017)]
Other genes	
C7orf73, *CCSER1*, *CD2AP*, *HK1*, *HRSP12*, *LUC7L3*, *MED21*, *REP15*	

Table D.3: The 32 potential breast cancer risk genes within 10kb of loci identified by MuGLasso and not the adjusted GWAS, together with information as to their biological relevance

Appendix E

Sparse Multitask group Lasso (SMuGLasso) supplementary materials

E.1 Population stratification adjustment in Arabidopsis thaliana dataset

In this part, we study the population stratification confounder in Arabidopsis thaliana dataset. We also compare the adjustment methods presented in Chapter 4 (Section 2.2): genomic control, PCA-based models and linear mixed models.

As mentioned before in Section 5, we study a quantitative phenotype in Arabidopsis thaliana dataset. To apply PCA-based methods, we use linear regression to correct for population stratification, instead of logistic regression when the phenotype is qualitative (case-control).

Inflation factor Before adjustment for population structure, the inflation factor was $\lambda = 9.06$. This value denotes a strong population stratification case as λ exceeds extremely 1.

Comparing population stratification correction methods We conduct a comparison study of population stratification adjustment techniques. However, [Grimm *et al.*(2017)] recommend to use linear mixed models, more precisely FastLMM. Indeed, FastLMM outperforms the other methods in decreasing the inflation factor. It gives the lowest value near to 1. We present in Table E.1 the values of the inflation factor obtained after correcting for population stratification used the different methods mentioned above. We show also in Figure E.1 the Q-Q plots before adjustment (Figure E.1a) and after adjustment using FastLMM (Figure E.1b).

Methods	Inflation factor λ
Before adjustment	9.0609
GC	4.3872
EIGENSTRAT	1.5910
Linear regression with 10 PCs as covariates	1.7050
Linear regression after phenotype adjustment	1.8901
FastLMM	1.1006

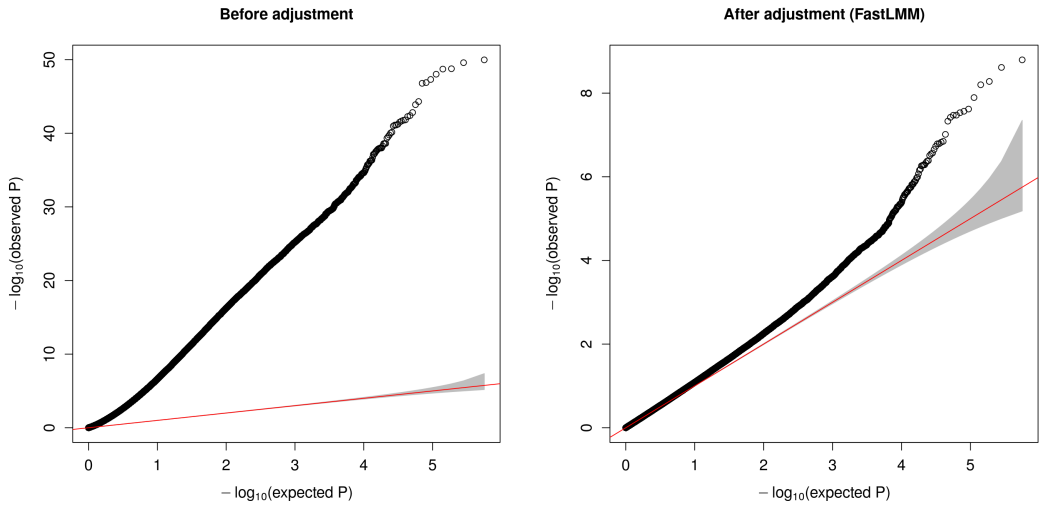
Table E.1: For Arabidopsis thaliana, inflation factor values are given before and after adjustment for population stratification

E.2 Breast cancer risk loci detected by SMuGLasso and MuGLasso on DRIVE

We present in Table E.2 the breast cancer risk genes findings using our developed models SMuGLasso and MuGLasso. We compare our discoveries to Adjusted GWAS findings.

E.3 DTF3 loci detected by SMuGLasso and MuGLasso on Arabidopsis thaliana dataset

We present in Table E.3 our genes findings related to the time to flowering duration in Arabidopsis thaliana plants using SMuGLasso and MuGLasso. We show also the discoveries of Adjusted GWAS.



(a) Before population stratification correction (b) After population stratification correction using FastLMM

Figure E.1: For Arabidopsis thaliana, Q-Q plots before and after population stratification adjustment

Adjusted GWAS	**ITPR1**, **MRPS30**, **MAP3K1**, **SETD9**, **MIER3**, **EBF1**, **FGFR2**, **TOX3**, **MKL1**
SMuGLasso	**ITPR1**, **MRPS30**, **MAP3K1**, **SETD9**, **MIER3**, **EBF1**, **FGFR2**, **TOX3**, **MKL1**, ADSL, ASTN2, CACNA1I, CCDC170, CCDC91, CDYL2, DIRC3, ELL, ESR1, FTO, GRHL1, HK1, HRSP12, KCNU1, NEK10, NUP205, PAX9, POP1, PPFIBP1, PTHLH, REP15, SGSM3, SSBP4, TGFBR2, TNRC6B, ZMIZ1, ZNF365.
MuGLasso	**ITPR1**, **MRPS30**, **MAP3K1**, **SETD9**, **MIER3**, **EBF1**, **FGFR2**, **TOX3**, **MKL1**, ADSL, ASTN2, C7orf73, CACNA1I, CCDC170, CCDC91, CCSER1, CD2AP, CDYL2, DIRC3, ELL, ESR1, FTO, GRHL1, HK1, HRSP12, KCNU1, LUC7L3, MED21, NEK10, NUP205, PAX9, POP1, PPFIBP1, PTHLH, REP15, SGSM3, SSBP4, TGFBR2, TNRC6B, ZMIZ1, ZNF365.

Table E.2: For DRIVE dataset, list of risk genes associated with breast cancer selected by SMuGLasso, MuGLasso and Adjusted GWAS. In bold are genes selected by Adjusted GWAS. CEU-specific selected genes are highlighted in blue and YRI-specific selected genes are highlighted in red. The others (in black) are risk genes shared across all populations

Methods	List of selected genes
Adjusted GWAS	**AT5G10100**, **AT5G45830**, **AT4G00730**, **AT4G00752**, **AT4G00630**, **AT4G00740**, **AT4G00750**.
SMuGLasso	**AT5G10100**, **AT5G45830**, **AT4G00730**, **AT4G00752**, **AT4G00630**, **AT4G00740**, **AT4G00750**, AT4G01915, AT5G15020, AT5G17710, AT5G27945, AT5G53410, AT3G58590, AT1G20130, AT3G29450, AT3G14490, AT1G03365, AT3G14470, AT1G28410, AT2G23430, AT3G27040, AT4G17970, AT4G09160, AT2G34890, AT4G30100, AT2G39990, AT4G35080, AT2G18500, AT3G46340, AT1G29300, AT3G27670, AT5G41820, AT2G38720, AT3G44610, AT4G33760, AT5G40450, AT1G27520, AT3G26140, AT4G16990, AT1G61360, AT3G61170, AT5G55910, AT2G25940, AT5G51830, AT1G43600, AT2G39310, AT4G34310, AT1G78970.
MuGLasso	**AT5G10100**, **AT5G45830**, **AT4G00730**, **AT4G00752**, **AT4G00630**, **AT4G00740**, **AT4G00750**, AT4G01915, AT5G15020, AT5G17710, AT5G27945, AT5G53410, AT3G58590, AT1G20130, AT3G29450, AT3G14490, AT1G03365, AT3G14470, AT1G28410, AT2G23430, AT3G27040, AT4G17970, AT4G09160, AT2G34890, AT4G30100, AT2G39990, AT4G35080, AT2G18500, AT3G46340, AT1G29300, AT3G27670, AT5G41820, AT2G38720, AT3G44610, AT4G33760, AT5G40450, AT1G27520, AT3G26140, AT4G16990, AT1G61360, AT1G43600, AT2G39310, AT4G34310, AT1G78970, AT4G33480, AT5G40290, AT1G12970, AT3G13550, AT2G32170, AT4G27290, AT1G59690, AT3G61170 AT5G55910, AT2G25940, AT5G51830.

Table E.3: For Arabidopsis thaliana dataset and for Adjusted GWAS, SMuGLasso and MuGLasso, list of selected genes associated with DTF3 trait. In bold are the genes selected by Adjusted GWAS. In blue are genes selected for specific populations. The others are shared genes selected across all populations

Bibliography

[Abdellaoui *et al.*(2013)] A. Abdellaoui et al.*Population structure, migration and diversifying selection in the Netherlands*, Eur J Hum Genet 21 (2013).

[Abegaz and et al.(2021)] F. Abegaz and F. V. L. et al., *Performance of model-based multifactor dimensionality reduction methods for epistasis detection by controlling population structure*, BioData Min (2021).

[Abegaz *et al.*(2018)] F. Abegaz, K. Chaichoompu, et al., *Principals about principal components in statistical genetics*, Bioinformatics (2018).

[Alexander and Lange(2011)] D. H. Alexander and K. Lange, *Stability selection for genome-wide association*, Genetic Epidemiology 35 (2011).

[Altshuler *et al.*(2010)] D. M. Altshuler, R. A. Gibbs, et al., *Integrating common and rare genetic variation in diverse human populations*, Nature (2010).

[Ambroise *et al.*(2019)] C. Ambroise et al., *Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics*, Algorithms Mol Biol (2019).

[Armitage(1955)] P. Armitage, *Tests for linear trends in proportions and frequencies.*, Biometrics (1955).

[Association(2009)] A. D. Association, *Diagnosis and classification of diabetes mellitus*, Diabetes Care (2009).

[Azencott *et al.*(2013)] C.-A. Azencott, D. Grimm, et al., *Efficient network-guided multi-locus association mapping with graph cuts*, Bioinformatics (2013).

[Bach(2008)] F. R. Bach, *Bolasso: Model consistent lasso estimation through the bootstrap*, International Conference on Machine Learning (2008).

[Bellon *et al.*(2016)] V. Bellon, V. Stoven, and C.-A. Azencott, *Multitask feature selection with task descriptors.*, Pacific Symposium on Biocomputing (2016).

[Bermingham *et al.*(2015)] M. L. Bermingham, R. Pong-Wong, et al., *Application of high-dimensional feature selection: evaluation for genomic prediction in man*, Scientific Reports (2015).

[Bersini *et al.*(2020)] S. Bersini, N. K. Lytle, et al., *Nup93 regulates breast tumor growth by modulating cell proliferation and actin cytoskeleton remodeling*, Life Sci Alliance (2020).

[Boehnke(2000)] M. Boehnke, *A look at linkage disequilibrium*, Nat Genet **25** (3), 246 (2000).

[Box and Cox(1964)] G. E. P. Box and D. R. Cox, *An analysis of transformations*, Journal of the Royal Statistical Society. Series B (Methodological) (1964).

[Byers(2008)] D. L. Byers, *Components of phenotypic variance*, Nature Education 1 (2008).

[Cantley and Ashcroft(2015)] J. Cantley and F. M. Ashcroft, *Q&a: insulin secretion and type 2 diabetes: why do beta-cells fail?*, BMC Biol (2015).

[Chiaretti *et al.*(2016)] S. Chiaretti, V. Astro, et al., *Effects of the scaffold proteins liprin-α1, β1 and β2 on invasion by breast cancer cells.*, Biol Cell. (2016).

[Cho *et al.*(2009)] S. Cho, H. Kim, et al., *Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis.*, BMC Proc (2009).

[Cho and Kim(2010)] S. Cho and K. Kim, *Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis*, Ann Hum Genet. (2010).

[Christopher *et al.*(2016)] A. Christopher, D. Joe, et al., *The oncoarray consortium: a network for understanding the genetic architecture of common cancers*, Cancer Epidemiol Biomarkers Prev. 2017 10.17863/CAM.6139 (2016).

[Climente-González *et al.*(2021)] H. Climente-González, C. Lonjou, et al., *Boosting gwas using biological networks: A study on susceptibility to familial breast cancer*, PLoS Comput Biol. (2021).

[Consortium(2003)] I. H. Consortium, *The international hapmap project*, Nature (2003).

[Consortium(2006)] M. Consortium, *The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements.*, Nat Biotechnol (2006).

[Consortium(2016)] T. . G. Consortium, *1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana*, Cell. (2016).

[Consortium(2015)] T. . G. P. Consortium, *A global reference for human genetic variation.*, Nature (2015).

[Consortium(2007)] T. W. T. C. C. Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*, Nature (2007).

[Cortes *et al.*(2015)] A. Cortes, S. L. Pulit, et al., *Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with erap1*, Nature Communications (2015).

[Davis *et al.*(2006)] C. A. Davis, F. Gerick, et al., *Reliable gene signatures for microarray classification: assessment of stability and performance.*, Bioinformatics (2006).

[Dehman *et al.*(2015)] A. Dehman, C. Ambroise, and P. Neuvial, *Performance of a blockwise approach in variable selection using linkage disequilibrium information*, BMC Bioinformatics (2015).

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

[Devlin and Roeder(1999)] B. Devlin and K. Roeder, *Genomic control for association studies*, Biometrics (1999).

[Dunne *et al.*(2002)] K. Dunne, P. Cunningham, and F. Azuaje, *Solutions to instability problems with sequential wrapper-based approaches to feature selection.*, Technical Report TCD-CS-2002-28, Trinity College Dublin, School of Computer Science (2002).

[Elhaik(2021)] E. Elhaik, *Why most principal component analyses (pca) in population genetic studies are wrong*, Preprint BioRxiv (2021).

[Fan *et al.*(2016)] Y. Fan, F. B. Rina, et al., *Selective inference for group-sparse linear models*, Adv Neural Inf Process Syst **29**, 2469 (2016).

[Fang *et al.*(2022)] J. Fang, P. Zhang, et al., *Artificial intelligence framework identifies candidate targets for drug repurposing in alzheimer's disease*, Alzheimer's Research & Therapy (2022).

[Fergus *et al.*(2020)] P. Fergus, C. C. Montanez, et al., *Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in african-american women*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020).

[Ferreira *et al.*(2019)] M. A. Ferreira, E. R. Gamazon, et al., *Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer*, Nature Communications (2019).

[Fu *et al.*(2011)] J. Fu, E. A. Festen, and C. Wijmenga, *Multi-ethnic studies in complex traits*, Hum Mol Genet. (2011).

[Garcia-Closas *et al.*(2013)] M. Garcia-Closas, F. J. Couch, et al., *Genome-wide association studies identify four er negative-specific breast cancer risk loci*, Nat Genet. (2013).

[GBD2013(2015)] GBD2013, *Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013*, Europe PMC Author Manuscripts (2015).

[Goh and Wong(2016)] W. W. B. Goh and L. Wong, *Evaluating feature-selection stability in nextgeneration proteomics.*, Journal of Bioinformatics and Computational Biology (2016).

[Grimm *et al.*(2017)] D. G. Grimm, D. Roqueiro, et al., *easygwas: A cloud-based platform for comparing the results of genome-wide association studies*, The Plant Cell (2017).

[Guo *et al.*(2018)] Q. Guo, Y. Wang, et al., *Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies*, Bone Res (2018).

[Gusareva and Steen(2014)] E. S. Gusareva and K. V. Steen, *Practical aspects of genome-wide association interaction analysis*, Hum Genet (2014).

[Guzman-Martinez *et al.*(2011)] R. Guzman-Martinez et al., *Feature selection stability assessment based on the jensen-shannon divergence.*, European Conference on Machine Learning. Springer (2011).

[Haury *et al.*(2011)] A.-C. Haury, P. Gestraud, and J.-P. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures*, PLoS ONE (2011).

[Haury *et al.*(2012)] A.-C. Haury et al., *TIGRESS: trustful inference of gene regulation using stability selection*, BMC Systems Biology (2012).

[Howie *et al.*(2009)] B. N. Howie et al., *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*, PLoS Genetics (2009).

[Hu *et al.*(2016)] X. Hu, W. Zhang, et al., *Group-combined p-values with applications to genetic association studies*, Bioinformatics (2016).

[Jacob *et al.*(2009)] L. Jacob, G. Obozinski, and J.-P. Vert, *Group lasso with overlap and graph lasso*, International Conference on Machine Learning (2009).

[Janssens and van Duijn(2008)] A. C. J. W. Janssens and C. M. van Duijn, *Genome-based prediction of common diseases: advances and prospects*, Hum Mol Genet. (2008).

[Jr. *et al.*(2019)] R. P. I. Jr., T. G. Kinzy, and J. N. C. Bailey, *Genetic risk scores*, Curr Protoc Hum Genet. (2019).

[Kalousis *et al.*(2005)] A. Kalousis, J. Prados, and M. Hilario., *Stability of feature selection algorithms.*, IEEE International Conference on Data Mining (2005).

[Kalousis *et al.*(2007)] A. Kalousis, J. Prados, and M. Hilario, *Stability of feature selection algorithms: A study on high-dimensional spaces.*, Knowledge and Information Systems (2007).

[Kang *et al.*(2010)] H. M. Kang, J. H. Sul, and S. K. S. et al., *Variance component model to account for sample structure in genome-wide association studies.*, Nature Genetics (2010).

[Kang *et al.*(2008)] H. M. Kang, N. A. Zaitlen, et al., *Efficient control of population structure in model organism association mapping*, Genetics (2008).

[Khan *et al.*(2020)] M. A. B. Khan, M. J. Hashim, et al., *Epidemiology of type 2 diabetes – global burden of disease and forecasted trends*, J Epidemiol Glob Health (2020).

[Kriti *et al.*(2010)] P. Kriti, K. Seyoung, and X. E. P, *Multi-population GWA mapping via multi-task regularized regression*, Bioinformatics **26** (12), i208 (2010).

[Krizek *et al.*(2007)] P. Krizek, J. Kittler, and V. Hlavac, *Improving stability of feature selection methods.*, CAIP (2007).

[Kuncheva(2008)] L. I. Kuncheva, *A stability index for feature selection*, IASTED ICAIA (2008).

[LaFramboise(2009)] T. LaFramboise, *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.*, Nucleic Acids Res (2009).

[Laurent El Ghaoui(2010)] T. R. Laurent El Ghaoui, Vivian Viallon, *Safe feature elimination for the lasso and sparse supervised learning problems*, Pacific Journal of Optimization (2010).

[Lausser *et al.*(2013)] L. Lausser, C. Mussel, et al., *Measuring and visualizing the stability of biomarker selection techniques.*, Computational Statistics (2013).

[Li and Li(2008)] C. Li and M. Li, *Gwasimulator: a rapid whole-genome simulation program*, Bioinformatics (2008).

[Li *et al.*(2010)] J. Li, K. Das, et al., *The bayesian lasso for genome-wide association studies*, Bioinformatics (2010).

[Li *et al.*(2020)] L. Li et al., *Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza A (H1N1) virus using mutations and variations in glycosylation of hemagglutinin*, BMC Bioinformatics **21**, 1 (2020).

[Li *et al.*(2018)] X. Li, L. Liu, et al., *Heterogeneity analysis and diagnosis of complex diseases based on deep learning method*, Scientific Reports (2018).

[Lin *et al.*(2014)] D. Lin et al., *Integrative analysis of multiple diverse omics datasets by sparse group multitask regression*, Front Cell Dev Biol **2**, 62 (2014).

[Lin and Zeng(2011)] D. Y. Lin and D. Zeng, *Correcting for population stratification in genomewide association studies*, J Am Stat Assoc (2011).

[Lippert *et al.*(2011)] C. Lippert et al., *FaST linear mixed models for genome-wide association studies*, Nat Methods (2011).

[Liu *et al.*(2012)] J. Liu, J. Huang, et al., *Incorporating group correlations in genome-wide association studies using smoothed group lasso*, Biostatistics (2012).

[Liu *et al.*(2021)] Y. Liu, H. Sun, et al., *Identification of a three-rna binding proteins (rbps) signature predicting prognosis for breast cancer*, Front Oncol. (2021).

[Loos and Janssens(2017)] R. J. F. Loos and A. C. J. W. Janssens, *Predicting polygenic obesity using genetic information*, Cell Metab. (2017).

[Lustgarten *et al.*(2009)] J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran, *Measuring stability of feature selection in biomedical datasets.*, AMIA Annu Symp Proc (2009).

[Manolio *et al.*(2009)] T. A. Manolio, F. S. Collins, and N. J. C. et al., *Finding the missing heritability of complex diseases.*, Nature (2009).

[Mayhew and Meyre(2017)] A. J. Mayhew and D. Meyre, *Assessing the heritability of complex traits in humans: methodological challenges and opportunities*, Curr Genomics (2017).

[Medina-Gomez *et al.*(2015)] C. Medina-Gomez, J. F. Felix, et al., *Challenges in conducting genome-wide association studies in highly admixed multiethnic populations: the generation r study*, Eur J Epidemiol. (2015).

[Meinshausen and Bühlmann(2009)] N. Meinshausen and P. Bühlmann, *Stability selection*, J. R. Statist. Soc. B (2009).

[Michailidou *et al.*(2015)] K. Michailidou, J. Beesley, et al., *Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer*, Nat Genet. (2015).

[Michailidou *et al.*(2013)] K. Michailidou, P. Hall, et al., *Large-scale genotyping identifies 41 new loci associated with breast cancer risk*, Nat Genet. (2013).

[Michailidou *et al.*(2017)] K. Michailidou, S. Lindström, et al., *Association analysis identifies 65 new breast cancer risk loci*, Nature (2017).

[Milne *et al.*(2017)] R. L. Milne, K. B. Kuchenbaecker, et al., *Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer*, Nature Genetics (2017).

[Naidoo *et al.*(2011)] N. Naidoo, Y. Pawitan, et al., *Human genetics and genomics a decade after the release of the draft sequence of the human genome*, Human Genomics (2011).

[Nakamoto *et al.*(2006)] K. Nakamoto, S. Wang, et al., *Linkage disequilibrium blocks, haplotype structure, and htsnps of human cyp7a1 gene*, BMC Genet. (2006).

[Ndiaye *et al.*(2017)] E. Ndiaye et al., *Gap safe screening rules for sparsity enforcing penalties*, Journal of Machine Learning Research 18 (2017).

[Need *et al.*(2009)] A. C. Need et al., *A genome-wide investigation of SNPs and CNVs in schizophrenia*, PLOS Genetics (2009).

[Nogueira and Brown(2015)] S. Nogueira and G. Brown, *Measuring the stability of feature selection with applications to ensemble methods*, International Workshop on Multiple Classifier Systems (2015).

[Nogueira and Brown(2016)] S. Nogueira and G. Brown, *Measuring the stability of feature selection*, Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2016).

[Nogueira *et al.*(2018)] S. Nogueira, K. Sechidis, and G. Brown, *On the stability of feature selection algorithms.*, Journal of Machine Learning Research (2018).

[Nolte *et al.*(2017)] I. M. Nolte, P. J. van der Most, et al., *Missing heritability: is the gap closing? an analysis of 32 complex traits in the lifelines cohort study*, European Journal of Human Genetics (2017).

[Novembre and Stephens(2008)] J. Novembre and M. Stephens, *Interpreting principal component analyses of spatial population genetic variation*, Nature genetics (2008).

[Obozinski *et al.*(2006)] G. Obozinski, B. Taskar, and M. Jordan, *Multi-task feature selection*, Technical report, UC Berkeley (2006).

[Ochoa and Storey(2021)] A. Ochoa and J. D. Storey, *Estimating fst and kinship for arbitrary population structures*, PLOS GENETICS (2021).

[Okser *et al.*(2013)] S. Okser, T. Pahikkala, and T. Aittokallio, *Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives*, BioData Min (2013).

[O'Brien and Szu(2017)] A. O'Brien and P. Szu, *Breaking the curse of dimensionality for machine learning on genomic data*, BAI@IJCAI (2017).

[Pardiñas *et al.*(2018)] A. F. Pardiñas, P. Holmans, et al., *Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection*, Nature Genetics (2018).

[Park(2019)] L. Park, *Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants*, Scientific Reports (2019).

[Patterson *et al.*(2006)] N. Patterson, A. L. Price, and D. Reich, *Population structure and eigenanalysis*, PLoS Genet (2006).

[Peloso and Lunetta(2011)] G. M. Peloso and K. L. Lunetta, *Choice of population structure informative principal components for adjustment in a case-control study.*, BMC Genetics (2011).

[Peloso *et al.*(2009)] G. M. Peloso, N. Timofeev, and K. L. Lunetta, *Principal-component-based population structure adjustment in the north american rheumatoid arthritis consortium data: impact of single-nucleotide polymorphism set and analysis method.*, BMC Proc (2009).

[Peres-Neto *et al.*(2005)] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, *How many principal components? stopping rules for determining the number of non-trivial axes revisited*, Computational Statistics & Data Analysis 49 (2005).

[Phan *et al.*(2017)] N. N. Phan, C.-Y. Wang, et al., *Voltage-gated calcium channels: Novel targets for cancer therapy*, Oncol Lett. (2017).

[Price *et al.*(2010)] A. L. Price, N. A. Zaitlen, et al., *New approaches to population stratification in genome-wide association studies.*, Nat Rev Genet (2010).

[Price *et al.*(2006)] A. L. Price et al., *Principal components analysis corrects for stratification in genome-wide association studies*, Nat Genet (2006).

[Privé(2021)] F. Privé, *Optimal linkage disequilibrium splitting*, Bioinformatics (2021).

[Privé *et al.*(2018)] F. Privé, H. Aschard, et al., *Efficient analysis of large-scale genome-wide data with two r packages: bigstatsr and bigsnpr*, Bioinformatics. (2018).

[Privé *et al.*(2020)] F. Privé, K. Luu, et al., *Efficient toolkit implementing best practices for principal component analysis of population genetic data*, Bioinformatics (2020).

[Purcell *et al.*(2007)] S. Purcell et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses*, Am J Human Genet (2007).

[Qizhai and Kai(2008)] L. Qizhai and Y. Kai, *Improved correction for population stratification in genome-wide association studies by identifying hidden population structures*, Genetic Epidemiology (2008).

[Reich *et al.*(2001)] D. E. Reich et al., *Linkage disequilibrium in the human genome*, Nature **411** (6834), 199 (2001).

[Rosenberg *et al.*(2010)] N. A. Rosenberg, L. Huang, et al., *Genome-wide association studies in diverse populations*, Nat Rev Genet. (2010).

[Rybicki and Elston(2000)] B. A. Rybicki and R. C. Elston, *The relationship between the sibling recurrence-risk ratio and genotype relative risk*, Am J Hum Genet. (2000).

[Sabourin *et al.*(2019)] J. A. Sabourin, C. D. Cropp, et al., *Compass-gwas: A method to reduce type i error in genome-wide association studies when replication data are not available*, Genet Epidemiol (2019).

[Sebastian *et al.*(2014)] O. Sebastian et al., *Regularized machine learning in the genetic prediction of complex traits*, PLoS Genetics **10** (11), e1004754 (2014).

[Shah and Samworth(2013)] R. D. Shah and R. J. Samworth, *Variable selection with error control: another look at stability selection*, J R Stat Soc B (2013).

[Silver and Montana(2011)] M. Silver and G. Montana, *Pathway selection for gwas using the group lasso with overlaps*, International Conference on Bioscience, Biochemistry and Bioinformatics (2011).

[Slatkin(2008)] M. Slatkin, *Linkage disequilibrium — understanding the evolutionary past and mapping the medical future*, Nat Rev Genet (2008).

[Slim *et al.*(2022)] L. Slim, C. Chatelain, and C.-A. Azencott, *Nonlinear post-selection inference for genome-wide association studies*, Pacific Symposium on Biocomputing (PSB) (2022).

[Somol and Novovicova(2010)] P. Somol and J. Novovicova, *Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality.*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2010).

[Sugiyama *et al.*(2014)] M. Sugiyama, C.-A. Azencott, et al., *Multi-task feature selection on multiple networks via maximum flows*, Proceedings of the 2014 SIAM International Conference on Data Mining (2014).

[Tak and Farnham(2015)] Y. G. Tak and P. J. Farnham, *Making sense of gwas: using epigenomics and genome engineering to understand the functional relevance of snps in non-coding regions of the human genome*, Epigenetics & Chromatin (2015).

[Tan *et al.*(2017)] T. Tan, K. Zhang, and W. Chen, *Genetic variants of esr1 and sgsm3 are associated with the susceptibility of breast cancer in the chinese population*, Breast Cancer. (2017).

[Teo *et al.*(2009)] Y. Y. Teo, A. E. Fry, et al., *Genome-wide comparisons of variation in linkage disequilibrium*, Genome Res. (2009).

[Tibshirani(1996)] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (1996).

[Tishkoff *et al.*(2006)] S. A. Tishkoff, F. A. Reed, et al., *Convergent adaptation of human lactase persistence in africa and europe*, Nature Genetics (2006).

[Visscher *et al.*(2017)] P. M. Visscher et al., *10 years of gwas discovery: Biology, function, and translation*, Am J Human Genet **101** (2017).

[Wald *et al.*(2013)] R. Wald, T. M. Khoshgoftaar, and A. Napolitano, *Stability of filter- and wrapper-based feature subset selection.*, International Conference on Tools with Artificial Intelligence. IEEE Computer Society (2013).

[Waldmann *et al.*(2013)] P. Waldmann, G. Mészáros, et al., *Evaluation of the lasso and the elastic net in genome-wide association studies*, Front. Genet. (2013).

[Wang *et al.*(2019)] H. Wang, T. Yue, et al., *Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies*, BMC Bioinformatics (2019).

[Wang *et al.*(2012)] H. Wang et al., *Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort*, Bioinformatics **28** (2), 229 (2012).

[Watanabe *et al.*(2017)] K. Watanabe, E. Taskesen, et al., *Functional mapping and annotation of genetic associations with fuma*, Nat Commun (2017).

[Wei *et al.*(2014)] W.-H. Wei, G. Hemani, and C. S. Haley, *Detecting epistasis in human complex traits*, Nature Reviews Genetics volume (2014).

[Weigel and Mott(2009)] D. Weigel and R. Mott, *The 1001 genomes project for arabidopsis thaliana*, Genome Biology (2009).

[WS and JH(2012)] B. WS and M. JH, *Chapter 11: Genome-wide association studies.*, PLoS Comput Biol (2012).

[Wu *et al.*(2011)] C. Wu, A. DeWan, et al., *A comparison of association methods correcting for population stratification in case–control studies*, Ann Hum Genet (2011).

[Wu *et al.*(2013)] C. Wu, D. Li, et al., *Genome-wide association study identifies common variants in slc39a6 associated with length of survival in esophageal squamous-cell carcinoma*, Nat Genet (2013).

[Wu and Chen(2009)] T. T. Wu and Y. F. Chen, *Genome-wide association analysis by lasso penalized logistic regression*, Bioinformatics (2009).

[Xiaoli *et al.*(2017)] L. Xiaoli et al., *Group guided sparse group lasso multitask learning for cognitive performance prediction of alzheimer's disease*, Int Conf on Brain Inform (2017).

[Xu *et al.*(2021)] J. Xu, Y. Hou, et al., *A network-based deep learning framework catalyzes gwas and multi-omics findings to biology and drug repurposing for alzheimer's disease*, Cold Spring Harbor Laboratory (2021).

[Yang and Wen(2020)] S. Yang and J. Wen, *Prioritizing genetic variants in gwas with lasso using permutation-assisted tuning*, Bioinformatics (2020).

[Yang *et al.*(2017)] T. Yang, P. Thompson, et al., *Identifying genetic risk factors via sparse group lasso with group graph structure*, arXiv:1709.03645v1 (2017).

[Yaohui and Patrick(2017)] Z. Yaohui and B. Patrick, *The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in R*, The R Journal (2017).

[Yiwei and Wei(2015)] Z. Yiwei and P. Wei, *Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements?*, Genet Epidemiol **39** (3), 149 (2015).

[Yu *et al.*(2006)] J. Yu et al., *A unified mixed-model method for association mapping that accounts for multiple levels of relatedness*, Nat Genet (2006).

[Yu *et al.*(2008a)] K. Yu, Z. Wang, et al., *Population substructure and control selection in genome-wide association studies.*, PLoS One (2008a).

[Yu *et al.*(2008b)] L. Yu, C. H, et al., *Stable feature selection via dense feature groups.*, KDD (2008b).

[Yuan and Lin(2006)] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser.B (2006).

[Zeggini *et al.*(2008)] E. Zeggini et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes*, Nat Genet (2008).

[Zhang and Zhang(2014)] C.-H. Zhang and S. S. Zhang, *Confidence intervals for low-dimensional parameters in high-dimensional linear models*, J. R. Stat. Soc. Ser. B Stat. Methodol. (2014).

[Zhang *et al.*(2009)] M. Zhang, L. Zhang, et al., *Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes.*, Bioinformatics (2009).

[Zhang *et al.*(2010)] Z. Zhang, E. Ersoz, et al., *Mixed linear model approach adapted for genome-wide association studies.*, Nature Genetics (2010).

[Zhao *et al.*(2018)] H. Zhao, N. Mitra, et al., *A practical approach to adjusting for population stratification in genome-wide association studies: Principal components and propensity scores (pcaps)*, Stat Appl Genet Mol Biol (2018).

[Zou and Hastie(2005)] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Statist. Soc. B (2005).

[Zubair *et al.*(2016)] N. Zubair, M. Graff, et al., *Fine-mapping of lipid regions in global populations discovers ethnic-specific signals and refines previously identified lipid loci*, Hum Mol Genet. (2016).

[Zucknick *et al.*(2008)] M. Zucknick, S. Richardson, and E. A. Stronach, *Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods.*, Statistical Applications in Genetics and Molecular Biology (2008).

[Zuk *et al.*(2012)] O. Zuk, E. Hechter, et al., *The mystery of missing heritability: Genetic interactions create phantom heritability*, Proceedings of the National Academy of Sciences (2012).

[Zurlo *et al.*(2019)] G. Zurlo, X. Liu, et al., *Prolyl hydroxylase substrate adenylosuccinate lyase is an oncogenic driver in triple negative breast cancer*, Nature Communications (2019).

RÉSUMÉ

Les études d'association pangénomiques, ou les GWAS ont pour objectif de détecter des polymorphismes nucléotidiques (SNPs) associés à un phénotype d'intérêt. Parmi ses défis, le problème de la grande dimensionnalité des données qui se manifeste par le faible nombre d'échantillons disponibles. D'autres facteurs limitants incluent notamment la corrélation entre les SNPs, à cause du déséquilibre de liaison (LD), la structure de la population, c'est-à-dire, la confusion due à l'ascendance génétique et la faible puissance statistique en détectant un nombre limité de SNPs significatifs. Les modèles d'apprentissage automatique basés sur l'analyse multivariée contribue à avancer la recherche en GWAS. Par conséquent, les modèles de sélection de variables réduisent la dimensionnalité des données en ne conservant que les variables pertinentes. Cependant, ces méthodes manquent de la stabilité, c'est-à-dire de la robustesse suite à des légères variations dans le jeu de données d'entrée, ce qui peut conduire à une fausse interprétation biologique. Par conséquent, nous nous concentrons dans cette thèse sur l'évaluation et l'amélioration de la stabilité de sélection comme il s'agit d'un indicateur important pour avoir de la confiance aux SNPs découverts. Dans cette thèse, nous développons deux nouvelles méthodes efficaces (multitask group lasso et sparse multitask group lasso) basées sur l'analyse multivariée de Lasso sur des données multi-populations. Chaque tâche correspond à une sous-population des données et chaque groupe à un LD-groupe. Cette formulation atténue le problème de fléau de la dimension et permet d'identifier des LD-groupes pertinents partagés entre les populations/tâches, ainsi que certains LD-groupes qui sont spécifiques à une population/tâche. De plus, nous utilisons la sélection de stabilité pour augmenter la robustesse de nos approches. Enfin, les règles "Gap Safe Screening Rules" accélèrent les calculs en permettant à nos méthodes de fonctionner à l'échelle génomique. En analysant plusieurs données, dont un ensemble de données sur le cancer du sein, l'efficacité des modèles développés a été démontrée dans la découverte de nouveaux gènes à risque liés à la maladie.

MOTS CLÉS

GWAS, apprentissage automatique, sélection de variables, multitask group lasso, sparse multitask group lasso, stabilité de sélection.

ABSTRACT

Genome-Wide Association Studies, or GWAS, aim at finding Single Nucleotide Polymorphisms (SNPs) that are associated with a phenotype of interest. GWAS are known to suffer from the large dimensionality of the data with respect to the number of available samples. Many challenges limiting the identification of causal SNPs such as dependency between SNPs, due to linkage disequilibrium (LD), the population stratification and the low of statistical of univariate analysis. Machine learning models based on multivariate analysis contribute to advance research in GWAS. Hence, feature selection models reduce the dimensionality of data by keeping only the relevant features associated with disease. However, these methods lack of stability, that is to say, robustness to slight variations in the input dataset. This major issue can lead to false biological interpretation. Hence, we focus in this thesis on evaluating and improving the stability as it is an important indicator to trust feature selection discoveries. In this thesis, we develop two efficient novel methods (multitask group lasso and sparse multitask group lasso) for the multivariate analysis of multi-population GWAS data based on a two multitask group Lasso formulations. Each task corresponds to a subpopulation of the data, and each group to an LD-block. This formulation alleviates the curse of dimensionality, and makes it possible to identify disease LD-blocks shared across populations/tasks, as well as some that are specific to one population task. In addition, we use stability selection to increase the robustness of our approach. Finally, gap safe screening rules speed up computations enough that our method can run at a genome-wide scale. By analyzing several data including breast cancer dataset, the efficiency of the developed models was demonstrated in discovering new risk genes related to disease.

KEYWORDS

GWAS, machine learning, feature selection, multitask group lasso, sparse multitask group lasso, stability selection.