PSL UNIVERSITÉ PARIS

Préparée à MINES ParisTech

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Network-guided genome-wide association studies

Études d'association génome entier guidées par des réseaux

Soutenue par

Héctor Climente González

Le 4 Février 2020

Fin de confidentialité

Le 4 Février 2021

École doctorale n^{o}621

Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Bio-informatique

Composition du jury :

Nadine ANDRIEU Mme., Institut Curie	*Présidente*
Kristel VAN STEEN Mme., Université de Liège	*Rapporteuse*
Antonio RAUSELL M., Imagine Institute	*Rapporteur*
Laura FURLONG Mme., Pompeu Fabra University	*Examinatrice*
Véronique STOVEN Mme., MINES ParisTech	*Directrice de thèse*
Chloé-Agathe AZENCOTT Mme., MINES ParisTech	*Co-encadrante*

Contents

Acknowledgements **xi**

1 Context **1**

1.1 The common disease/common variant framework 2

1.2 Genome-wide association studies 4

1.2.1 Challenges 5

1.2.1.1 Low statistical power 5

1.2.1.2 Choice of encoding 6

1.2.1.3 Estimating individual risk 6

1.2.1.4 Population structure 7

1.2.1.5 Interpretability 7

1.3 Epistasis 8

1.4 Genome-wide association interaction studies 8

1.5 Diseases studied in this thesis 9

1.5.1 Breast cancer 10

1.5.1.1 The GENESIS dataset 11

1.5.2 Inflammatory bowel disease 12

1.5.2.1 The IIBDGC dataset 12

1.6 Network view of complex diseases 12

1.6.1 Networks in disease 14

1.6.2 Network-guided approaches to disease study 15

1.6.2.1 High-score subnetwork search 15

1.6.2.2 Module detection 15

1.6.2.3 Aggregation of networks 17

1.7 Contributions 17

2 Combining network-guided GWAS to discover susceptibility mechanisms for breast cancer **21**

2.1 Introduction 22

2.2 Methods 23

2.2.1 Data preprocessing and quality control 23

2.2.2 High-score subnetwork search algorithms 25

2.2.2.1 SNP and gene association 25

2.2.2.2 Mathematical notation 25

2.2.2.3 Methods used 26

2.2.2.4 Gene-gene network 29

2.2.2.5 SNP networks 30

2.2.2.6 Consensus network 30

2.2.3 Evaluation of methods 30

2.2.3.1 Classification accuracy of selected biomarkers 30

2.2.3.2 Biological relevance of the genes 31

2.2.4 Code availability 32

2.3 Results 32
2.3.1 A conventional GWAS shows that FGFR2 is strongly associated with familial breast cancer 32
2.3.2 Network methods successfully identify genes associated with breast cancer 32
2.3.3 heinz retrieves a small, highly informative set of biomarkers in a fast and stable fashion 37
2.3.4 No solution is perfect 42
2.3.5 Aggregating solutions provides insights into the biology of cancer 45
2.4 Discussion 47

3 The *martini* R package 53
3.1 Introduction 54
3.2 Improvements over SConES 54
3.2.1 Covariates and additional measures of association 54
3.2.2 Hyperparameter optimization 54
3.2.3 Network-based simulations 56
3.2.4 Interface, documentation and quality assurance 58
3.3 The `scones.nf` pipeline 58
3.4 Conclusions 59

4 Boosting interpretability and statistical power in epistasis detection by using prior biological knowledge 61
4.1 Introduction 63
4.2 Materials and methods 63
4.2.1 Dataset and initial quality control 63
4.2.2 Gene interaction detection procedure 63
4.2.2.1 Functional SNP pre-filtering 64
4.2.2.2 Post-filtering quality control 65
4.2.2.3 SNP-level epistasis detection and multiple test correction 65
4.2.2.4 From SNP-level to gene-level epistasis 66
4.3 Results 67
4.3.1 Type I error 67
4.3.2 Chromatin contacts map more SNPs per gene than other mappings 67
4.3.3 The *Physical* protocol does not recover any SNP interaction . 67
4.3.4 Gene-level network 69
4.3.5 *Chromatin* and *Standard* mappings partially replicate previous studies on IBD 72
4.4 Discussion 72

5 High-order epistasis detection through fusion of epistasis networks 77
5.1 Introduction 78
5.2 Materials and methods 78

5.2.1 Data, quality control and preprocessing 78
5.2.2 Epistasis detection methods 79
5.2.2.1 Linear regression 79
5.2.2.2 MB-MDR 79
5.2.2.3 EpiHSIC 79
5.2.3 High-order epistasis detection 80
5.2.4 Code availability 80
5.3 Results 80
5.3.1 Epistasis detection methods produce relatively similar results 80
5.3.2 High-order epistasis interactions in IBD 81
5.3.3 Mapping SNP to genes involves the complement system . . . 83
5.4 Discussion 83

6 Conclusions 85

Funding acknowledgments 89

References 91

A Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data 105

B The Functional Impact of Alternative Splicing in Cancer 115

C Systematic Analysis of Splice-Site-Creating Mutations in Cancer 129

D Susceptibility genes to breast cancer 147
D.1 Homologous recombination repair 147
D.2 Replication fork stability 147
D.3 Transcription-replication collisions 147
D.4 Mismatch repair 148
D.5 DNA damage signaling, checkpoints and cell death 148