



# From Statistical to Biological Interactions via Omics Integration

Kyrylo Bessonov

Promotor:	Prof. Dr. Dr. Kristel Van Steen
Chair:	Prof. Dr. Pierre Geurts
Internal Jury:	Prof. Dr. Vincent Bours Prof. Dr. Patrick Meyer
External Jury:	Prof. Dr. Monika Stoll Prof. Dr. Benno Schwikowski

Université de Liège  
Faculté des Sciences Appliquées  
Département d'Electricité, Electronique et Informatique

Thèse présentée en vue de l'obtention du Grade de

*Docteur en Sciences de l'Ingénieur*

Juin 2016



## Summary

The XXI century opened a new ‘Big Data’ era in which, thanks to rapid technological advancements and appearance of high throughput technologies, vast amounts of unstructured omics data (e.g., transcriptomic, genomic, etc.) are generated every day. This thesis mainly focuses on solving the problems related diverse omics data integration and interaction identification tasks. Particular attention is given to useful knowledge extraction in the context of complex diseases including pathological mechanisms with the development of software tools and pipelines. The diseases covered included glioblastoma multiforme, asthma, and ankylosing spondylitis.

Interactions detection in genomic data requires standardization of the protocols. In Chapter 3, we tested the impact of different settings in a genome-wide association interaction study (GWAIS). Some of the settings included marker selection strategy, the LD pruning, lower order effects adjustment, analytical tool. We were able to show that even small changes in each setting can have drastic impacts requiring careful assessment of proper settings and results comparisons across several analysis protocols. The greatest impact was attributed to the input dataset composition highlighting the importance of the marker selection strategy and use of prior knowledge.

Expression of genes can be affected by nearby (‘*cis*’) or distant (‘*trans*’) genotypes. Thus, we developed methodology to identify complex *trans/cis* regulatory mechanisms between expression and genotype data in the context of asthma (CAMP data). Significant overlap between ‘*trans*’ and ‘*cis*’ gene regulatory components related to immune and signaling pathways was clearly identified matching asthma disease pathology. The semi-parametric Model-Based Multifactor Dimensionality Reduction (MB-MDR) method was applied for the first time in the context eQTL study achieving low false discovery and family-wise error rates (FDR and FWER).

Identification of a meaningful data structure from omics data is a pressing topic nowadays. Gene regulatory networks (GRN) conveniently summarize large amounts of data allowing for useful knowledge generation. GRN inference is especially attractive for deciphering of complex diseases mechanisms allowing biologists to formulate a better hypothesis. We were able to generate GRNs from a single source (e.g., microarray expression data) using conditional inference forest (*CIF*)

with more attractive features compared to classical Random-Forest (*RF*) including unbiased node variable selection even in the context of highly correlated variables particularly relevant in transcriptomics. The *CIF* methods provided attractive features and performance characteristics coupled to valuable pathological insights into type 1 diabetes.

## Résumé

Le XXI<sup>e</sup> siècle a ouvert une nouvelle ère du «Big Data». Grâce aux progrès rapides et à l'apparition des technologies à haut débit, de vastes quantités de données omiques non structurées (par exemple transcriptome, génomique, etc.) sont générées chaque jour. Cette thèse s'axe principalement sur la résolution des problèmes liés à l'identification des interactions et l'intégration de divers données omiques. Une attention particulière a été accordée à l'extraction de connaissances «utiles» dans le contexte des maladies complexes, y compris les mécanismes pathologiques, ainsi qu'au développement de logiciels et de pipelines. Les maladies couvertes incluent le glioblastome multiforme, l'asthme et la spondylarthrite ankylosante.

La détection des interactions dans les données génomiques exige la standardisation du protocole. Nous avons testé l'impact des différents paramètres sur la composition des résultats finaux dans une étude d'interaction association pangénomique (GWAIS) sur l'ensemble du génome. Certains des paramètres en questions sont la sélection de la stratégie des marqueurs de sélection, le déséquilibre de liaison (LD), le faible ajustement des effets principaux et l'outil d'analyse choisi. Nous avons pu montrer que chaque paramètre pourrait avoir des effets drastiques qui nécessitent une évaluation attentive des paramètres appropriés et d'analyse comparative des résultats entre plusieurs pistes. Le plus grand impact a été attribué à la composition de l'ensemble de données lié à la stratégie de sélection des marqueurs et à l'utilisation de connaissance préalable.

L'expression des gènes pourrait être affectée par génotypes à proximité ('*cis*') ou à distance ('*trans*'). Ainsi, nous avons cherché à identifier des mécanismes mixtes *trans/cis* existants entre les données d'expression et de génotypes dans le contexte de l'asthme (données CAMP). Un chevauchement important existe entre les composants de régulation '*trans*' et '*cis*' liés au système immunitaire et à la signalisation correspondant à la pathologie de la maladie de l'asthme. La méthode semi-paramétrique Model-Based Multifactorielle Dimensionnalité Réduction (MB-MDR) a été appliqué pour la première fois dans l'étude eQTL, ce qui a permis d'atteindre un taux de faux positifs bas.

La recherche d'une structure de données significatives à partir de plusieurs sources hétérogènes de données omiques est un sujet de recherche important à l'heure actuelle. Les réseaux de régulation

des gènes (GRN) résument facilement de grandes quantités de données permettant la production de connaissances utiles. L'inférence GRN est particulièrement attrayante pour déchiffrer des mécanismes de maladies complexes permettant aux biologistes de formuler des hypothèses plus exactes. Nous avons été en mesure de produire un GRN à partir d'une seule source (par exemple, les données de biopuces d'expression) en utilisant des forêts d'inférence conditionnelle (*CIF*) avec des caractéristiques plus attrayantes par rapport à des forêts aléatoires classiques (Random Forests). Les avantages comprennent l'impartialité de sélection de variables liées à un noeud, l'impartialité même dans le contexte de variables corrélées particulièrement pertinente pour les données transcriptomique. Les *CIF* méthodes possèdent des caractéristiques attrayantes et conduisent à de bonnes performances. Ces méthodes fournissent des idées sur les mécanismes pathologiques du diabète de type 1.

## **Acknowledgements**

This thesis work would not be possible without coordinated work and support of my advisor, colleagues, collaborators and funding agencies. It was a path of self-discovery and adaptation to a new research environment and a new country, Belgium. Without their help and encouragement this thesis would not see the light of day.

I first extend my sincere thanks to promoter, Prof. Dr. Dr. Kristel Van Steen for providing me with her kind advice, insightful debates, rigorous manuscript reviews, excellent research intuition, and, above all, trust in me. I am also thankful for being able to meet and work with so many different professionals thanks to her impressive professional network. I've got exposed to challenging aspects of teamwork and life in the dry lab. I have learned new skills including administrative, presentation and teaching skills that I hope will be useful in my future life journey.

My PhD journey started in Sept 2012 in the lab of Kristel Van Steen. Everything seemed so new and unusual including language, cuisine, lifestyle and new lab mates. These were the hardest months, but thanks to the patience of my advisor and my lab-mates I was able to succeed. I especially thankful to Elena Gusareva who gave me shelter for the first few weeks.

I would never forget my lab colleagues including Ramouna Fouladi, Kridsakorn Chaichoompu, Francois Van Lishout, Francesco Gadaleta and Elena Gusareva, Jestinah Mahachie and Bärbel Maus who brightened my day and provided to me with a valuable feedback and lots of smiles. Thanks to a very friendly team, we were able not only collaborate scientifically but also to get together on various occasions and create friendships and nice moments to remember. I am also grateful to Saints Alexander Nevsky and Seraphim of Sarov Orthodox Church in Liege and especially to Oksana Konoshonkina.

I am also thankful to Pierre Geurts who provided me with valuable machine learning angle and sincere feedback on my research. I've enjoyed attending scientific events together.

Last but not least, I am grateful to my parents for their faith in my success and support.



# Contents

<b>List of Figures .....</b>	<b>vii</b>
<b>List of Tables .....</b>	<b>xiii</b>
<b>List of Abbreviations .....</b>	<b>xv</b>
<b>Chapter 1: Introduction .....</b>	<b>1</b>
1.1. Big Data .....	1
1.2. Biological systems .....	3
1.3. Interactions .....	6
1.4. Complex diseases .....	11
1.5. Integration of omics data .....	12
1.6. Aims and goals of the thesis .....	16
1.7. Reader's guide .....	18
1.8. Summary of main achievements .....	19
1.9. References .....	21
<b>Chapter 2: Background .....</b>	<b>25</b>
2.1. Different omic measurement types .....	27
2.1.1. Single Nucleotide Polymorphisms and modes of inheritance .....	27
2.1.2. Epigenetic markers .....	28
2.1.3. Gene expression measurements .....	29
2.2. Genome-wide association studies (GWAS) .....	32
2.3. Genome-wide association interaction studies (GWAIS) .....	33
2.3.1. Epishell .....	35

2.3.2.	Model-Based Multifactor Dimensionality Reduction (MB-MDR) .....	35
2.3.3.	Highlights .....	37
	<i>Recursive Partitioning</i> .....	37
	<i>Tree ensembles</i> .....	39
	<i>Random Forest</i> .....	39
	<i>Conditional Inference Forest (CIF)</i> .....	41
2.4.	Networks.....	43
2.4.1.	Network syntax .....	43
2.4.2.	Network medicine .....	45
2.4.3.	Network inference via tree ensembles .....	47
2.5.	Integration strategies.....	49
2.6.	Quality Control.....	51
2.6.1.	Genotypic Data Quality Control .....	51
	<i>Hardy-Weinberg Equilibrium (HWE)</i> .....	52
	<i>Minor Allele Frequency (MAF)</i> .....	52
	<i>Call Rate</i> .....	52
	<i>Linkage-Disequilibrium (LD)</i> .....	53
2.6.2.	Methylome Data Quality Control.....	53
2.6.3.	Expression Data Quality Control .....	53
2.7.	References .....	54
<b>Chapter 3: Genome-Genome Interactions The impact of protocol changes for genome-wide association SNP x SNP interaction .....</b>		<b>59</b>
3.1.	Chapter summary .....	61
3.2.	Introduction .....	63
3.3.	Methods .....	64

---

3.3.1. Data Quality Control .....	64
3.3.2. Additional data handling .....	66
3.3.3. Interaction testing.....	66
3.3.4. Assessing consistencies between protocols .....	67
3.3.5. Biological relevance.....	68
3.4. Results.....	69
3.4.1. Consistency between interaction results derived from different GWAIS protocols....	69
3.4.2. AS pathology relevance .....	75
3.5. Discussion.....	78
3.6. Conclusions .....	83
3.7. Acknowledgments.....	84
3.8. Chapter highlights .....	85
3.9. Appendix .....	86
3.10. References .....	88
<b>Chapter 4: Trans-eQTL epistasis protocol .....</b>	<b>91</b>
4.1. Chapter summary .....	93
4.2. Introduction .....	96
4.3. Methods .....	99
4.3.1. Data .....	99
4.3.2. eQTLs epistasis mapping.....	100
4.3.3. Controlling false positives .....	102
4.3.4. SNP to gene mapping and pathway enrichment .....	103
4.3.5. Network analysis .....	104
4.3.6. Differential network analysis .....	104
4.4. Results.....	105

---

4.4.1. Pathway enrichment .....	105
4.4.2. Control of type I error rates .....	106
4.4.3. Network analysis .....	110
4.4.4. Differential network analysis .....	110
4.5. Discussion.....	114
4.6. Conclusions .....	121
4.7. Chapter highlights .....	122
4.8. Acknowledgements and funding .....	122
4.9. Appendix .....	123
4.10. References .....	124
<b>Chapter 5: Gene expression networks .....</b>	<b>129</b>
5.1. Chapter summary .....	131
5.2. Introduction .....	134
5.3. Methods .....	136
5.3.1. Data sources.....	136
5.3.2. <i>CIT/CIF</i> -based network inference methodologies .....	138
5.4. Results.....	142
5.4.1. Evaluation of <i>CIT/CIF</i> -based GRN inference with DREAM4 data .....	142
5.4.2. Evaluation of <i>CIF<sub>mean</sub></i> -based GRN inference with DREAM2 data.....	146
5.4.3. Application of <i>CIFs</i> to DREAM5 data .....	146
5.4.4. Case Study: T1D data .....	147
5.5. Discussion.....	149
5.6. Conclusions .....	153
5.7. Chapter highlights .....	154
5.8. Acknowledgements and funding.....	154

---

5.9. Appendix .....	155
5.10. References .....	160
<b>Chapter 6: Contributed work .....</b>	<b>163</b>
6.1. Chapter summary .....	165
6.2. Identification of asthma sub-types via breath profiles analysis (patent #203-17) .....	167
6.2.1. Section summary .....	167
6.2.2. Introduction .....	168
6.2.3. Methods .....	169
6.2.4. Results .....	171
6.2.5. Discussion .....	176
6.2.6. Conclusions .....	177
6.2.7. Section highlights .....	178
6.2.8. Acknowledgments and funding .....	178
6.3. <i>Regression2Net</i> - Integration of Gene Expression and Methylation to unravel biological networks in glioblastoma patients .....	179
6.3.1. Section summary .....	179
6.3.2. Introduction .....	180
6.3.3. Method .....	181
6.3.4. Data .....	184
6.3.5. Results .....	185
6.3.5.1. Data and network characteristics .....	185
6.3.5.2. Annotation of the <i>ANDnet</i> and <i>XORnet</i> unique gene lists .....	186
6.3.5.3. Enrichment analysis .....	187
6.3.6. Discussion .....	192
6.3.7. Conclusions .....	193

---

6.3.8. Section highlights.....	194
6.3.9. Acknowledgments and funding.....	194
6.3.10. Appendix.....	194
6.3.11. References .....	194
<b>Chapter 7: General Discussion.....</b>	<b>197</b>
7.1. Introduction .....	199
7.2. Optimal GWAIS protocol.....	199
7.3. <i>Trans</i> -eQTL epistasis protocol for eQTL detection .....	201
7.4. GRN inference via trees from microarray expression data .....	205
7.5. Integration.....	207
7.6. Perspectives .....	208
7.7. References .....	209
<b>Chapter 8: CV and Publications .....</b>	<b>213</b>
8.1. Publication list (2012-2016).....	215
8.2. Curriculum Vitae.....	216

# List of Figures

<b>Figure 1.1:</b> Hierarchical organization and compartmentalization of living organisms. Arrows represent hierarchical ordering from lower to higher complexity / scale. ....	5
<b>Figure 1.2:</b> Central dogma of molecular biology and cellular compartmentalization with some of the key cellular components. ....	6
<b>Figure 1.3:</b> Biological and statistical epistasis differences. Statistical epistasis results are not always readily display 1 to 1 correspondence to biological epistasis. Legend: star (*) - mutation at particular locus; G1 – gene 1; P1 – protein 1, etc. Figure adapted from [22]. ....	10
<b>Figure 1.4:</b> Components of complex diseases: O – omics caused by changes in genomic DNA sequence, expression levels of the key marker genes, methylation profiles and others; E – environmental component including non-genomic variables (e.g., smoking), physical characteristics (e.g., weight, sex, BMI) and others. Phenotype (P) – observable characteristics of a disease helping in disease sub-phenotyping (e.g., tumor morphology, propagation rate, predominance of immune cells). ....	11
<b>Figure 1.5:</b> Homogeneous and heterogeneous data integration. Each layer represents a similarity matrix for a data source: mRNA (expression), DNA (genotype data), and methylation. “G” refers to genes (e.g., “G1” – gene 1). ....	15
<b>Figure 1.6:</b> Thesis main thesis topics areas, achievements and paper titles. Chapters 3 and 4 deal will statistical view of epistasis followed by Chapters 5,6 and 7 that also incorporate biological views of epistasis in the context of transcriptional gene regulatory networks. The Chapter keys are the main topics. ....	19
<b>Figure 2.1:</b> SNP, allele and locus definitions. DNA is represented as double helix respecting base complementarity A-T and C-G. Eye color is taken as an example of a phenotype. Genotype, a specific set of alleles (e.g., ‘Aa’) with potential impact on phenotype. ....	28
<b>Figure 2.2:</b> <i>cis</i> and <i>trans</i> gene expression regulation. Black line represents genomic DNA sequence. The orange square represents the coding region of the target get (TG) with 0 representing the beginning of the transcription start site (TSS). The 200 bp upstream (-200 bp) of the TSS the transcription binding site is located represented by green and pink rectangles. The TSS binds transcription factors (TF) modifying expression of the nearby or distant target genes. A) TG expression is regulated by nearby <i>cis</i> -regulatory sequence (green square); TG expression is regulated by TF expressed by a distant <i>trans</i> gene. B) <i>cis</i> eQTL - the locus (*) located near the protein coding sequence in yellow controls gene expression; <i>trans</i> eQTL - the locus (*) controls expression of a distant gene via <i>trans</i> -acting TF. C) epistatic <i>trans/cis</i> eQTL showing <i>trans</i> and <i>cis</i> loci (*) affecting the TG expression.. ....	31

<b>Figure 2.3:</b> Graphical summary of MB-MDR epistasis detection for a binary response (e.g., case/control). Each pair of SNPs is tested for their strength of association to the response variable (i.e. trait) summarized by a permutation-based $p$ -value. ....	36
<b>Figure 2.4:</b> Decision trees: A) structure of a typical decision tree; B) hypothetical tree applied to smokers ( $S$ ) and non-smokers ( $NS$ ) classification problem based on heterogeneous types of variables .....	40
<b>Figure 2.5:</b> Conditional inference tree node variable selection and slitting steps implemented during the tree growth.....	42
<b>Figure 2.6:</b> Graphs: key elements .....	43
<b>Figure 2.7:</b> Example of the shortest path. The shortest path between $v_i$ and $v_j$ nodes is 3 .....	45
<b>Figure 2.8:</b> Example of network inference via tree-based methods. The output is represented by $i$ th gene ( $g_i$ ) and input by all other genes except the $i$ th gene ( $g^{-i}$ ). The tree ensemble is built. The weight $w_{ij}$ measures the strength of interaction between the $i$ th and $j$ th genes ( $g_i - g_j$ ). The $w_{ij}$ value calculation depends on the chosen method and is summarized by $VIM$ . The above procedure was repeated $p$ times corresponding to the total number of genes by re-assigning a new output. A list of all gene-gene interactions is obtained with corresponding $w_{ij}$ . The pairwise gene-gene interaction list represents a gene network where $w_{ij}$ represents a weight of an edge.....	48
<b>Figure 3.1:</b> Summary of 10 GWAIS protocols included in this study and applied to AS data, the ankylosing spondylitis dataset from [17]. The number of SNPs retained at each step is shown in parenthesis. The bottom nodes refer to GWAIS protocol abbreviations and chosen parameters, following protocol components as described in [21] [8]. The abbreviations <i>additive</i> and <i>co-dominant</i> refer to SNP main effects correction encodings in MB-MDR (see [10]). The abbreviation <i>gammaMAXT</i> and <i>MAXT</i> refer to the SNP x SNP interaction significance assessment strategies implemented in MB-MDR [15] (see Methods).....	65
<b>Figure 3.2:</b> Euler diagram capturing significant SNP pairs identified in each of the 10 GWAIS protocols (Table S3.1). Each circle represents a set of the significant SNP pairs in the corresponding GWAIS protocol. Protocol numbers match the protocol referencing used in Figure 3.1. ....	70
<b>Figure 3.3:</b> Consistency between GWAIS protocols based on 207 common SNPs. Each SNP pair has a protocol-specific rank, which is stored in a protocol-specific vector. The dendrogram shows the distance between protocols, obtained via hierarchical clustering of 10 vectors (referring to the 10 GWAIS protocols included in this study) of length 207 and the Euclidean distance measure. The Euclidean distances themselves are listed in Table 3.2.....	71
<b>Figure 3.4:</b> Effect of SNP MAFs on ranked epistasis results. For each protocol, the top 1000 epistasis results are presented. Each SNP pair was ordered such that the SNP with the largest MAF was assigned to locus A, and the SNP with the lowest MAF to locus B. The numbers in red refer	

to the # of SNP pairs that were assigned to each 2-dimensional MAF bin. Protocol numbers match the protocol referencing used in Figure 3.1. .... 74

**Figure 4.1:** Definition of *cis* and *trans* SNPs with respect to the open reading frame (ORF) of an eQTL gene. ORF is the region of DNA that codes for protein and includes intron and exon sequences. A *trans* x *cis* eQTL pair refers to a SNP pair involved in a *trans* x *cis* eQTL interaction and is defined by *trans* and *cis* SNPs defined schematically by this diagram ..... 100

**Figure 4.2:** General workflow diagram of the *trans/cis* eQTL methodology. A total number of 1763 *cis* eQTLs identified in step 2 were used as “seeds” for subsequent *trans/cis* eQTL analysis. Thus, the 1364 *trans/cis* eQTLs SNP pairs contained one of the previously identified *cis* eQTLs. .... 101

**Figure 4.3:** Overlap between the “*cis*” and “*trans*” eQTL gene sets of the significant 1364 *trans/cis* eQTL SNP pairs. The “*trans*” and “*cis*” gene set refers to genes associated with the significant *trans* SNPs of the 1364 *trans/cis* eQTL SNP pairs. The numbers represent the unique gene counts. The overlap area was 2.014 % (30) of the total combined *trans* and *cis* areas. Common genes are listed in Table S4.2. .... 105

**Figure 4.4:** Overlap between “*cis*” and “*trans*” significantly enriched pathways obtained from the list of 1364 *trans/cis* eQTLs. The numbers refer to significantly enriched pathway counts. The overlap area is 18.7% (71) of total combined *trans* and *cis* areas. .... 106

**Figure 4.5:** Distribution of  $FWER_{within}$  and  $FWER_{between}$  in subplots A) and B), respectively. A) the  $FWER_{within}$  mean and medians are 0.0506 and 0.04; B) the  $FWER_{between}$  mean and medians are 0.0506 and 0.0501; The density function is shown in red.  $FWER_{within}$  and  $FWER_{between}$  see Section 4.3.3. Note we consider a false positive result any *trans/cis* loci pair with  $p$ -value  $< 0.05$ . The FWER values are computed based on complete 100 permutation-based *trans/cis* eQTL replica runs on the null data where only response variable was permuted as described in Section 4.3.3. .... 108

**Figure 4.6:** Distribution of false discovery rate (FDR) per 100 permutation replicas each containing 1763 *trans/cis* eQTL runs. The FDR was defined as number of false positives within each replica permutation run containing approximately 877,615 *trans/cis* loci pairs. FDR per replica run is defined in Section 4.3.3. The false positive result is any *trans/cis* loci pair with  $p$ -value  $< 0.05$ . The red line represents the density function. .... 109

**Figure 4.7:** Distribution of the Pearson correlation values of all unique *cis* eQTL gene pairs (1,554,084). The mean and median was 0.0104 and 0 with standard deviation of 0.21. .... 109

**Figure 4.8:** Directed *trans/cis* eQTL network  $G$  composed of 1459 nodes. Nodes with degree  $\geq 2$  are shown. No cliques containing more than 2 nodes were found. Nodes with degree  $\geq 8$  are shown in red while nodes with degree  $\geq 2$  but  $< 8$  are shown in orange. The node names correspond to gene symbols representing SNPs mapped to the nearest genes (see Methods). .... 112

<b>Figure 4.9:</b> Total degree distribution for the <i>trans/cis</i> eQTL network $G$ . The total node degrees represent the sum of the “in” and “out” degrees. The associated genes of the selected highest degree nodes are indicated. The total degree distribution is also available in online Table S4.4. ....	113
<b>Figure 4.10:</b> Differential undirected network $G_D$ built from $G_{NS}$ and $G_S$ networks. The $G_D$ highlights the three nodes <i>LEPR</i> , <i>BCL11A</i> , <i>PPP1R13L</i> , members of the unique largest clique of size 3. No other cliques were present in the $G_D$ . The node names correspond to gene symbols. ....	113
<b>Figure S4.1:</b> The <u>complete</u> weighted directed gene network $G$ built using the list of 1364 significant <i>trans/cis</i> eQTLs complimenting Figure 4.8. (See the online supplement) .....	123
<b>Figure S4.2:</b> The <u>complete</u> weighted directed gene network $G_{NS}$ built using the list of 1552 significant <i>trans/cis</i> eQTLs complimenting Figure 4.10. (See the online supplement) .....	123
<b>Figure S4.3:</b> The <u>complete</u> weighted directed gene network $G_S$ built using the list of 707 significant <i>trans/cis</i> eQTLs complimenting the differential network $G_D$ shown in Figure 4.10 (See the online supplement) .....	123
<b>Figure 5.1:</b> Gene regulatory network framework based on <i>CIT /CIF</i> , adapted from [17,30] ....	137
<b>Figure 5.2:</b> DREAM4 performance results – $mtry=k/3$ . <i>AUROC</i> and <i>AUPR</i> expressed performance of considered GRN inference methodologies for each of the 5 DREAM4 networks included in the study and described in the methods Section 5.3.2. Table S5.1 complements this figure with specific <i>AUROC</i> and <i>AUPR</i> values.....	143
<b>Figure 5.3:</b> DREAM4 performance results – variable $mtry$ . The performance of the $CIF_{mean}$ methods at various $mtry$ values assessed via the DREAM4 overall score. Overall scores are averages over 5 networks. ....	144
<b>Figure 5.4:</b> DREAM2 performance results – variable $mtry$ . a) The performance of the $CIF_{mean}$ methods based on the total area of <i>AUROC</i> and <i>AUPR</i> . b) A more detailed view of the <i>AUROC</i> and <i>AUPR</i> dynamics as a function of the $mtry$ parameter. Table S5.2 complements this figure with specific <i>AUROC</i> and <i>AUPR</i> values.....	145
<b>Figure 5.5:</b> DREAM5 performance results - $mtry=k/3$ . The GRN inference performance levels across $CIF_{mean}$ methodologies. Performance is quantified via the DREAM5 overall score as defined in for instance [26]. Table S5.3 complements this figure. ....	146
<b>Figure 5.6:</b> DREAM5 performance results – variable $mtry$ . The performance of the $CIF_{mean}$ methods at various $mtry$ values were assessed based on DREAM5 overall score averaged over 3 DREAM5 networks. ....	147
<b>Figure 5.7:</b> The T1D Case study performance results – variable $mtry$ . a) Performance of the $CIF_{mean}$ methods based on the <i>AUROC</i> and <i>AUPR</i> . b) A more detailed view of the <i>AUROC</i> and <i>AUPR</i> dynamics as a function of the $mtry$ parameter. As the gold standard the verified set of TF/TG sets from [27].....	148

<b>Figure S5.1:</b> DREAM 4 performance results - $mtry=k/3$ . The GRN inference performance levels across the 8 methodologies described in methods section. Performance is quantified via the DREAM4 overall score as defined in for instance [26].	155
<b>Figure S5.2:</b> DREAM2 performance results - $mtry=k/3$ . The performance of the $CIF_{mean}$ and $RF$ methods based on the total area of $AUROC$ and $AUPR$ using the default settings with the $mtry=k/3$ .	155
<b>Figure S5.3:</b> DREAM5 performance results - $mtry=k/3$ showing $AUROC$ and $AUPR$ per each network.	156
<b>Figure S5.4:</b> T1D Case study performance results - $mtry=k/3$ . The performance of the $CIF_{mean}$ methods based on the total area of $AUROC$ and $AUPR$ using the default settings with the $mtry=k/3$ .	157
<b>Figure S5.5:</b> DREAM4 GS networks. The DREAM4 GS networks size 100 from 1 to 5 along with the basic network measures.	158
<b>Figure 6.2.1:</b> variable importance measure ( $VIM$ ) in $E/N$ , $E/P$ and $N/P$ binary sub-asthma classification scenarios calculated from conditional inference forests ( $CIFs$ ). Legend: $E$ – eosinophilic, $N$ – neutrophilic, $P$ – paucigranulocytic asthma. For chemical identities of highlighted VOCs please refer to Table 6.2.1	172
<b>Figure 6.2.2:</b> $AUROC$ and $AUPR$ curves from the $CIFs$ binary classification of the $N/P$ , $E/P$ and $N/P$ scenarios where $E$ -eosinophilic, $N$ - neutrophilic and $P$ -paucigranulocytic asthma sub-types. The red line diagonal line indicates random guess – the minimal classifier threshold. Table 6.2.2 provides areas under the curves.	173
<b>Figure 6.2.3:</b> VOCs box plots across three sub-types of asthma (Eosinophilic, Neutrophilic and Paucigranulocytic). The red line indicates mean area under the peak and ‘eos’, ‘neutro’ and ‘pauci’ are asthma types.	174
<b>Figure 6.2.4:</b> Accuracy and precision of each individual VOC under $N/P$ , $E/P$ and $N/P$ scenarios where $E$ - eosinophilic, $N$ - neutrophilic and $P$ -paucigranulocytic asthma sub-types. Table 6.2.1 provides VOC number to chemical name conversion.	176
<b>Figure 6.3.1:</b> General workflow diagram of the <i>Regression2Net</i> methodology consisting of 4 stages.	181
<b>Figure 6.3.2:</b> Total degree distribution of the <i>ANDnet</i> network of edges present in both <i>EEnet</i> and <i>EMnet</i> , Node index represents the node number	185
<b>Figure 6.3.3:</b> Total degree distribution of the <i>XORnet</i> network of edges present in <i>EMnet</i> but not in <i>EEnet</i>	186
<b>Figure 6.3.4:</b> <i>ANDnet</i> network overlap with the significant pathways. The highlighted genes belong to the significant pathways indicated in Table 6.3.1 while non-highlighted (white) genes	

have not been linked to any significant pathway. The size of each node is determined by betweenness, defined as the number of shortest paths going through the node .....	188
---	-----

# List of Tables

<b>Table 1.1:</b> Information flows suggested by the central dogma of molecular biology .....	5
<b>Table 1.2:</b> Interaction examples .....	8
<b>Table 2.1:</b> the main types of biological networks .....	46
<b>Table 3.1:</b> Most significant SNP pairs across 10 adopted GWAIS analysis protocols. All <i>p</i> -values are multiple testing corrected, either Bonferroni-based (BOOST protocols) or re-sampling based (MB-MDR protocols). .....	72
<b>Table 3.2:</b> Significant pairs containing one of the 49 SNPs associated to main effects [17], obtained via the 10 GWAIS protocols. ....	73
<b>Table 3.3:</b> Statistically significant SNP x SNP interactions that contain a SNP occurring in at least one of 102 SNP pairs listed in Supplementary Table 5 in Evans <i>et al.</i> [17]*. ....	76
<b>Table 3.4:</b> Top 10 Significant GO terms related to top 1000 SNP pairs per GWAIS protocol, based on Fisher's combined <i>p</i> -value at a significance level of 0.05. Protocol-specific <i>p</i> -values are also reported. ....	77
<b>Table S3.1:</b> Parameters used to run ten GWAI protocols.....	86
<b>Table S3.2:</b> Euclidean distances amongst GWAI protocols (ref. to Figure 3.1).....	87
<b>Table S3.3:</b> Significant SNP pairs with multiple testing adjusted <i>p</i> -values (<0.05). (See the online supplement) .....	87
<b>Table S3.4:</b> List of common 207 SNP pairs amongst 10 GWAI protocols findings (including significant and non-significant SNP pairs). (See the online supplement). ....	87
<b>Table S3.5:</b> Annotated Evans' <i>et al.</i> (2011) 38 SNP pairs out of 102 listed in Supplementary Table 5 of [17]. These pairs contain one SNP (in bold) that was present amongst the significant findings of the 10 GWAI protocols. (See the online supplement) .....	87
<b>Table S3.6:</b> Significant GO terms related to top 1000 SNP pairs per GWAI protocol, based on Fisher's combined <i>p</i> -value at a significance level of 0.05. Protocol-specific <i>p</i> -values are also reported. (See the online supplement) .....	88
<b>Table 4.1:</b> Top 20 significantly enriched Reactome pathways in <i>trans</i> and <i>cis</i> sets from 1364 <i>trans/cis</i> eQTLs.....	111
<b>Table S4.1:</b> The complete list of significant <i>trans/cis</i> eQTLs. (See the online supplement) .....	123
<b>Table S4.2:</b> The list of 30 common genes of the <i>trans/cis</i> and <i>cis</i> eQTL genes sets. See online supplement. (See the online supplement) .....	123

---

<b>Table S4.3:</b> The 71 common significantly enriched pathways between <i>trans/cis</i> and <i>cis</i> eQTL genes sets. (See the online supplement) .....	123
<b>Table S4.4:</b> Total degree distribution of nodes of the <i>trans/cis</i> eQTL network $G$ (see Figure 4.5). (See the online supplement) .....	123
<b>Table S4.5:</b> The common list of nodes between $G_{NS}$ and $G_S$ networks with the corresponding total degrees. (See the online supplement).....	123
<b>Table S4.6:</b> The 3 significant epistatic <i>trans/cis</i> eQTL pairs both at Bonferroni and $M_{eff}$ method [59] thresholds .....	123
<b>Table 5.1:</b> Data characteristics .....	138
<b>Table 5.2:</b> Runtime estimates of the family of CIF methods on a single CPU .....	144
<b>Table 5.3:</b> $CIF_{mean}$ $p$ -value ( <i>Monte Carlo</i> ) significant pairs from the T1D dataset .....	149
<b>Table S5.1:</b> DREAM4 methodology rankings - default settings.....	159
<b>Table S5.2:</b> DREAM2 methodology rankings - default settings .....	159
<b>Table S5.3:</b> DREAM5 methodology rankings - default settings .....	159
<b>Table 6.2.1:</b> VOC mappings to compound names .....	172
<b>Table 6.2.2:</b> $CIFs$ classifier performance in asthma type binary classification .....	175
<b>Table 6.2.3:</b> Single VOC classification rules based on area units (AU).....	175
<b>Table 6.3.1:</b> $ANDnet$ enriched pathways .....	189
<b>Table 6.3.2:</b> $XORnet$ enriched pathways.....	189
<b>Table 6.3.3:</b> $INTnet$ enriched pathways .....	191
<b>Table S6.3.1:</b> $ANDnet$ 284 gene annotations. (See the online supplement). .....	194
<b>Table S6.3.2:</b> $XORnet$ 447 gene annotations. (See the online supplement). .....	194

# List of Abbreviations

*MB-MDR* – Model-Based Multifactor Dimensionality Reduction methodology

*MBMDR* – Model-Based Multifactor Dimensionality software tool

*NGS* – Next Generation Sequencing

*miRNA* – micro RNA

*RNAseq* – mRNA sequencing

*BD2K* – *Big Data* to Knowledge

*GWAS* - genome-wide association study

*GWAIS* – genome-wide association interaction study

*GRN* – gene regulatory network

*eQTL* - expression quantitative trait locus

*SNF* – similarity network fusion

*CIF* – conditional inference forest

*RF* – random forest

*QC* – quality control

*LD* – linkage disequilibrium

*AS* - ankylosing spondylitis

*GBM* - glioblastoma multiforme

*GRN* – gene regulatory network

*VIM* – variable importance measure

*DT* – decision trees

*Big Data* - massive amounts of raw domain-specific information



# **Chapter 1: Introduction**



# 1. Introduction

The XXI century is marked by massive amounts of high-dimensional, heterogeneous and complex data [1-3]. The rapid technological advancements in IT and biology-related fields made it relatively easy to autonomously acquire, store, transmit and analyze massive amounts of unstructured raw data presenting new challenges for data processing and visualization [4].

This introductory section provides the general context and terminology associated with the biological data handling and integration. This thesis first introduces the complexity of biological datasets from the interactions point of view followed by the discussion of different omics data types in the context of complex diseases. This Chapter will be concluded by introduction of the central topic linked to omics data integration.

## 1.1. Big Data

Nowadays many biological fields including genomics, proteomics, molecular biology, statistical genetics and others are witnessing arrival of *Big Data* era. Big Data term has several definitions, but in this thesis we define it as an act of collecting vast amounts of data exceeding the processing capacity of conventional systems [3]. Velocity, volume, variety, variability and complexity are the main characteristics of *Big Data* present not only in biology-related fields, but also in many others some of which include social, financial, business and meteorology sciences [5]. Big corporations, such as Facebook and Google, are constantly collecting and analyzing vast amounts of client's *Big Data* to better understand social behaviour, interaction patterns [6], reduce costs, make smarter decisions and to improve personalized recommendation systems (e.g., Netflix, Amazon.com) [7].

Along with many new exciting opportunities mentioned above, *Big Data* poses new challenges to statisticians and bioinformaticians who need to deal with massive sample sizes, high-dimensionality of data, data heterogeneity, scalability issues, transmission and storage bottlenecks [1].

The data “explosion” is currently witnessed by biological data repositories such as the European Bioinformatics Institute (EBI) in Hinxton that, as stated by the 2014 report, stored 2.58 petabytes of sequencing data (2,580 TB) [8]. Next Generation Sequencing (NGS) of one human DNA sample with 30x coverage requires 310 gigabytes of total data storage [9]. According to recent reports, the EBI’s sequencing data doubles in size every year [4]. The Beijing Genomics Institute (BGI) that generates 6 terabytes of genomic data every day [4]. New compression algorithms and cloud services may provide a promising solution to these problems some of which include storage and computational power. Acquisition and storage of raw and under-analyzed data are of little use to society and science requiring concrete solutions and efficient Big Data to Knowledge (BD2K) analysis pipelines. Lawrence E. Hunter from the University of Colorado states that “getting the most out of the data requires all relevant prior knowledge”[4]. The hypothesis-driven studies can greatly reduce the Big Data analysis obstacles by providing specific targets and hypotheses. Prior knowledge most often comes from multiple sources including, for example, protein-protein, expression, metabolic, clinical data. The hypothesis-free studies are more challenging compared to hypothesis-driven ones as the search space can be vast. Therefore, there is a strong need in intelligent and computationally efficient algorithms and methodologies to integrate and extract useful transferable practical knowledge in a given domain. The data integration will be further discussed in detail in the Sections 1.5 and 2.5.

Fortunately, there is a large global investment and efforts in order to mitigate and address the pressing issues imposed by ‘Big Data’. The key efforts concentrate in increasing accessibility and openness in scientific research. European Union had invested in the European life-sciences Infrastructure for biological Information (ELIXIR) project [10] aiming at providing access to large biological datasets and computational facilities. Increased accessibility brings obvious benefits allowing researchers to compare in-lab generated results on a global scale. For example, results obtained on one type of cancer can be compared and contrasted with the other types strengthening and possibly adjusting the original hypothesis(es). For example, several integrative studies comparing data patterns across several cancers and tissues are being completed using The Cancer Genome Atlas (TCGA) resource [11-13]. In addition, provision of the computational capabilities and IT infrastructure by ELIXIR via cloud services and field experts allows labs with limited resources and technical skills to contribute meaningfully to large-scale complex research projects.

The unprecedented availability of large biological data combined with ever-improving communications technologies of the XXI century is changing scientific practices and bringing new opportunities and challenges. One of them is the development of effective BD2K pipelines implementing creative and computationally efficient solutions addressing the data integration needs of heterogeneous and complex biological data. In the ever-increasing effort of “making sense” of Big Data, sharing of expertise, valuable ideas and “know-how” from different domains is essential, yet there are significant accessibility barriers for less tech-savvy users. Laboratories can readily generate data but face significant limitations implementing the BD2K data processing pipelines and methodologies. Galaxy platform [14] provides attractive GUI and proper documentation allowing users to implement and to share their data processing pipelines more efficiently while avoiding complexities associated with software installation and computational requirements. Another collaborative BD2K project related to genomic material sequencing is EasyGenomics<sup>TM</sup> [15] developed by BGI. The platform offers users the ability to easily access and to process large amounts of NGS, exome, RNA-*seq*, miRNA and other types of data along with already developed bioinformatics workflows such *de novo* genome assembly tools among others.

In the next section, we describe properties of biological data focusing over interactions and interdependencies between variables and data sources.

## **1.2. Biological systems**

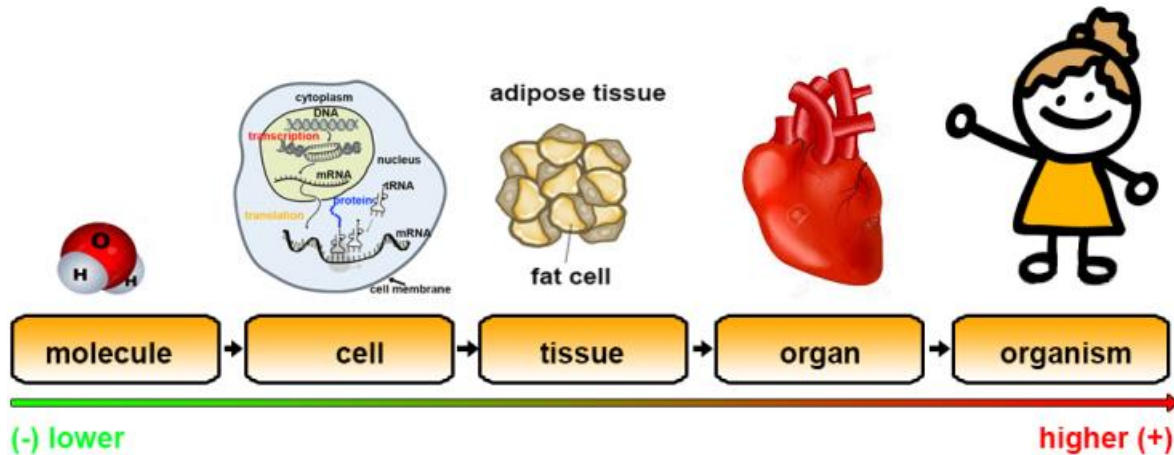
Interactions are universal amongst living organisms. They are typically complex and are dependent on many hidden and observable variables. By permeating almost every aspect of our everyday life, they can be found in the events such as an increase in global temperature, extinction of particular species, changes in the quality of life and lifestyle habits. Living organisms can be characterized by strong inter-dependency, hierarchy and constant need of interactions (e.g., communications). Effective interaction through communication and “emotional intelligence” can significantly increase productivity and emotional well-being of the whole group. In the clinical context, it was shown that effective communication and interaction between cancer patients, nurses, and doctors can significantly decrease emotional distress, improve emotional well-being, increase accuracy and completeness of medical data [16].

Interconnectedness is the central concept in systems biology implying that all system components are interconnected at some organizational level. Since biological systems are highly organized, the notion of hierarchy is essential as it includes allocation of roles and tasks subdivision creating smaller interconnected components. The interaction between system components assumes circulation of vital information allowing an orchestrated functioning of the whole system as one unit. Due to ever-changing conditions, biological systems are also dynamic and adaptive relying on rapid communication between its components which can be further logically subdivided into information layers represented by biological processes. Each organizational level interacts with the other in organized and controlled manner to maintain the necessary equilibrium and well-being (i.e. homeostasis) of the whole system (i.e. organism). The typical homeostatic response to stress involves changes in biological networks involving cellular constituents. Changes in internal network wiring of any biological system impact the functioning of its cellular components. These components often include mRNA, proteins (enzymes, TFs), miRNA and many others. Thus, at each organizational level, there is a complex network of inter and intra interactions such as cell-cell, molecule-cell and many others allowing bi-directional flow of information. These concepts are further developed in subsequent paragraphs starting with biological interactions (Section 1.3).

For biological systems, interactions are essential as they underline the basic survival and homeostasis principles. The essential components of any living organism are hierarchically organized in modules from lower to higher complexity as shown in Figure 1.1. The modularity in organization of living organisms increases their adaptability to external environment. For example, cancer cells identified in a particular tissue can be isolated and removed without disruption of other vital biological processes thanks to modularity design. Thus, any biological system can be understood at different levels of abstraction.

Biological systems are dynamic due to ever-changing environment and adaptation stimuli inherent to living organisms. One example of a above-mentioned dynamic information flow at the cellular level is addressed by “the central dogma of molecular biology”, introduced by Francis Crick in 1956 [17]. According to this “central dogma”, the DNA, mostly located inside the nucleus (Figure 1.2), is translated to mRNA, which is subsequently exported to cytoplasm, where it is finally translated to amino acids – the building blocks of any protein (Table 1.1) . Gene expression refers

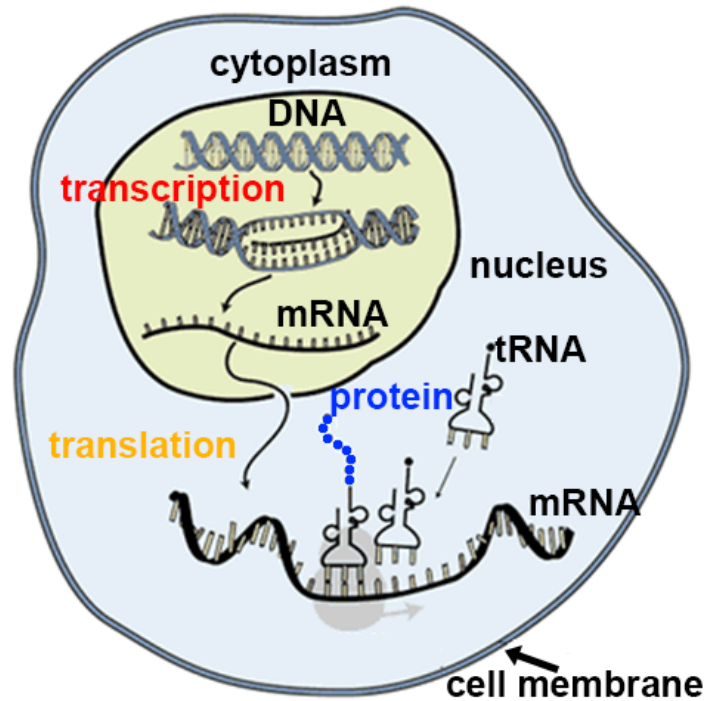
to genes being transcribed into RNA (DNA  $\rightarrow$  RNA). It is a very complex process and involves many intermediate steps. Some of the gene expression control mechanisms are discussed in the Section 2.1.3.



**Figure 1.1:** Hierarchical organization and compartmentalization of living organisms. Arrows represent hierarchical ordering from lower to higher complexity / scale.

**Table 1.1:** Information flows suggested by the central dogma of molecular biology

Flow of information	Process Name	Context
DNA $\rightarrow$ DNA	Replication	Copy of genetic material during cell replication (mitosis)
DNA $\rightarrow$ RNA	Transcription	mRNA production
RNA $\rightarrow$ protein	Translation	Protein production



**Figure 1.2:** Central dogma of molecular biology and cellular compartmentalization with some of the key cellular components.

At the cell level, biological processes are compartmentalized into several compartments (i.e. organelles) the most important of which are nucleus, endoplasmic reticulum, Golgi, mitochondria and others (Figure 1.2). Each organelle has its unique function such as nucleus which stores cellular genome (i.e. DNA). Information flow between cellular entities (genes, proteins, metabolites, etc.) and system components is not isolated and involves interactions. In the next Section 1.3, we explain what is meant by “interactions” in different contexts.

### 1.3. Interactions

Some of the previously referenced relationships involve biological interactions. Historically biologists utilized reductionist approach by focusing on a single cellular component (e.g., protein, organelle, and metabolite). Advances in network biology showed that this approach is oversimplification as biological functions are rarely governed by a single compound, but, instead, involve numerous inter and intracellular interactions. Some of such 2-way interactions include

protein-protein, protein-DNA, protein-metabolite, transcription factor – target gene (TF-TG) and many others.

In 1998 Bruce Albers and Andres Murray highlighted limitations of the reductionist approach and suggested to study cellular components as functional groups and modules [18]. This higher order view was slowly adopted by systems biology researchers. The current research practices readily utilize this new modular hierarchical paradigm to identify and further study protein complexes instead of single proteins. These studies proved a presence of higher order interactions in biological systems spanning further than a single cellular membrane [19]. Future efforts in systems biology related fields focused on developing methodologies to deal with ever-increasing complexities of the datasets. Current studies focus on inter-species interactions, group genes into functional groups (i.e. modules), involve host-pathogen-drug interactions and many others.

The interactions considered in this thesis are the ones shown in Table 1.2. These interactions are quite diverse in nature and span fields of statistical genetics, molecular biology and biochemistry amongst others. The transcript-transcript and protein-protein interactions have a physical interpretation that is either reflected in mRNA levels or in creation of protein complexes. On the other hand, under the context of complex diseases gene-gene ( $G \times G$ ) and gene-environment ( $G \times E$ ) interactions are described by statistical models that might not be readily interpretable and provide a clear biological or clinical meaningless [20]. Thus, it is important to both interpret interactions from both perspectives and verify via wet-lab experiments.

Protein-protein interactions are usually identified with help of protein chips and yeast two-hybrid (Y2H) screens. Protein chips provide information on protein quantity while the Y2H screens tell whether two proteins physically bind to each other and form a complex. The gene-gene interactions are often monitored via mRNA levels with help of microarrays and, most recently, via exome sequencing - RNA-seq. One type of gene-gene interaction involves the transcription factor (TF) - target gene (TG). Practically, these interactions are identified via chromatin immunoprecipitation (ChIP) technology that locates genome regions bound by transcription factor – transcription factor binding site (TFBS). The TFBS sequence can identify genes potentially targeted by transcription factor via analysis of the 5'-untranslated regions of the potential target genes [21]. The statistical

protein-protein interactions can be identified via simple correlation measures but are not in 1 to 1 correspondence with biological protein-protein interactions [22-24].

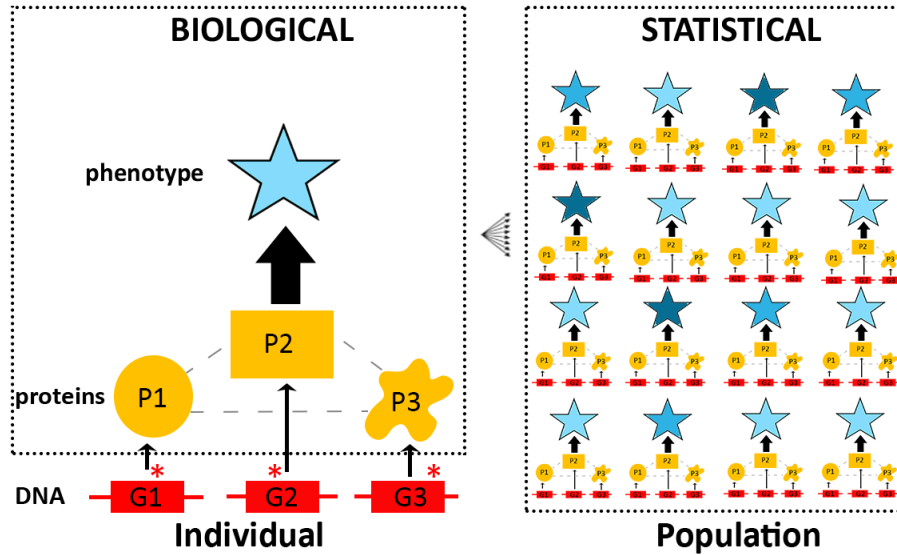
**Table 1.2:** Interaction examples

Interaction	Source	Data type	Technology
gene – gene ( $G \times G$ )	mRNA	expression	microarrays, RNA- <i>seq</i>
protein-protein	protein	expression	protein arrays
gene-environment ( $G \times E$ )	mutations / environment	genotypes / clinical	SNP arrays
gene-gene ( $SNP \times SNP$ )	mutations	genotypes	SNP arrays

Networks or graphs naturally represent pairwise interactions. Graphs are mathematical structures used to represent information. The strength of graphs lies in their visual aspect allowing to capture entire set of interactions. Graph theory was introduced and developed by mathematicians. In 1736 Leonid Euler introduced the concept of a graph via Königsberg Bridge problem. Since then, graph problems were continuously studied. In the XXI century, the most well-known graph is World Wide Web (WWW) linking thousands of web pages and resources on a global scale. Computer scientists extensively study the WWW which became a primary global communication medium. Graphs are also heavily used in sociology under the context of social network analysis (SNA). The most famous contribution of SNA was the “*Small World*” phenomenon introduced by Stanley Milgram in 1967 which states that an average path to connect two strangers is surprisingly very low – on average, only 6 degrees of separation [25]. The SNA further developed new analysis tools and methods [26]. For example, collaboration graphs with signed edges representing friendship/hatred, respectively, are used to predict dynamics of a complex set of relationships in a company. Analysis of non-conserved content spread (i.e. information flow) through a social network is used in businesses administration and, even, in epidemiology fields. Recent epidemiological study predicted the spread of HIV via the human contact patterns/socialization habits [27]. These examples from different contexts highlight a convenient generalization nature of graphs. The next paragraphs will briefly introduce graph application under biological contexts.

Depending on the biological context and involved entities (e.g., proteins, genes, metabolites) there are several types of biological networks. The most widely known graphs are protein-protein networks that describe physical interactions between proteins. Another kind of graphs are gene regulatory networks (GRNs) which describe gene-gene regulatory mechanisms. Some proteins exhibit regulatory function (i.e. TFs) impacting target gene mRNA levels. GRNs describe such transcription factor  $\rightarrow$  target gene (TF  $\rightarrow$  TG) directional interactions where TG is a gene whose expression levels are dependent on TF. The GRNs have directional edges as the reverse direction TG  $\rightarrow$  TF is not possible since the TG alone can not impact the TF mRNA levels unless the TG also acts as TF with respect to other TG. Finally, gene-gene  $G \times G$  and gene-environment  $G \times E$  interactions in the context of complex diseases can be described via epistatic networks defined in Section 2.4.2.

In molecular and human genetics, interactions can contribute to *epistasis* (from Greek *epistasis* - the act of stopping) which is defined as the interaction between different genes under biological perspective [28]. The term epistasis has many conflicting meanings often lacking precise definition [28]. Under classical statistical view that follows Fisher's definition [29], epistasis can be viewed as an interaction between a two loci,  $X_1$  and  $X_2$ , resulting in a non-additive contribution to a phenotype/trait ( $Y$ ). Epistasis from the biological point of view refers to gene-gene interaction among biomolecules (mRNA, protein, etc.) in which the phenotypic effect of one gene is being modified by the other. This results in a “*joint*” phenotype incorporating effects of both genes (e.g., eye color, metabolic reaction rates, etc.). Another definition of biological epistasis is a departure from Mendel's Laws of inheritance in which a pair of epistatic loci results in different from the expected phenotypic 9:3:3:1 ratio (in the case of non-interacting loci). Although to the present date there exists a limited number of studies on biological epistasis, it is evident that epistasis ubiquity underlines complex diseases. Hinkley *et al.* shown that by considering epistatic effects between 200 loci associated to *HIV-1* drug resistance, the predictive power of the viral replicative capacity improves by the 18.3% [30]. It is important to realize that there is no direct 1:1 correspondence between statistical and biological epistasis occurring at the population and individual levels, respectively [20] (Figure 1.3). The key challenge is to bridge the gap existing between these two definitions by the development of novel methodologies supported by the wet-lab experimental data.



**Figure 1.3:** Biological and statistical epistasis differences. Statistical epistasis results are not always readily display 1 to 1 correspondence to biological epistasis. **Legend:** star (\*) - mutation at particular locus; **G1** – gene 1; **P1** – protein 1, etc. Figure adapted from [22].

Proper definitions of linear and non-linear statistical epistatic models are particularly relevant in the context of complex diseases where higher-order complex multicollinearity (i.e. correlation) patterns within and between biological layers are present (see Section 1.5). It is believed that in such scenarios non-linear and non-parametric models can better capture the ‘true’ nature of correlated and interdependent biological interactions as highlighted by a large number of systems biology and functional genomics studies [31-34].

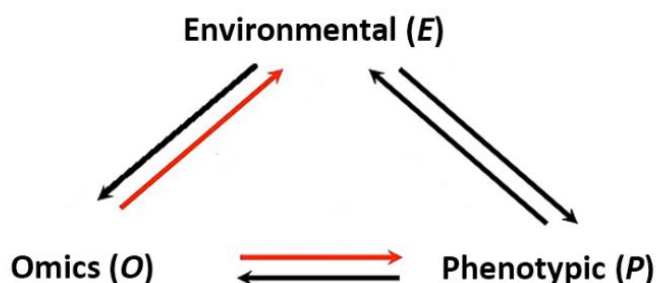
Complex diseases, caused by a combination of genetic and environmental factors, assume interactions between genes. Elucidation of statistical epistasis networks highlighting  $G \times G$  and  $G \times E$  interactions is an area of an active research aiming at better understanding of a genetic architecture of complex diseases. Some of the first efforts in the elucidation of statistical epistasis networks [24] in bladder cancer were able to capture the global context of gene-gene interactions allowing to better understand the pathology of the disease [24]. Further explanation of epistasis in the context of Genome-Wide Interaction Studies (GWAIS) is provided in Section 2.3.

## 1.4. Complex diseases

What is understood by a ‘complex disease’? Strictly speaking, complex diseases are those that may result from a mixture of genetic and environmental and lifestyle factors (Figure 1.4). Complex diseases do not follow classical Mendelian patterns of inheritance. Most of these factors and causal mechanisms remain to be identified.

Often the genetic component in these diseases is minor highlighting importance of environment and lifestyle factors. Unfortunately, these factors are highly variable from one individual to another. In addition, the response to environmental factors greatly varies across population requiring population stratification adjustments. Some examples of a complex disease include ankylosing spondylitis, Alzheimer's disease, glioblastoma cancer, asthma, Crohn's disease and many others.

Rather than studying each component separately (Figure 1.4), the complex diseases need to be addressed under integrative holistic context taking into account interactions existing between all three constituents show in Figure 1.4. Specifically, this thesis will focus on the omics component composed of transcriptomics, genomics, methylomics, as well as, other data types. Also, will consider phenotypic and environmental components of complex diseases such as asthma and ankylosing spondylitis.



**Figure 1.4:** Components of complex diseases: **O** – omics caused by changes in genomic DNA sequence, expression levels of the key marker genes, methylation profiles and others; **E** – environmental component including non-genomic variables (e.g., smoking), physical characteristics (e.g., weight, sex, BMI) and others. Phenotype (**P**) – observable characteristics of a disease helping in disease sub-phenotyping (e.g., tumor morphology, propagation rate, predominance of immune cells).

## 1.5. Integration of omics data

Integration means different things to different researchers due to the lack of a solid conceptual framework [35]. Some interpret data integration as elimination of redundancies in data; others use additional data sources to validate findings; others, meta-analytic efforts to aggregate results across datasets. In this section, we will provide partial systematization and survey of different integration strategies used by the community. Statistical and biological epistasis under the context of omics data integration is defined in [36]. Perhaps the most general definition of data integration can be formulated as an integration of systems components via relationship and mathematical models [37,38]. In our case ‘system’ can be considered a biological system – an organism. More generally data integration can be also defined as combination and aggregation of various data sources in order to provide a unified view of the data. Indeed, one of the purposes of integration is to provide a unified and common view to heterogeneous data. In addition, integration naturally complements current systematic view of biological systems discussed in previous Sections 1.1-1.4. The reductionist view decomposes the complex system into its components obtaining a catalogue of elements; meanwhile, integrated view tries to integrate those components providing a bigger and more truthful picture of the system. The integrated view tries to understand relationships existing between data sources and principles governing those. The main goal of data integration approaches is to increase the posterior data utilization via an efficient BD2K integrative pipeline. Another support towards the integrated view of biological systems and its components is the presence of synergy [39]. Synergy dictates that the whole is greater than the sum of its parts. The new properties and functional understanding of the whole system can be achieved by considering additive effects of its individual parts. For example, Takahashi *et al.* systematically studied and successfully identified combinations of transcription factors required to reprogram a somatic cell to undifferentiated pluripotent cell [40]. Their success would be limited if they only considered each transcription factor in isolation neglecting additive effects. Thus, a full understanding of complex biological systems can be only done considering all known layers of omics data ultimately having a synergistic impact on the phenotype. A biological system can be decoupled into several components that can include genome, epigenome, transcriptome, proteome, metabolome, phenome, and others. The inferred model ideally should not only consider interactions between individual elements within a given data layer, but also interactions occurring between layers under integrative data context. Data integration is a fuzzy term, but in this thesis, it refers to

methodologies that effectively collect evidence from different omics layers (i.e. data sources) resulting in either one integrative model or several models with a posterior aggregation step. These data integration strategies are termed meta-dimensional analysis and multi-staged analysis respectively [41] and are further discussed in Section 2.5. The integration term should not be confused with data fusion mentioned in Section 2.5.

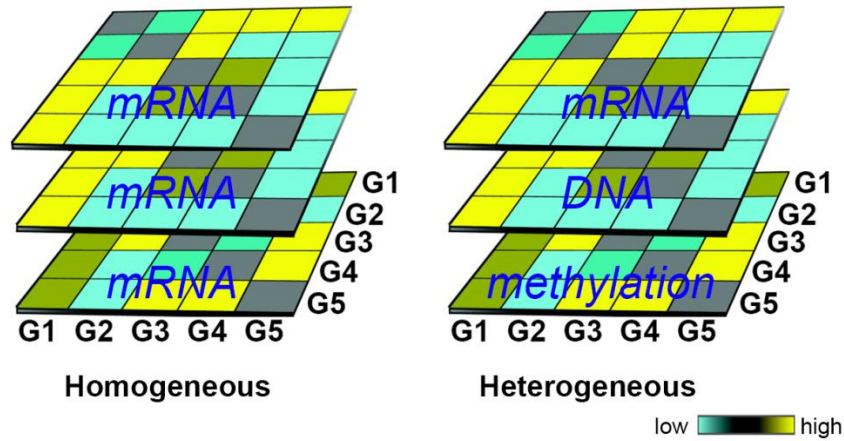
One type of relationship is between genotypes and phenotypes (Figure 1.4). Genome-wide association studies (GWAS) are primary tools to analyze complex diseases. On a genome-wide scale, they assess genetic variation in a population in relation to a given phenotype/trait (e.g., disease status). In other words, GWAS measures the strength of association between a single locus and phenotype (i.e. main effect). To account for higher order interactions, the genome-wide association interaction studies (GWAIS) provide a better alternative to classical GWAS thanks to a more realistic disease model which takes into account gene-gene (*GxG*) and gene-environment (*GxE*) epistatic interactions. In the case of the *GxG* GWAIS, the association between phenotype (i.e. trait) and two loci (i.e. a SNP pair) is being quantified while, in the case of the *GxE* GWAIS, the environmental variable (e.g., sex, age, etc.) together with a single locus is taken into account (Figure 1.4). The main challenges associated with GWAIS include small sample sizes, generally “weak” epistatic signals and high statistical significance threshold requirements ( $p\text{-value} < 10^{-13}$ ) due to an enormous number of hypotheses. One needs to consider all SNP pairs in datasets often containing more than 1 million of markers.

Interplay of many system components and processes results in observable phenotypes such as a disease. Since a single network can conveniently represent a large number of interactions, network biology is trying to systematize and better understand the interplay between each of the system components including cell-cell, protein-protein, drug-protein, drug-patient and many others. The central role and main challenge of network biology based approaches is to understand the impact and dynamics between cell components interactions [42]. Greater understanding of interaction dynamics between DNA, RNA, proteins and small molecules greatly improves characterization and understanding of living organism dynamics under the variety of contexts. The gained knowledge can be used in the production of drugs, more accurate patient classification and improved disease risk assessments among other things. An increased understanding of inter-

dependency and dynamics of biological systems has many benefits including improved quality of life, better disease control and management. Unfortunately, biological systems are very complex and the research breakthroughs take decades to happen. Nevertheless, with rapidly improving technology advancements in many essential fields, new break-throughs are likely to occur at shorter time lapse. One of such breakthroughs included sequencing of the human genome [43] and its public availability via publicly accessible databases such as NCBI Genome [44].

Accounting for the interaction context between system components is highly relevant for data originating from biological systems. The main methods used to integrate biological data can be broadly classified into methods integrating homogeneous and heterogeneous data types (Figure 1.5). The first type of integration involves homogeneous data types and is not new. Data integration has been already pioneered by meta-analytic studies on gene expression data [45,46]. In addition, GWAS meta-analytic studies are also gaining popularity. Rare and common complex diseases benefit from integration of several GWAS datasets due to increased power in detecting weak signals associated to disease risks [47,48]. Commonly the integration in meta-analytic studies involves combining  $p$ -values from various GWAS studies via Fisher's method [49]. The second type of integration involves heterogeneous data and, perhaps, is the most challenging. Compared to homogeneous data integration, heterogeneous data integration is much more problematic. Some examples of data integration on heterogeneous data can potentially include expression (microarrays), genotype (SNP), methylation, gene copy number variation (CNV). These data are measured on continuous and discrete scales. Integration of heterogeneous data generally involves scaling and conversion to common format/data space.

Some of the current approaches used to integrate data include kernel based, component-based (reduction) and network-based (dynamics and visualization). The kernel based approaches can be represented by the two main steps. The first one involves selection of appropriate kernel to summarize each data type while the second one involves combining different kernels to represent data comprehensively. For example, Lanckriet *et al.* [50] used SVN kernel-based approach for protein classification, while Reverter *et al.* [51] applied kernel based PCA to integrate successfully gene expression and fatty acid concentration data.



**Figure 1.5:** Homogeneous and heterogeneous data integration. Each layer represents a similarity matrix for a data source: mRNA (expression), DNA (genotype data), and methylation. “G” refers to genes (e.g., “G1” – gene 1).

Kernel based methods are computationally intensive and are faced with scaling issues requiring data transformation and reductionist approaches such as PCA [51]. In this case, data is plotted along new eigenvectors followed by the selection of several eigenvectors resulting in dimensionality reduction at the cost of information loss. Nevertheless, the limitation of reductionist approaches is that they prevent scientists to account for relationships existing between system components. Reductionist approaches such as PCA operate under the principle that complex systems or phenomena can be better understood by analysis of simpler individual components [52]. Data integration approaches face several issues including complexity and heterogeneity of data meaning that each system component can be represented by heterogeneous data (e.g., continuous, categorical, ordinal). Heterogeneous data integration is the novel area of active research. Bellow I will mention some of the key challenges currently faced by the community: 1) relevant data discovery (i.e. sources) to support biologically plausible models; 2) inadequate data standardization, annotation and storage; 3) strong generalization of biological phenomena while ignoring exceptions and individual contexts; 4) lack of user-friendly tools for wider community data accessibility and interpretation; 5) different data acquisition platforms that are not readily convertible to a common format; 6) others.

Our contribution to this data integration domain is several fold including a proposal of novel methodologies (*trans-eQTL MB-MDR epistasis protocol*,  $CIF_{mean}$  and *Regression2Net*) ready to

deal with homogeneous and heterogeneous omics data integration under different contexts exemplified over complex diseases including ankylosing spondylitis, asthma, type 1 diabetes and glioblastoma. Chapter 3 will cover integration of clinical and genotype data in the context of GWAIS. Chapter 4 will cover integration of genotype and expression data via mining of *trans/cis* epistatic eQTLs and construction of statistical epistasis networks highlighting gene regulatory mechanisms in asthma patients. Chapter 5 will introduce tree-based omics data integration with the posterior inference of gene expression regulatory model summarized by gene regulatory networks (GRNs).

## 1.6. Aims and goals of the thesis

In this thesis we focus on the problem of interactions identification in biological and clinical data under variety of contexts and fields. Specifically, we combined principles of statistics, machine learning and systems biology to develop omics-based genetic and epistatic statistical networks. The main contributions are summarized by the Figure 1.6. Our global main aim is to develop bioinformatics methodologies that integrate different layers (i.e. sources) of information to improve gene mapping of complex diseases (Figure 1.5). We investigate role of interactions and effect modifiers in gene mapping (i.e. causal disease genes) using the three cases. The first case covered in Chapter 3 uses genomic data and disease phenotypes to explore genetic epistasis (Figure 1.6). The second case uses genomic and gene expression data to explore intersection of statistical and biological epistasis covered in Chapter 4 (Figure 1.6). The third case uses omics data to explore gene network inference from single or multiple data sources explored in Chapters 5-6.

The developed epistasis detection methodologies are related to  $\text{SNP} \times \text{SNP} \rightarrow \text{disease phenotype}$ , *trans*  $\text{SNP} \times \text{cis} \text{ SNP} \rightarrow \text{gene expression}$  and gene regulatory network inference contexts. Figure 1.6 provides an integrated ‘Big Picture’ summary view of the data integration based on statistical and biological epistatic views.

The first aim is addressed by Chapter 3. There we check the robustness of the previously established genome-wide association interaction (GWAIS) protocol [53]. The role of each protocol tuning parameter related to data preparation and filtering is investigated (LD pruning, marker pre-

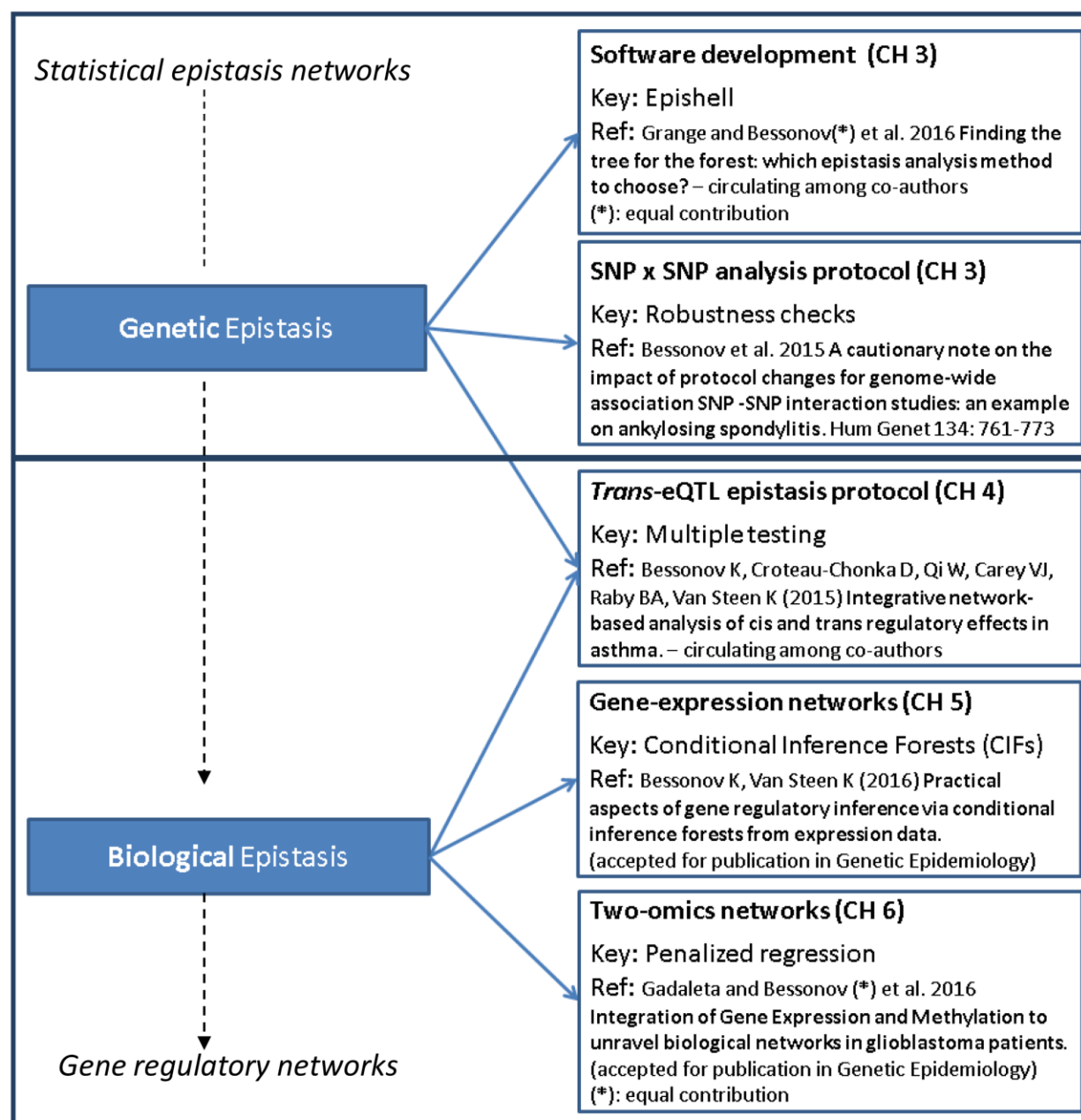
selection, epistatic method). The discussion is in the context of ankylosing spondylitis (AS). Specifically, the assessment of SNP x SNP interaction impact(s) on disease phenotypes is done via 1) integration of genotype and phenotype data; 2) suggestion of an optimal GWAIS protocol to detect statistical epistasis in genotype data; 3) identification of a biological relevance of epistatic SNP x SNP interactions in the context of ankylosing spondylitis.

The second aim covered in Chapter 4 identifies relevant epistatic genotypic SNP x SNP interactions and their impact on gene expression levels linked to a complex disease phenotype. Specifically, the joint effect of *trans*- and *cis*- gene expression regulation via genotypic component is studied in the context of asthma. In addition, the global impact of smoking (i.e. environmental variable) on the gene-gene interactome is assessed via network-based approaches. Due to extremely large number of possible SNP x SNP pairs and interaction hypotheses, the multiple testing correction is an issue. To this end, we designed a step-wise methodology to measure 2-way genotype *trans* SNP x *cis* SNP interaction impact(s) on gene expression via identification of *trans/cis* eQTLs. The gene regulatory mechanisms are inferred from *trans/cis* gene-gene regulatory network with posterior literature validation in the context of asthma.

The third aim gradually covered in Chapters 5-6 and involves the integration of multiple omics data types (expression, methylation, genotypes) via tree-based methodology and a gene regulatory network (GRN). The advantage of GRNs is that they allow to visualize a large number of interactions between genes and their regulators. Initially, a single data source (e.g., expression) is used to assess superior theoretical properties of conditional inference forest (*CIF*) compared to Random Forest (*RF*). There we identify an alternative to *RF* tree-based methods, the *CIF* variants, to infer GRNs from expression data. Specifically, we compare the gene regulatory network inference performance of *CIF*s to that of *RF* (Chapter 5). Finally, the *Regression2Net* methodology, based on penalized regression with inference of a single integrated GRN from multi omics data sources, is also presented in contributions section of the thesis - Section 6.3.

## 1.7. Reader's guide

This thesis is divided into 8 chapters , the relationships between some of these are shown in Figure 1.6. Chapter 2 covers the key concepts related to genome-wide association studies, gene regulation, tree-based feature selection, networks and other concepts. Chapter 3 validates and explores the impact of the key parameters of the GWAIS protocols on the final outcomes in ankylosing spondylitis data from WTCCC2. Chapter 4 extends the GWAIS protocol to *trans/cis* eQTL context of asthma based on CAMP expression and genotype real-life data. Chapter 5 addresses GRN inference via tree-based methods such as conditional inference forests (*CIFs*) including an introduction of the *CIF<sub>mean</sub>* method. Chapter 6 presents our contributions to other collaborative projects. Section 6.2 will cover patent describing the application of *CIFs* in the feature selection and classification contexts exemplified on asthma VOCs data. Section 6.3 introduces *Regression2Net* – penalized regression-based methodology – applied in the context of expression and methylation data integration. Chapter 7 provides a general discussion, conclusions and future perspectives of the thesis Chapters 3-6. Finally, Chapter 8 contains a list of all publications produced in the context of this thesis, together with author's curriculum vitae.



**Figure 1.6:** Thesis main topics areas, achievements and paper titles. Chapters 3 and 4 deal with statistical view of epistasis followed by Chapters 5, 6 and 7 that also incorporate biological views of epistasis in the context of transcriptional gene regulatory networks. The Chapter keys are the main topics.

## 1.8. Summary of main achievements

Our first main contribution is the validation of an earlier proposed GWAIS protocol [53] for the detection of detect gene-gene and gene-environment interactions. In particular, we have addressed the following questions (Chapter 3):

1. What is the impact of slight changes in the GWAIS protocol?
2. What are time requirements? In particular, can we handle genome-wide scales when MB-MDR is used in different contexts?
3. Which GWAIS protocol settings tend to produce the higher number of biologically relevant gene-gene interactions?
4. Which recommendations can we give to data analysts performing a GWAIS?

In an extension of the GWAIS protocol [53], which assumes a single trait and explores gene-gene interactions that affect this trait, we investigated *trans/cis* eQTL interactions. Hence, here we are dealing with multiple traits (gene expressions) and SNP-SNP interactions where at least one of the SNPs has a significant main effect on the trait. The following questions were addressed in Chapter 4:

1. Is MB-MDR suitable to detect *trans/cis* eQTL interactions? That is, can the false positive rate be kept under control? Is the multiple testing approach built in the MBMDR software adequate? Does it need to be adapted?
2. What is the extent of interaction between *trans* and *cis* regulatory gene expression components?
3. Can our *trans/cis* eQTL methodology detect disease relevant gene-gene interactions from genotypic and transcriptomic data?

Next, we developed methodologies for GRN inference, allowing the detection of ‘hidden’ structure in omics data. In particular, we investigated the use of tree-based approaches to infer GRNs from microarray expression data (Chapter 5). We also considered penalized regression approaches to infer GRNs from integration such data with epigenetic data (Chapter 6). The following questions were of interest:

1. Can tree-based techniques such as random forest (*RF*) and conditional inference forests (*CIFs*) provide suitable GRN inference performance?
2. What are the advantages of using penalized regression methods for GRN inference?
3. Can the best characteristics of *CIFs* and penalized regression based GRN inference be combined to give optimal performance?

Overall, this thesis addresses a very broad spectrum of problems related to biological and statistical interactions. My thesis covers the fields of statistical genomics, machine learning and molecular biology. The covered topics, particularly in the areas of omics data integration, merit further investigations.

## 1.9. References

1. Fan J, Han F, Liu H (2014) **Challenges of Big Data Analysis**. *Natl Sci Rev* 1: 293-314.
2. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, et al. (2015) **Deep learning applications and challenges in big data analytics**. *Journal of Big Data* 2: 1.
3. Dumbill E (2013) **Making sense of big data**. *Big Data* 1: 1-2.
4. Marx V (2013) **Biology: The big challenges of big data**. *Nature* 498: 255-260.
5. Sagioglu S, Sinanc D. **Big data: A review**; 2013. IEEE. pp. 42-47.
6. Ortigosa A, Carro RM, Quiroga JI (2014) **Predicting user personality by mining social interactions in Facebook**. *Journal of Computer and System Sciences* 80: 57-71.
7. Liu J, Dolan P, Pedersen ER. **Personalized news recommendation based on click behavior**; 2010. ACM. pp. 31-40.
8. EMBL-EBI (2014) **The European Bioinformatics Institute (EMBL-EBI) Annual Scientific Report 2014**. 47 p.
9. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) **The real cost of sequencing: higher than you think!** *Genome biology* 12: 125.
10. Crosswell LC, Thornton JM (2012) **ELIXIR: a distributed infrastructure for European biological data**. *Trends Biotechnol* 30: 241-242.
11. Gevaert O, Villalobos V, Sikic BI, Plevritis SK (2013) **Identification of ovarian cancer driver genes by using module network integration of multi-omics data**. *Interface Focus* 3: 20130013.
12. Zhu J, Shi Z, Wang J, Zhang B (2015) **Empowering biologists with multi-omics data: colorectal cancer as a paradigm**. *Bioinformatics* 31: 1436-1443.
13. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) **Similarity network fusion for aggregating data types on a genomic scale**. *Nat Methods* 11: 333-337.
14. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) **Galaxy: a platform for interactive large-scale genome analysis**. *Genome research* 15: 1451-1455.
15. EasyGenomics (2012) [www.genomics.cn/en/news/show\\_news?nid=99014](http://www.genomics.cn/en/news/show_news?nid=99014).
16. Fallowfield L, Jenkins V (1999) **Effective communication skills are the key to good cancer care**. *European Journal of Cancer* 35: 1592-1597.
17. Crick FH. **On protein synthesis**; 1958. pp. 138.
18. Alberts B (1998) **The cell as a collection of protein machines: preparing the next generation of molecular biologists**. *Cell* 92: 291-294.
19. Fischbach MA, Krogan NJ (2010) **The next frontier of systems biology: higher-order and interspecies interactions**. *Genome Biol* 11: 208.
20. Boulesteix A-L, Janitza S, Hapfelmeier A, Van Steen K, Strobl C (2014) **Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests**. *Briefings in bioinformatics*: bbu012.
21. Lenhard B, Wasserman WW (2002) **TFBS: Computational framework for transcription factor binding site analysis**. *Bioinformatics* 18: 1135-1136.
22. Moore JH (2005) **A global view of epistasis**. *Nature genetics* 37: 13-14.

23. Greene CS, Penrod NM, Williams SM, Moore JH (2009) **Failure to replicate a genetic association may provide important clues about genetic architecture.** *PLoS One* 4: e5639.
24. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, et al. (2011) **Characterizing genetic interactions in human disease association studies using statistical epistasis networks.** *BMC bioinformatics* 12: 364.
25. Milgram S (1967) **The small world problem.** *Psychology today* 2: 60-67.
26. Wasserman S, Faust K (1994) **Social network analysis: Methods and applications:** Cambridge university press.
27. Rothenberg RB, Potterat JJ, Woodhouse DE, Muth SQ, Darrow WW, et al. (1998) **Social network dynamics and HIV transmission.** *Aids* 12: 1529-1536.
28. Cordell HJ (2002) **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Human molecular genetics* 11: 2463-2468.
29. Fisher RA (1919) **XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.** *Transactions of the royal society of Edinburgh* 52: 399-433.
30. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, et al. (2011) **A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase.** *Nature genetics* 43: 487-489.
31. Solvang HK, Lingjærde OC, Frigessi A, Børresen-Dale A-L, Kristensen VN (2011) **Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer.** *BMC bioinformatics* 12: 197.
32. Yuan H, Westwick DT, Ingenito EP, Lutchen KR, Suki B (1999) **Parametric and nonparametric nonlinear system identification of lung tissue strip mechanics.** *Annals of biomedical engineering* 27: 548-562.
33. Kim SY, Imoto S, Miyano S (2003) **Inferring gene networks from time series microarray data using dynamic Bayesian networks.** *Briefings in bioinformatics* 4: 228.
34. Markowetz F, Spang R (2007) **Inferring cellular networks—a review.** *BMC bioinformatics* 8: S5.
35. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, et al. (2009) **Data integration in genetics and genomics: methods and challenges.** *Hum Genomics Proteomics* 2009.
36. Van Steen K (2012) **Travelling the world of gene-gene interactions.** *Brief Bioinform* 13: 1-19.
37. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merckenschlager M, et al. (2014) **Data integration in the era of omics: current and future challenges.** *BMC Syst Biol* 8 Suppl 2: I1.
38. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, et al. (2015) **Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer.** *PLoS genetics* 11: e1005689-e1005689.
39. MacLellan WR, Wang Y, Lusis AJ (2012) **Systems-based approaches to cardiovascular disease.** *Nat Rev Cardiol* 9: 172-184.
40. Takahashi K, Yamanaka S (2006) **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 126: 663-676.
41. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) **Methods of integrating data to uncover genotype-phenotype interactions.** *Nature Reviews Genetics* 16: 85-97.
42. Barabasi AL, Oltvai ZN (2004) **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 5: 101-113.
43. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) **Initial sequencing and analysis of the human genome.** *Nature* 409: 860-921.
44. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic acids research* 40: D130-D135.
45. Choi JK, Yu U, Kim S, Yoo OJ (2003) **Combining multiple microarray studies and modeling interstudy variation.** *Bioinformatics* 19 Suppl 1: i84-90.
46. Jiang H, Deng Y, Chen HS, Tao L, Sha Q, et al. (2004) **Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.** *BMC Bioinformatics* 5: 81.

47. Ioannidis JP, Patsopoulos NA, Evangelou E (2007) **Heterogeneity in meta-analyses of genome-wide association investigations.** *PLoS One* 2: e841.
48. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) **Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.** *Nat Genet* 33: 177-182.
49. Kost JT, McDermott MP (2002) **Combining dependent P-values.** *Statistics & Probability Letters* 60: 183-190.
50. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) **A statistical framework for genomic data fusion.** *Bioinformatics* 20: 2626-2635.
51. Reverter F, Vegas E, Oller JM (2014) **Kernel-PCA data integration with enhanced interpretability.** *BMC Syst Biol* 8 Suppl 2: S6.
52. Fang FC, Casadevall A (2011) **Reductionistic and holistic science.** *Infection and immunity* 79: 1401-1404.
53. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet* 133: 1343-1358.



## Chapter 2: Background





## 2. Background

This chapter provides a detailed background of the main thesis concepts related to genome-wide association studies (GWAS) and genome-wide association interaction studies (GWAIS), network inference via tree-based methods, different types of omics data integration and quality control before further analysis. We first start with basic concepts associated with both genomics and genetics.

### 2.1. Different omic measurement types

#### 2.1.1. Single Nucleotide Polymorphisms and modes of inheritance

Genetic material is stored in cell nucleus by double-stranded deoxyribonucleic acid (DNA). This molecule of life is composed of nucleotides that can contain one of the four bases including cytosine (C), guanine (G), adenine (A), or thymine (T). Thus, DNA is coded by 4-letter alphabet. DNA material is not static, its elementary units, nucleotides, can undergo alterations also known as mutations. Mutations can be beneficial, neutral or harmful occurring naturally throughout the life of any organism.

Single Nucleotide Polymorphisms (*SNPs*) can be defined as single nucleotide change (i.e. base) at a specific DNA location present within at least >1% of population (Figure 2.1). SNPs are one type of *genetic markers*, as they can be used to assess genetic differences between individuals. For example, let us take DNA sequences from two individuals at the same genomic location and DNA strand: ATTCC and ATGCC. Then at the highlighted position 3 there is a *bi-allelic* T/G SNP meaning that there exist two possible bases at that genomic location. In general, an allele is a version (variant) of a DNA segment defined by a genetic locus (*plural* ‘loci’) - the physical DNA location (with lengths ranging from 1 to  $\geq 1$  nucleotides). Since cells in the human body, except for gametes, are diploid (i.e. containing two versions for each chromosome), it is possible to consider an individual’s combination of alleles or *genotype* at a particular locus. Based on statistical genetics context, *genotypes* are defined as homogeneous wild type ‘AA’, heterogeneous ‘Aa’, and homozygous ‘aa’, when allele ‘A’ is the most frequent genotype. The allele frequency for the least frequent allele (‘a’) is referred to as the *minor allele frequency* and denoted by ‘MAF’.



**Figure 2.1:** SNP, allele and locus definitions. DNA is represented as double helix respecting base complementarity A-*T* and C-G. Eye color is taken as an example of a phenotype. Genotype is a specific set of alleles (e.g., ‘*Aa*’) with potential impact on phenotype.

A *genetic model* captures the relationship between an allele or genotype and a *phenotype* (any visible characteristic of an individual or organism) or a *trait* (coded phenotype). For instance, if the risk to a binary trait conferred by an allele is increased  $k$ -fold for heterozygotes and  $k^2$  ( $2k$ )-fold for homozygotes, then the corresponding genetic risk model is *multiplicative* (*additive*). If the risk on a binary trait induced by the ‘*Aa*’ genotype (heterozygote) individuals lies between that of ‘*AA*’ (wildtype homozygote) and ‘*aa*’ (minor allele homozygote) individuals, but not in the specific relationship of a multiplicative or additive model, the genetic model is *codominant*.

### 2.1.2. Epigenetic markers

Several mechanisms exist that may change or regulate gene expression. One such mechanism is epigenetics. Epigenetic events affect gene expression without making changes in the DNA sequence. Two molecular mechanisms are histone modification and DNA methylation. For the latter a methyl ( $\text{CH}_3$ ) group is added to the DNA (in particular, a cytosine nucleotide). DNA sequences with a high number of di-nucleotide CpG repeats, at least 200 bp long, are referred to as CpG islands. These islands are often located in the promoter regions of genes. Methylation of CpG islands in the vicinity of gene’s coding region usually blocks initiation of transcription. Conversely, methylation in the gene’s coding region might stimulate transcription elongation resulting in gene’s

activation [1]. In addition to gene expression regulation, methylation plays an essential role in embryo development and accessibility to genomic DNA by controlling compaction of the heterochromatin [2,3]. Therefore, for a more complete and comprehensive picture of gene expression patterns and their links to complex diseases phenotypes, it is also important to consider methylation data.

### 2.1.3. Gene expression measurements

Gene expression data referred to as transcriptome, represents the amount mRNA in the sample. Following the central dogma of molecular biology, genomic DNA is translated to mRNA after recruitment and binding of transcription machinery (which includes RNA polymerase II) to the genomic region preceding the transcription start site (TSS) of a given gene. The RNA polymerase II synthesises a new strand of mRNA that can be translated to a protein associated with a biological function. Thus, the amount of mRNA in the sample represents the expression level of a given gene. Changes in gene expression are often associated with the external environmental factors such as responses to a drug treatment, smoking, gene knock-outs, and presence or absence of a particular disease/phenotype [4].

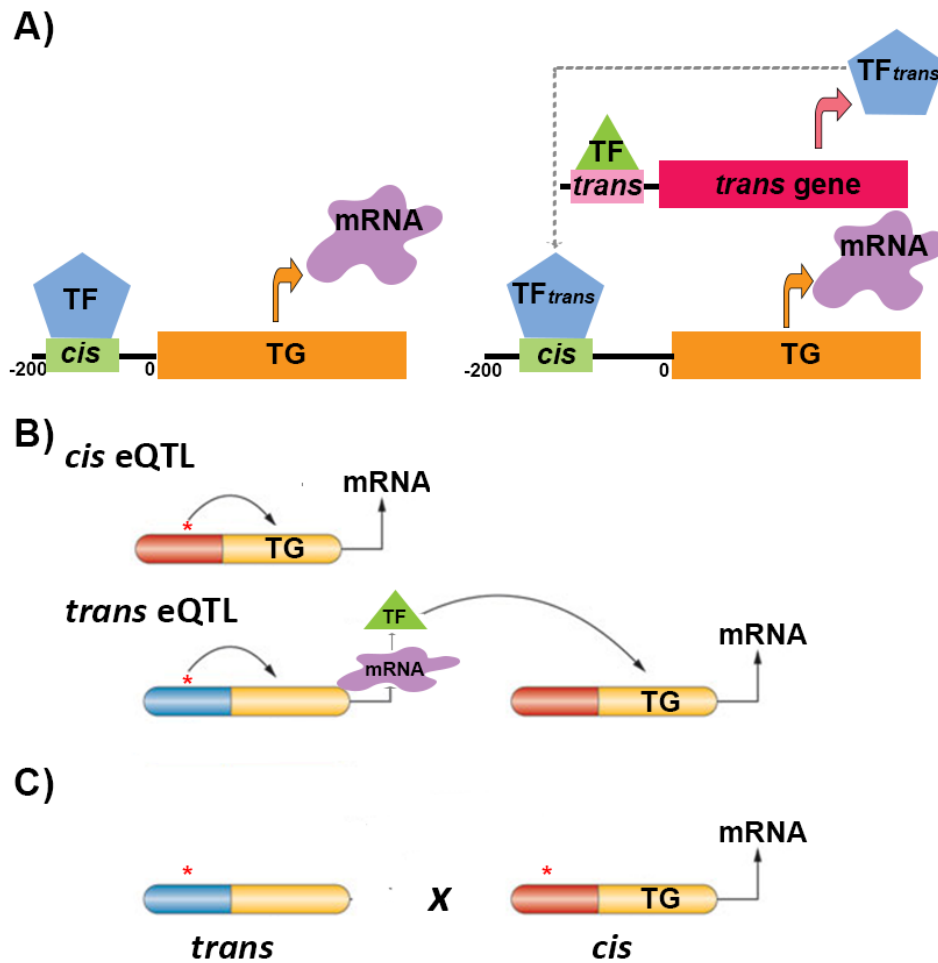
The human transcriptome is large containing thousands of protein-coding genes (~22 000) that can be conveniently monitored via microarrays providing almost a complete coverage [5]. Most of the publically available human expression datasets available at Gene Expression Omnibus (GEO) contain a limited number of samples averaging at hundreds of individuals. This means that expression datasets are high-dimensional with each gene represented by a handful of samples. To improve the quality of biological hypothesis(es) and quality of gene mapping efforts, one needs to apply well established and reliable quality control protocol.

Gene expression can be regulated via *cis* and *trans* regulatory sequences found outside the protein coding sequence – open reading frame (e.g., in the promoter region). These DNA regulatory sequences bind regulatory proteins commonly referred to as transcription factors (TFs). The *cis*-regulatory sequences are located in the vicinity of the target gene (TG) approximately within ~200 bp upstream of the protein coding region binding a selected set of TFs that regulate expression of

the TG (Figure 2.2A). The TG gene expression can be also governed by distantly located *trans*-regulatory sequences controlling the expression of *trans*-acting TF (Figure 2.2A). In this case, the TG expression is said to be controlled in *trans* since it is dependent on the expression of the *trans* TF gene (Figure 2.2A).

Locus that controls transcript expression levels (mRNA) is called expression quantitative trait locus (eQTL). Different alleles of the locus can impact binding ability of the regulatory proteins, such as TFs, and, hence, the TG mRNA levels. If eQTL is associated with a closely located locus with respect to the TG (e.g., in a promoter region), then such eQTL acts in *cis* (Figure 2.2B). In turn, *trans* eQTL involves distantly located locus that controls the expression of the *trans*-acting TF acting on a TG (Figure 2.2B). *Trans* locus can be located on the same or even another chromosome. Thus, *trans* eQTL involves either distant or indirect expression regulation of the target gene. Typical eQTL analysis treats expression as phenotypic trait (i.e. response) while the loci alleles as genotypic predictors. The eQTL analysis links genetic variation and expression considering inter layer interactions between genomic and transcriptomic data sources. In particular, eQTL studies were useful in providing insights into gene regulatory mechanisms and pathways in cancer linked to genotypic variation [6,7].

In Chapter 4 we consider the joint epistatic effects of *trans* and *cis* loci on the expression trait depicted schematically by Figure 2.2C. Predicted interaction between a given *trans/cis* loci eQTL pair also has a biological context. There is evidence that epistatic interaction between *trans* and *cis* loci can occur in biological systems [8]. Due to chromatin looping of DNA segments even distantly located loci can be brought closer and jointly regulate gene expression [9]. In contrast to individual *trans* and *cis* eQTL scenarios (Figure 2.2B) exploring only marginal effects, the *trans/cis* epistatic eQTL screenings search for 2-way interactions associated with the expression phenotype (Figure 2.2C). While there are a lot of studies analyzing effects of individual markers on expression traits, only recently the joint loci effects are being considered to a greater extent [10].



**Figure 2.2:** *cis* and *trans* gene expression regulation. Black line represents genomic DNA sequence. The orange square represents the coding region of the target gene (TG) with 0 representing the beginning of the transcription start site (TSS). The 200 bp upstream (-200 bp) of the TSS the transcription binding site is located represented by green and pink rectangles. The TSS binds transcription factors (TF) modifying expression of the nearby or distant target genes. **A)** TG expression is regulated by nearby *cis*-regulatory sequence (green square); TG expression is regulated by TF expressed by a distant *trans* gene. **B)** *cis* eQTL - the locus (\*) located near the protein coding sequence in yellow controls gene expression; *trans* eQTL - the locus (\*) controls expression of a distant gene via *trans*-acting TF. **C)** epistatic *trans/cis* eQTL showing *trans* and *cis* loci (\*) affecting the TG expression. Due to interaction, both loci impact final mRNA levels of the TG.

## 2.2. Genome-wide association studies (GWAS)

In the past decade, a considerable effort was devoted to the linkage of the genetic loci to disease effects and complex traits. Previously, causal variant studies tried to use linkage analysis together with pedigree data in order identify causal mutations that contribute to the disease risk [11]. These studies used prior scientific evidence and experimental work suggesting that a given causal locus (i.e. SNP) is relevant to a disease trait [11]. Thus, candidate polymorphism studies used a SNP functional analysis to determine the impact of genotype on a complex trait - disease status. Unfortunately, the scale of these studies was rather limited.

With the advent of high-throughput technologies, it was possible to obtain genotypic information simultaneously for a larger number of SNPs and individuals. This promoted an extensive use of genome-wide association studies (GWAS). The goal of GWAS is to identify statistically significant associations between a genetic marker (or markers) and a trait in a given population. Most frequently, the tests that are carried out in GWAS contexts are equivalent to those obtained from a regression model, where the genetic marker information at locus  $i$  is captured by a variable  $X$  (see Eq. 2.1 for a continuous trait (e.g., body mass index ), where  $\beta$ 's are regression coefficients and  $c_i$  is the error term). GWAS can also be done using related individuals in family-based designs such as in [12].

$$Y_{trait} = \beta_o + \beta_1 X_{locus\ i} + c_i \quad \text{Eq. 2.1}$$

The initial interest to GWAS owns to the advent of SNP arrays and high hopes that single gene disorders (e.g., muscular dystrophy, cystic fibrosis) can be extrapolated to multigenic complex disorders (e.g., diabetes) [11]. Despite several criticisms [11], GWAS were successful in the discovery of novel biomarkers and provided new insights into the etiology of complex diseases. According to the 2013 statistics, the Catalog of Published Genome-Wide Association Studies contained 1778 curated GWAS studies that discovered 12,123 statistically significant SNP-trait associations at  $5.0 \times 10^{-8}$  significance level [13]. Unfortunately, the significant SNP-trait associations identified by GWAS account only a limited amount of the genetic variance in a population, raising a problem of missing heritability [14]. For example, height is 80-90% heritable

trait, yet GWAS studies only account 5% of the height heritability. The main causes of missing heritability include causal alleles with small effect sizes, rare variants, epigenetic effects, epistatic interactions and others [15]. This indicates that 1-dimensional association tests (Eq. 2.1) in GWAS are rather limited and that more advanced epistatic models accounting for gene-gene interactions are required to achieve a better performance [16].

### 2.3. Genome-wide association interaction studies (GWAIS)

Detection of epistatic signals is not trivial and is more complicated than in the case of GWAS that considers only the main effects (i.e. single locus and a trait interaction). GWAIS are faced with the same issues present in GWAS at a greater scale. Identification of true epistatic signals is statistically and experimentally challenging [17]. The experimental challenge is linked to the large sample sizes requirements due to the extremely large landscape of genetic interactions and low signal-to-noise ratio. A panel of 100 SNPs (resp. 1000, 10,000) would require assessment of 4950 SNP x SNP (resp. 499,500, 49,995,000) interactions. Thus, the number of genetic interactions grows exponentially and is typical of large-scale epistatic studies accentuating the ‘curse of dimensionality’ problem [18]. In statistics, the ‘curse of dimensionality’ relates to the problem of slow convergence of the estimated statistic to the true value in high dimensional space. That is additional variables in the model require new samples. The statistical challenge lies in the severe penalty incurred due to a large number of statistical tests requiring very small  $p$ -values (i.e. strong signals) to reach statistical significance [19]. A typical significance threshold for GWAIS is at  $p$ -value  $10^{-13}$  while that of a GWAS is only of  $10^{-5}$  [20]. Finally, the computational challenge lies in the number of tests to be evaluated raising scalability issues requiring a clever use of information storage and information coding/compression that takes into account the computer architecture specifics. For example, BOOST exploits bitwise operations allowing a better use of CPU cycles while promoting memory and storage efficiency [21].

The complexity of diseases calls for more advanced models accounting for epistatic interactions. Indeed, the effect of one locus can be masked or modified by the other reducing the detection power of the first locus (Section 1.3). These loci interactions in the context of the phenotype are referred to as epistasis that was initially discussed in Section 1.3 and illustrated in Figure 1.3. The epistatic

interactions have statistical and biological interpretation introduced in Section 1.3 and Figure 1.3. Cordell *et al.* provides clear definitions and examples of both epistasis types in [22].

The number of methods to detect epistatic effects in complex human disease is steadily growing with the development of novel tools and GWAIS methodologies [21,23-25]. Epistasis is proven to exist in model organisms [17], is wide-spread, and should be accounted for in the experimental design.

There are a large number of tools allowing identification of epistatic effects [21,24,26]. Parametric regression-based approaches are common [19]. The classical approaches use a linear regression-based framework that not only accounts for main effects, but also for interactions between loci pairs (Eq. 2.2), where  $X_{locus\ i}$  and  $X_{locus\ j}$  refer to the  $i$ th and  $j$ th loci and  $\beta_3$  measures the interaction strength). For example, FastANOVA [26] uses two-locus ANOVA test to explore the interaction space of SNP x SNP interactions via implementation of heuristics during the calculation of the upper bound for a group of SNP pairs (e.g.,  $X_{locus\ i}$  and  $X_{locus\ j}$ ).

$$Y_{trait} = \beta_o + \beta_1 X_{locus\ i} + \beta_2 X_{locus\ j} + \beta_3 X_{locus\ i} * X_{locus\ j} + c_i \quad \text{Eq. 2.2}$$

Regression-based methods cannot always optimally account for the main effects and present difficulties when working with ‘rare’ loci (MAF < 0.05). Rare or absent genotypes in high-dimensional data spaces are particularly problematic to deal with [19]. Other popular epistatic methods including BOOST and MB-MDR use alternative methods to cope with these GWAIS burdens. For instance, BOOST [21] tests for significance of the SNP x SNP interaction effect via likelihood ratio test while MB-MDR [24] uses semi-parametric dimensionality reduction methods discussed in a greater in the subsequent Section 2.3.2.

Finally, the lack of systematic studies, the standardized GWAIS protocol and gold standard data complicates methods comparison. To address this void, Gusareva *et al.* developed GWAIS protocol [27]. In this thesis work, we will further test the impact of parameter choices of the GWAIS protocol under the context of the complex disease - ankylosing spondylitis (Chapter 3).

### 2.3.1. Epishell

I was involved in software testing and improvement of the tool initially developed by Tom Cattaert. My involvement included debugging and testing tasks, improvements in the software ability to handle genome-wide datasets ( $> 65,535$  markets), inclusion of degrees of freedom in association tests and others. This led to the development of improved epistasis detection tool, EpiShell, based on popular ‘Boolean Operation-based Screening and Testing’ (BOOST) method [21] with a few enhancements related to missing data handling. Compared to classical BOOST [21], the significance of the association test statistic is more accurately estimated by taking into account missing genotype(s). For this purpose, the degrees of freedom (df) are estimated for each candidate SNP pair. EpiShell offers several ways to calculate statistical epistasis including using the score and the log-likelihood ratio tests referred to as ‘BOOST’ and ‘MB-MDR like’ modes. Another novelty of the BOOST method lies in Boolean representation of genotype data and bitwise operations to obtain SNP x SNP contingency tables in line with the computer hardware design. Boolean data representation allows BOOST efficiently store data and efficiently utilize CPU resources significantly improving the calculation speed decreasing the overall run-time requirements.

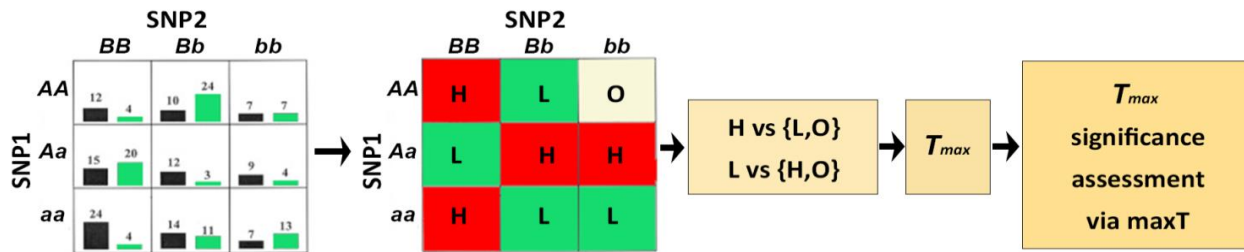
In the BOOST-like mode, EpiShell handles binary traits and fits a full generalized linear model with the main SNP effects (2 degrees of freedom (df) for each main effect) and SNP x SNP interaction effects (4 df). Significant (specific) interactions are identified via a Log-Likelihood Ratio Test (LRT) based on 4 df. The Bonferroni correction is proposed as *a posteriori* multiple testing corrective measure (not implemented). For additional details and performance indicators as compared to other epistasis detection methods (PLINK, EPIBLASTER and FORCE) in simulated data please refer to [28]. The EpiShell binary and its source code can be freely downloaded from <https://bitbucket.org/kbessonov/epishell> and <http://www.statgen.ulg.ac.be>.

### 2.3.2. Model-Based Multifactor Dimensionality Reduction (MB-MDR)

The Model-based multifactor dimensionality reduction (MB-MDR) method was introduced by Calle *et al.* [29] extending the MDR method of Ritchie *et al.* [30]. The MB-MDR method can currently deal with binary, continuous and censored traits, correct for confounders or lower-order

effects. The MB-MDR binary can be freely downloaded from [www.statgen.ulg.ac.be/software.html](http://www.statgen.ulg.ac.be/software.html). The MB-MDR code is written by François Van Lishout [31]. Briefly, MB-MDR carries out a dimensionality reduction procedure by pooling risk-alike multi-locus genotype combinations together into the low-dimensional construct. Its final test statistic contrasts ‘high risk’ versus ‘low-risk’ multi-locus genotypes (H/LO and L/HO) as shown in Figure 2.3. While correcting for multiple testing, the significance of the epistatic interaction is assessed via the resampling-based strategy proposed by Westfall *et al.* [32].

In more detail, a bi-allelic marker has maximum a total of 9 multi-locus genotypes (Figure 2.3). Depending on the trait (e.g., binary or continuous), each multi-locus cell is tested against the remaining cells using a Student’s t-test or Chi-square test at the 0.1 significance level ( $p\text{-value} < 0.1$ ). Second, based on the obtained results from such association tests, the genotype cells are labeled as high risk (‘H’), low risk (‘L’) and no evidence – ‘O’. Thus, now the interaction between two loci can be summarized by this lower dimensional *HLO* construct since explanatory multi-locus variable with 9 levels is reduced to 3. Third, the subsequent two association tests comparing ‘H’ versus ‘L’ and ‘O’ cells and one comparing ‘L’ against ‘H’ and ‘O’ cells are performed. The maximum of the two statistics is selected representing the observed statistic  $T_{max}$  of the SNP x SNP association to the trait (Figure 2.3). In step 3, the significance is assessed by adopting a permutation-based *MAXT* correction [32]. The MB-MDR is able to correct for main effects minimizing spurious SNP x SNP interactions. To do so, the MB-MDR first regresses out the main effects from the trait and uses the resulting residuals as new traits. The two extreme modes of correction are implemented including additive and co-dominant. For more details please refer to [24,29,33].



**Figure 2.3:** Graphical summary of MB-MDR epistasis detection for a binary response (e.g., case/control). Each pair of SNPs is tested for strength of association to the response variable (i.e. trait) summarized by a permutation-based  $p$ -value.

### 2.3.3. Highlights

#### ***Recursive Partitioning***

Recursive Partitioning methods are widely used for classification, regression, and variable importance assessment tasks (e.g., feature selection). Decision trees [34] are directed acyclic graphs used to recursively partition and explore data. They reconstruct a relationship between  $Y$  and  $X$  representing response and prediction variables, respectively. Decision trees (DT) can also be viewed as predictive models represented by a set of rules used to calculate an output (class(es) or continuous variable(s)). Their main advantages include ease of interpretation and an intuitive visual structure (Figure 2.4A), ability to explore the feature space non-linearly, ability to handle multiple outputs, ability to work with heterogeneous data, and others [35].

The structure of the tree is shown in Figure 2.4A. The ‘root node’ is located on the upper most level of the tree and represents the entire learning sample ( $LS$ ) portion of the dataset. Node  $N$  can be split into two resulting in left child ( $N_L$ ) and right child ( $N_R$ ) nodes. The source node that produced the child nodes is referred to as ‘parental node’ and is located one level higher. The ‘leaf’ or ‘terminal’ nodes ( $N_{leaf}$ ) do not contain any children nodes and are located at the bottom of the tree. The  $N_{leaf}$  nodes are usually assigned to majority class or mean of  $Y$  depending on the classification or regression contexts. For example, let us consider hypothetical decision tree used for classification of smokers and non-smokers based on ‘alcohol per month’ consumption and ‘jogging’ variables (Figure 2.4B). Initially, the data is split based on the ‘alcohol per month’ variable producing left child terminal node 2 and right child node 3 representing subsets of data containing  $n_L$  and  $n_R$  observations. Node 2 is rather ‘pure’ with 18% of  $S$  and 88%  $NS$  samples equivalent to  $p(S/N_{leaf})=0.18$  and  $p(NS/N_{leaf})=0.88$  where  $N_{leaf}$  denotes a leaf node. Node 3 assigned to jogging binary variable is further split producing terminal nodes 4 and 5. Node 5 attains maximum purity being entirely composed of  $S$  samples ( $p(S/N_{leaf})=1, p(NS/N_{leaf})=0$ ). Logically the Figure 2.4B tree shows that the lifestyle habits have impact on the individual smoking status.

Algorithms that learn decision trees can be dated back to 1984. The two most popular algorithms for classification and regression tasks introduced by Braiman *et al.* were CART [34] and C4.5 [36]. Their non-parametric approach, intuitive interpretability and being able to deal with small  $n$  large

$p$  datasets made these methods popular [35]. The CART utilizes early stopping and tree pruning to avoid over-fitting issues. The algorithm searches for the “best” split minimizing  $I(N)$  impurity function – i.e. splits leading to more pure/homogeneous nodes are preferred. CART and C4.5 offer GINI index and entropy as impurity measure, respectively [37]. The downside of CART is the lack of a formal statistical test to guide the feature selection process and bias towards features with many possible splits [37]. We refer to this issue as threshold selection problem for the candidate list of top ranked features. Conditional Inference Forest (*CIF*) algorithm, introduced in the subsequent section, overcomes this limitation by performing association test between response and predictor variables ( $Y \sim X$ ) reducing the variable selection bias [35].

Trees can grow indefinitely leading to data over-fitting problem. Thus, it is essential to apply stopping rules which can include a) minimum number of observations in nodes; b) maximum levels in a tree (i.e. depth); c) threshold for the minimum change in the impurity measure and others.

Most of the decision tree based algorithms rely on the concept of node purity and homogeneity in order to implement further splits. The change of impurity at node  $N$  for a binary split is defined by general function  $\Delta I(N)$  (Eq. 2.3) where  $I(N)$ ,  $I(N_L)$  and  $I(N_R)$  represent individual node impurity functions of node  $N$  and its left and right children ( $N_L$  and  $N_R$ ). Small  $n_L$  and  $n_R$  denote number of samples in the left and right nodes.

$$\Delta I(N) = I(N) - \frac{n_L}{n_L + n_R} I(N_L) - \frac{n_R}{n_L + n_R} I(N_R) \quad \text{Eq. 2.3}$$

The node impurity common measures are entropy (Eq. 2.4), GINI index (Eq. 2.5), and variance (Eq. 2.6). The maximum purity for GINI is 0 equivalent to 0 entropy. Variance measure also referred to as *least squares deviation* (LSD) [38], is commonly used in regression problems when  $Y$  is continuous.

$$\text{Entropy} = - \sum_{i=1}^K p_i \log_2 p_i \quad \text{Eq. 2.4}$$

where  $K$  is the total number of classes and  $p_i$  is probability of the  $i$ th class

$$GINI = 1 - \sum_{i=1}^K p_i^2 \quad \text{Eq. 2.5}$$

$$Variance = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Eq. 2.6}$$

where  $n$  is the number of samples and  $\bar{y}$  is the mean of the response variable  $Y$

### ***Tree ensembles***

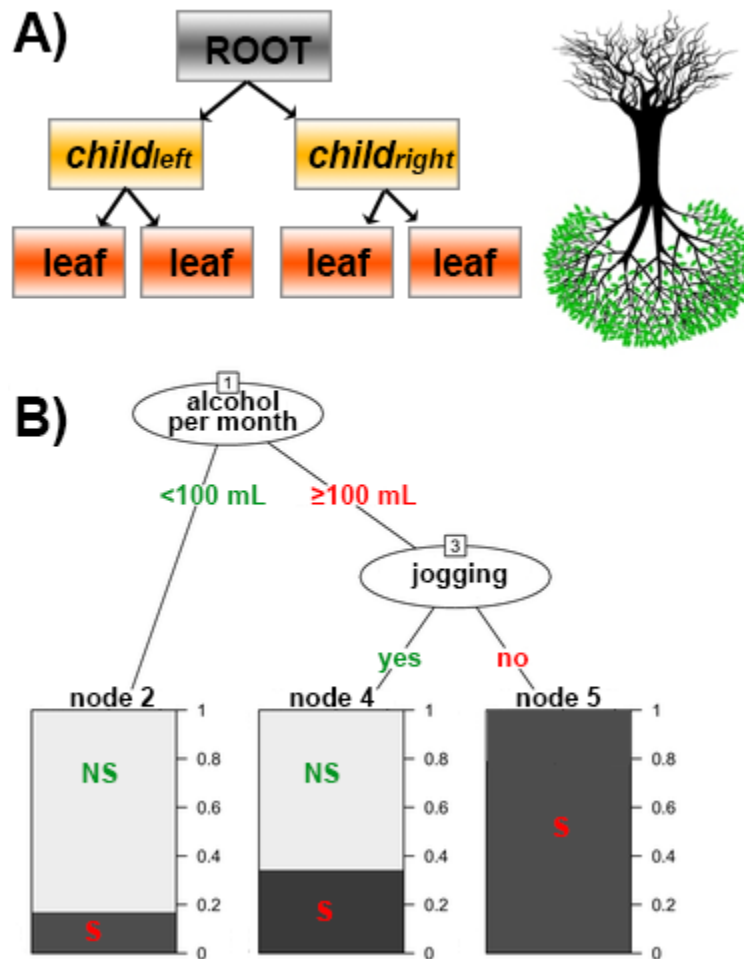
Decision trees show high variance indicating a high dependency on data and lack of robustness to outliers [39]. Thus small changes in data strongly impact decision trees inference. Each subsequent split of the decision tree depends on previous ones propagating error in case of sub-optimal split and increasing variability between trees. Several empirical studies show that ‘tree ensembles’ reduce variance by averaging predictions over several trees in the ensemble [40-42]. Thus, the power of tree ensembles lies in combining predictions across trees of the ensemble achieving higher prediction accuracy. In addition, averaging predictions across the tree ensembles smoothens the decision boundaries separately defined by each tree node [35].

In the scope of this thesis we will consider methods that introduce randomization to tree ensembles including well-known random forest (*RF*) [43] and condition inference forest (*CIF*) algorithms [44,45].

### ***Random Forest***

Most of the classical GWAS analyses are univariate since they analyze a single genetic marker with relation to the trait. Random Forest can accommodate large number of variables and build complex models of epistatic interactions making them ideal for analysis of biological data. Winham

*et al.* demonstrated that *RFs* have higher power to detect especially non-interacting SNPs in genetic data [46].



**Figure 2.4:** Decision trees: **A)** structure of a typical decision tree; **B)** hypothetical tree applied to smokers (S) and non-smokers (NS) classification problem based on heterogeneous types of variables: ‘alcohol per month’ consumption on continuous and ‘jogging’ activity on binary scales.

The logic behind *RF* is as follows, the random forest builds ensembles of trees utilizing bootstrapping that is sampling with replacement meaning that some samples can occur more than once. The  $n$  bootstrap samples are drawn creating  $n$  learning samples (*LS*). One tree is built for each bootstrap sample. Keeping in mind that each individual tree is highly dependent on each *LS*,

there can be high variability between individual trees allowing to explore differently the data space of the  $LS$ .

The  $RF$  advantages are their non-parametric feature allowing approximation of any unknown functions  $f(x)$  without knowing their *apriori* shapes. The ensembling feature makes  $RF$  suitable for datasets with complex  $n$ -order interactions and high number of dimensions typical of large  $p$  small  $n$  datasets. In comparison, the classical parametric regression-based approaches tend to over-fit to a greater extent such complex datasets requiring definition of additional parameters. Extra randomization conferred by a random pre-selection of the pool of the splitting variables via the definition of the  $mtry$  parameter, allows production of more diverse trees while giving a chance to lower ranked predictor variables (i.e. with lower score) being included as nodes of the growing trees [47]. This pre-selection amongst  $X$  gives an extra chance for other variables to appear in the context of other covariates.

In addition, to all these advantages,  $RF$  also suffers from several pitfalls including use of GINI variable importance measure (Eq. 2.5) shown to be biased in the presence of predictors with many possible splits (e.g. continuous, multi-categorical) discussed in [48]. This selection bias is still present even in the case of the permutation-based variable importance measure ( $VIM$ ) calculated from the previously built tree ensemble for each predictor variable ( $X$ ) [48]. Strobl *et al.* suggested use of sub-sampling, sampling without replacement, in combination with unbiased split criterion implemented in conditional inference forests [48]. In addition,  $RF$  showed selection bias towards correlated variables even after application of permutation-based  $VIM$  [48].

### ***Conditional Inference Forest (CIF)***

The unbiased conditional inference forest ( $CIF$ ) algorithm was suggested as alternative to  $RF$ . The strengths of  $CIF$  include unbiased variable selection criterion, minimization of over-fitting by provision of stopping rule, ability to deal with additional response variable scales, robustness to different variable scales, conditional permutation-based  $VIM$  dealing with correlated predictors [44,45]. To achieve this,  $CIF$  divides variable selection and splitting into separate steps minimizing

bias and overfitting [48]. Figure 2.5 describes the main node variable selection steps as implemented in *CIF*.

In addition, The *CIF* brings statistical notions to tree ensembles while using concepts from *RFs* such as permutation-based *VIM* and randomness conferred by *mtry* parameter. Importantly, the *CIF* addresses the *RF* bias and over-fitting issues [48].

The R library *party* [49] implements conditional inference forests providing several options to control tree building including variable selection approaches including bootstrapping (*replace=T*) and sub-sampling (*replace=F*) options, multiple-testing Bonferroni (*testtype="Bonferroni"*) and Monte Carlo (*testtype="MonteCarlo"*) correction and minimal significance threshold definition (*mincriterion=0.95*), minimum number of observations in a node (*minsplit=20*) amongst others.

```

Select randomly  $m$  variables from  $X$  defining  $X_m = \{x_1 \dots x_m\}$  set
where  $m$  is specified by the mtry parameter
Select one covariate  $x_j \in X_m$  via estimation of  $c_{max}^1$ 
Assign  $x_j$  to a tree node where  $x_j$  is a node variable
Search for the best split of the  $x_j^*$  via estimation of  $c_{max}^2$ .
Repeat previous steps if none of the stopping criteria are met

```

**Figure 2.5:** Conditional inference tree node variable selection and splitting steps implemented during the tree growth.  
Legend:  $p$  is the total number of variables and  $q$  is the number of samples of the  $p \times q$  input data matrix.

The *RF* and *CIF* ensemble algorithms will be further used to infer gene networks from diverse biological data in the context of complex diseases (Chapter 5).

---

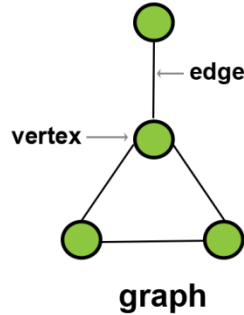
<sup>1</sup> Standardized univariate linear test statistic of association between  $X$  and  $Y$   $c = |\frac{t-\mu}{\Sigma}|$ . The statistic is computed for every  $m$  selected variables in  $X_m$ .  $c_{max}$  is the maximum out of all  $c$  statistics of  $X_m$ . Further details on  $c_{max}$ ,  $t$ ,  $\mu$  and  $\Sigma$  definitions are described in [49].

<sup>2</sup> Standardized univariate linear two-sample statistic  $c = |\frac{t^A-\mu}{\Sigma}|$  quantifying the split quality of  $x_j^*$ . The statistic is computed on the response  $Y$ . The  $c_{max}$  measures the discrepancy between subsets  $A_1^*$  and  $A_2^* \in \mathbb{R}^q$  after a split at a cut-point (\*). For further details on  $c_{max}$  and  $t^A$  statistics in the context of splitting refer to [49].

## 2.4. Networks

### 2.4.1. Network syntax

Mathematically networks are referred to as graphs ( $G$ ) composed of a set of nodes ( $V$ ) and edges ( $E$ ) connecting nodes defined as  $G = (V, E)$ . That is  $V = \{v_1, v_2, v_3, \dots\}$  and  $E = \{(v_1, v_2), (v_2, v_3), \dots\}$ . Depending on the context, nodes can be referred to as vertices and edges as links (Figure 2.6).



**Figure 2.6:** Graphs: key elements

Graphs which contain *directed* edges are *directed* while those that do not are *undirected*. Figure 2.6 displays *undirected* graph. The *directed* graphs should be used when the connection direction  $v_i \rightarrow v_j$  does not imply existence of an opposite  $v_j \rightarrow v_i$  direction.

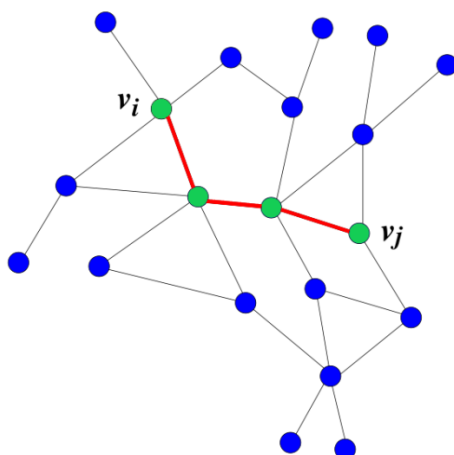
In addition, graph edges ( $E$ ) can carry a weight  $w \in W$  such that  $w \in \mathbb{R}$ . These graphs are *weighted* and can be expressed by  $G = (V, E, W)$ . The weights  $w_{ij}$  and  $w_{ji}$  measure the association strength between the node pairs  $(v_i, v_j)$  and  $(v_j, v_i)$ .

The subgraph  $S = (V', E')$  is part of  $G = (V, E)$  such that  $V' \subseteq V$ ,  $E' \subseteq E$  and  $E' \subseteq V' \times V'$ . A particular type of subgraphs is a *clique* when all its node pairs being connected. If clique contains the largest number of graph edges such clique is the *maximum clique*. The  $G$  is ‘*complete*’ when all graph nodes are being connected by  $n(n-1)/2$  number of edges where  $n$  is the total number of nodes in  $G$ .

The key graph theory concepts include *graph size*, *order*, *density*, *node degree*, *shortest path*, and *betweenness*. The graph size  $n$  is simply the number of nodes ( $v$ ) in the set  $V$ . The order  $l$  is the number of edges in  $E$ . Node degree  $d$  is the number of nodes a given node is linked to. The range

of  $d$  is from 0 to  $n-1$ . The *degree distribution* is vector  $k=(k_0, k_1, \dots, k_{n-1})$  where  $k_i$  represents the degree of the  $i$ th node. Thus,  $p(k)$  is the proportion of nodes in a graph with degree  $k$  ( $p(k) = \# \text{ nodes with degree } k / \text{total } \# \text{ number of nodes}$ ). The *shortest path* ( $L_{ij}$ ) is the minimum number of edges/steps required to connect a pair of nodes  $\{v_i, v_j\}$  (Figure 2.7). For a pair of nodes, the shortest path requires to use three edges shown in red to connect  $v_i$  with  $v_j$ . Lastly, the *node betweenness centrality* ( $C_i$ ) is defined as the number of shortest paths  $L$  that include  $i$ th node  $v_i$ . The *node betweenness* measure is indicative of the node centrality in a graph and is linked to the traversal frequency based on the random walks between nodes [51].

The degree distribution of nodes determines the network topology. It can be visualized by plotting  $p(k)$  against  $k$  on linear or logarithmic scales. The *random networks* are characterized by random edges. In these types of networks the node degrees follow Poisson distribution  $p(k) = \frac{e^{-c} c^k}{k!}$  (where  $c$  is the constant). This indicates that nodes have approximately the same number of links approaching mean of  $k$ . *Scale-free* networks are heterogeneous with respect to the degree distribution with small number of highly connected (i.e. hubs) and large number of sparsely connected nodes. The degree distribution of such networks follows the power law  $p(k) = k^{-\gamma}$  where  $\gamma$  is the degree exponent. This indicates that in scale free networks the probability that node is highly connected is higher than in random networks. The scale-free networks are present everywhere from *E.coli* gene networks to World Wide Web (WWW). In scale free networks the  $\gamma$  typically takes values between 1.9-2.3 [52]. In addition, scale free networks have a shorter average path length ( $L$ ) compared to *random networks*.



**Figure 2.7:** Example of the shortest path. The shortest path between  $v_i$  and  $v_j$  nodes is 3

### 2.4.2. Network medicine

Section 1.2 presented biology as information science with a hierarchical organization of its components. Networks simplify complex systems by presenting system elements by nodes and their interactions by edges (Figure 2.6). Biological networks model diverse interactions between macromolecules (DNA, mRNA, proteins, metabolites) and other species. The systems biology challenge is to model accurately and integrate networks spanning multiple information layers (Figure 1.1). From complex disease etiology point of view understanding of biological network functionality and dynamics is essential since complex diseases are caused by network perturbations [53]. Many studies identified marker genes and susceptibility loci, but functional information is still missing. This prompted emergence of network medicine field that applies systems biology and network science approaches to complex disease data [53,54]. The interactome networks can visualize and analyze many biological processes since nodes can be assigned to entities such as disease states, proteins, genes, patients while edges can represent physical interactions, transcriptional regulation, and similarity of disease phenotypes or gene expression profiles. In addition, networks incorporate intrinsic biological concept of robustness to perturbations. For example, transcriptional gene regulatory networks allow globally quantify impacts of introduced point mutations (perturbations) measured as changes of mRNA levels. Network medicine efforts

lead to promising results with better disease classifications and sub-typing via to integration of multiple omics data [55]. Perhaps future network medicine methods would allow disease classification not at organ resolution, but would also merge common disease etiologies allowing for more effective cure and patient prognosis [53].

Biological networks can be broadly classified into three main types described by Table 2.1 [56]. In this thesis work, we will mainly consider genetic networks. These networks represent expression regulation between genes. Regulatory proteins such as transcription factors (TFs) impact mRNA levels of their target genes. Such gene-gene networks are commonly referred to as gene regulatory networks (GRNs). The GRN can be seen as directed graph where a directed edge is drawn in regulatory protein expressed by  $i$ th gene ( $g_i$ ) impacts the expression of the  $j$ th gene ( $g_j$ ). Thus, the directionality of the TF→TG is important. GRNs can be interpreted as a bipartite graph with two sets of genes divided into TF and TG groups.

Epistatic networks also consider gene-gene interactions conditioned on a phenotype (e.g., disease phenotype) [57]. Such statistical networks are of a genetic type (see Section 1.3 and Figure 1.3). Nodes in these networks can be SNPs or genes, and edges are undirected but weighted. Weights indicate the strength of the association between SNP x SNP pairs and the phenotype. For example, the results of the GWAIS can be represented by an epistatic network after mapping of SNP x SNP interactions to the gene space and selection of appropriate significance threshold.

**Table 2.1:** the main types of biological networks

Type	Description
proteinic	nodes are proteins and edges represent functional links or physical binding (i.e. protein complexes) or enzymatic reactions (metabolic networks)
genetic	nodes are genes (i.e. gene symbols) and edges represent regulatory links between gene pairs
metabolic	nodes are proteins (enzymes) or chemical species and edges represent the speed of reaction (i.e. flux). A pair of nodes is equivalent to a chemical reaction converting one entity into another
ecologic	nodes are species and connections are predator-prey relationships

An important property of many biological networks is their scale-free topology [54]. Such network architecture confers robustness to stress, redundancy and evolutionary advantage. The redundancy

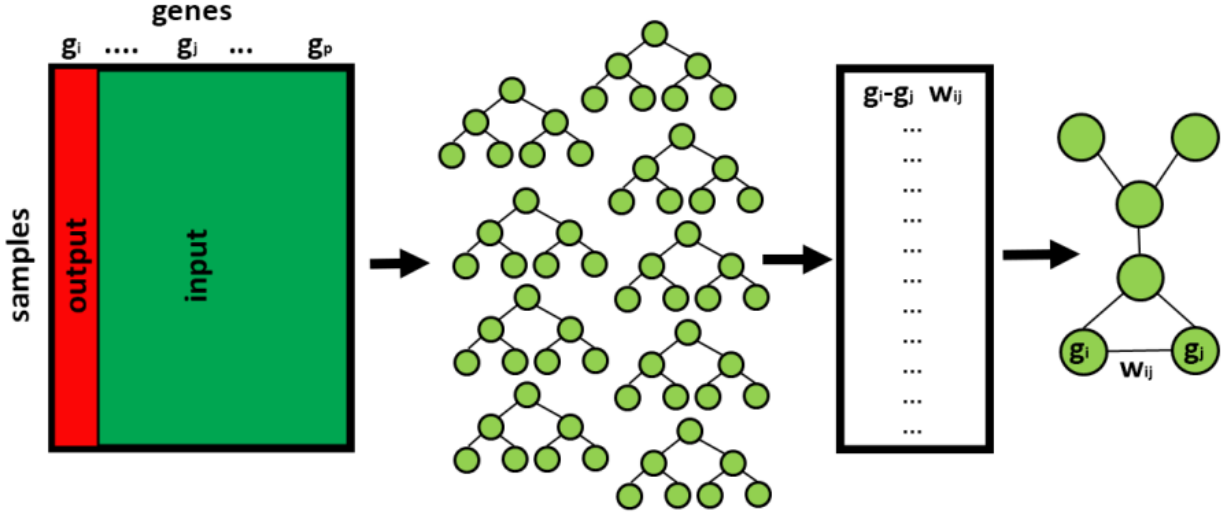
is important in cases when several pathways can provide similar functions and several routes can be taken to attain homeostatic conditions (e.g., such as adequate mRNA levels).

### 2.4.3. Network inference via tree ensembles

The power of tree-based methods can be combined with graph theory resulting in network inference. The tree-based methods like *CIFs* and *RFs* are used to rank potential gene-gene interaction edges between the  $i$ th and  $j$ th genes. Thus, a network inference algorithm consists in the assignment of weights ( $w$ ) to putative regulatory links that can be considered as feature selection problem. In the case of classical *CIF* and *RF* methods (see Chapter 5) these weights are summarized by variable importance measure (*VIM*). After assignment of a threshold value ( $w \leq t$ ) one can obtain a weighted directed network such as gene regulatory network (GRN). GRNs are transcriptional regulatory networks summarizing relationships existing between transcription factor and target gene(s) [58]. Figure 2.8 summarizes the essential network inference steps.

The inference of a GRN assumes that expression value of the  $i$ th gene is the result of the joint action of all other genes and factors spread across different omics data layers. Thus, these impacts on the expression of the  $i$ th gene can originate from multiple omics layers represented, for example, by genotype, expression, methylation and other data types. In this thesis, only the first three data types are being considered.

There exist many methods to infer GRNs some of which are based on sparse linear regression [59], Bayesian-networks [60], Mutual Information [61], Correlation frameworks [62]. Using correlation framework one can simply calculate correlation between transcription factors and target genes. For a more detailed review of GRN methods please refer to the supplementary section of [63].



**Figure 2.8:** Example of network inference via tree-based methods. The output is represented by  $i$ th gene ( $g_i$ ) and input by all other genes except the  $i$ th gene ( $g^{-i}$ ). The tree ensemble is built. The weight  $w_{ij}$  measures the strength of interaction between the  $i$ th and  $j$ th genes ( $g_i - g_j$ ). The  $w_{ij}$  value calculation depends on the chosen method and is summarized by *VIM*. The above procedure was repeated  $p$  times corresponding to the total number of genes by re-assignment of a new output. A list of all gene-gene interactions is obtained with corresponding  $w_{ij}$ . The pairwise gene-gene interaction list represents a gene network where  $w_{ij}$  is a weight of an edge.

*GENIE3* [64] is a tree-based method for GRN inference using tree ensembles from the expression data. In its core, the method uses *RF* algorithm to assign weights to all potential gene-gene regulatory edges without employment of permutation-based *VIM*. In contrast, *GENIE3* uses a total reduction of variance of the output variable in the child nodes after the split implementation of the parent node averaged over all trees in the ensemble. Thus, the final  $w$  is non-permutation-based *VIM* representing the strength of a given gene-gene interaction. Initially, *GENIE3* was applied on the synthetic expression data from DREAM4 and 5 challenges [63,65,66], but recently was extended to multi-source context integrating genotype and expression data. The two integrative methods were suggested including *GENIE3-SG-sep* and *GENIE3-SG-joint*. The *GENIE3-SG-sep* separately builds two different models (i.e. tree ensembles) for each data source while *GENIE3-SG-joint* builds a single joint model for both data sources (expression and genotype data) simultaneously. In simulated data of DREAM5 SysGenA challenge (Synapse:syn2820440) the

*GENIE3-SG-sep+product* method utilized a product of weights from expression and genotype data source ( $w_e$  and  $w_m$ ). The *GENIE3-SG-sep+product* showed top performance [67].

We explored similar in concept tree-based ensemble methods to infer GRNs relying on the *CIF* framework. These methods including *CIF<sub>mean</sub>* are presented in detail in Chapter 5. The proposed network inference methods simultaneously borrow ideas from both statistical and machine-learning disciplines. In addition, the performance of *CIF<sub>mean</sub>* and *CIF* variants is tested in simulated and real-life data.

The *CIF<sub>mean</sub>* method relies on the *CIF* framework introduced in Section 5.3.2 providing a greater modularity and stricter tree growth rules. Contrary to *RF*-based methods, the method allows to select and split node variables based on statistical significance. The original permutation-based *VIM* of original *CIF* implementation is computationally too intensive. Similar to *GENIE3*, *CIF<sub>mean</sub>* calculates the  $w_{ij}$  by averaging over all trees in the ensemble. Thus,  $w_{ij}$  is simply an average of  $p$ -values. The results of Chapter 5 indicate a competitive performance of *CIF<sub>mean</sub>* complemented with the relative ease of threshold definition and high scalability.

Similarly to *GENIE3-SG-sep*, *CIF<sub>mean</sub>* can build separate models for each data source. Nevertheless, *CIF<sub>mean</sub>* has a limitation. Contrary to *GENIE3-SG-joint*, *CIF<sub>mean</sub>* cannot simultaneously infer a joint model from multiple data sources.

## 2.5. Integration strategies

Integration is loosely defined term and can mean different things in different contexts. Oxley *et al.* defines integration as “the process of connecting systems into a larger system”[68]. *Integration* is often confused with *fusion* which is more closely related to concatenation of objects. Data fusion is a mapping of several objects onto a single one in an optimal fashion, whereas integration is more general and is defined as the process of connecting system components into a larger system which might have fused components in them [68]. Thus, data fusion and integration are not equivalent terms. Integration of omics data allows building systems which have components, interactions and functional states. Single omics data analysis is an assessment of interactions between all system

components at a particular layer including all cellular biochemical and molecular processes [69]. Examples of data layers include genomics, transcriptomic, proteomics and metabolomics. For example, if considering transcriptomic layer, the elementary components are expression levels of genes and interactions are association measures between expression profiles describing transcription regulation. Instead of looking at all within interactions, one can consider interactions with other layers. Exploration of complementarity between layers can be beneficial as demonstrated by better prediction of disease states and tumorigenesis mechanisms in [70]. Thus, a comprehensive integrative analysis should consider within and between interactions of multiple layers integrating them into one “*integrated system*”. The advantages of studying such integrated systems go beyond benefits of merging data allowing to infer holistic models considering all types of interactions between the information layers. In our view, integration strategies can be divided into three main strategies [69,71,72]

1. Change representation of each data source prior to analysis via generation of new constructs based on dimensionality principles. These can be represented by summary components (e.g., principal components, projections). For example in [33], a given gene can have information from several sources including mapped SNP, expression and functional data that can be aggregated into one feature such as *genomic region of interest* (ROI). The ROI method uses kernel PCA with clustering capabilities detailed in [33]. When different input data sources are considered, this new feature – ROI – is an integrated representation of multi-source data.
2. Fuse/concatenate input data accounting for structure between omics layers. This method is especially useful when considering interactions between different types of omics data of a different scale (e.g., genomic – transcriptomic). For example, after concatenation of genomic and transcriptomic data, *GENIE3-SG-joint* builds a single joint model represented by ensemble of trees incorporating variables originating from different data sources [67]. Another example, uses Bayesian approach to study joint impact of the genome (SNPs) and transcriptome (expression probes) on the phenotype trait after concatenation of genomic and transcriptomic data [73].
3. Analyze each data source separately generating a data-specific solution for each while ignoring inter-connections prior to obtaining an integrative result. For example, integrate results via concatenation of networks obtained separately from each data source generating

the final integrated network. The Similarity Network Fusion (SNF) networks [74] are a clear example of this strategy. Specifically, Wang *et al.* [74] showed how a fusion of several individual patient networks from each data source can result in a similarity network fusion allowing to better highlight different subject groups. Another example is *GENIE3-SG-sep* which calculates separately weights for each gene pair from each data source prior to integration step [67]. In addition, Section 6.3 presents *Regression2Net* method that utilizes this strategy.

In this thesis in Chapter 6 we propose solutions that are based on strategy 3. The integrated omics analysis can be subdivided into four steps 1) definition of biological problem and context; 2) data characterization and pre-processing; 3) integration analytics with validation and replication; 4) results interpretation. Each step has its own issues which are more accentuated compared to single omics data analysis. Chapters 6 and 7 provide suggestions to some of the issues with integrative omics analysis including replication, validation, visualization and collection of clinically relevant results, data storage and lack of adequate computational processing power [69].

## 2.6. Quality Control

The phrase ‘garbage in, garbage out’ is especially true for genotypic data as it might contain many hidden unknown confounding factors leading to increased false positives.

### 2.6.1. Genotypic Data Quality Control

In order minimize false positives due to detection of false positive associations, genotypic data from SNP arrays need to be cleaned and preprocessed prior to any statistical analysis [14]. Typical genotypic data quality control parameters include Hardy-Weinberg equilibrium (HWE), minor allele frequency (MAF) threshold, call rate of genotypes commonly referred to as Travemünde criteria [75]. Our lab developed the minimal GWAIS protocol for genome-wide association interaction studies [27] taking into account these criteria together with others including linkage-disequilibrium levels (LD). Chapter 3 will further explain the importance of proper quality control steps exemplified in the ankylosing spondylitis WTCCC2 data. Below we describe in detail the main criteria related to genotypic data QC.

### *Hardy-Weinberg Equilibrium (HWE)*

Hardy-Weinberg equilibrium (HWE) refers to a constant undisturbed genetic variation in a population linked to frequencies of alleles at a given locus. Assuming complete independence, a given set of alleles should follow HWE. Due to specific factors impacting segregation of alleles such as non-random mating, natural selection, random genetic drift the allele frequencies are not constant from generation to generation and, thus, deviate from equilibrium. Since HWE describes an idealized state of genetic variation, it is important to check whether each SNP follows HWE in order to detect issues associated with hidden population structure and to minimize false positive results. Generally, during quality control filtering, SNPs that do not pass HWE threshold are excluded from the further analysis. Typical HWE  $p$ -value threshold is generally set at  $10^{-4}$  [76]. At this threshold SNPs with HWE  $p$ -values  $< 10^{-4}$  would be eliminated from further analysis.

### *Minor Allele Frequency (MAF)*

The minor allele frequency (MAF) refers to the relative frequency of the least frequent allele in a population. Allele frequency is calculated based on how many times a given allele occurs in the population at a given locus divided by the total number of alleles (major and minor). Changes in allele frequencies might be indicative of population structure and non-random allele segregation deviating from HWE. If MAF of a given SNP is in the 0.005 - 0.05 range, such SNP is considered to be rare, otherwise common [77]. MAFs can be also expressed on the % scale.

### *Call Rate*

The genotypic call rate refers to the proportion of genotypes that are successfully genotyped out of the total number of samples. For example, given 100 samples and one SNP, the total of 95 genotypes are obtained. The call rate of the considered SNP is 0.95. Commonly the call rate threshold is set at 0.95 level [14]. Markers that do not meet such threshold are considered to have too many missing genotypes and are removed from further analysis.

### *Linkage-Disequilibrium (LD)*

Linkage disequilibrium refers to the non-random association of alleles at different loci. LD again is linked to allele frequencies and factors impacting segregation of alleles. The levels of LD are measured by the deviation coefficient ( $D$ ), normalized coefficient ( $D'$ ) and correlation coefficient ( $r$ ). If two loci are in equilibrium, their genotypes are independent and  $D = 0$ . The genetic distance between a given pair of loci impacts LD as chromosomes can be considered as mosaics affected by cross-over events (i.e. recombination). The further a given pair of loci is located, the higher are the chances of a recombination event and, subsequently, the lower is the LD. This is because the increased recombination rates tend to reduce dependency between loci pairs and, consequently, drive down the LD measure [78].

#### 2.6.2. Methylome Data Quality Control

DNA methylation data is accessible due to a relatively cheap cost of bisulfite sequencing (Bi-Seq) and methylation arrays processing. The Bi-Seq data acquisition and processing protocol steps include assessment of per-sequence and per-base quality plots, removal of adapter sequences from the 3' ends of the bisulfite treated DNA reads, followed by the alignment and methylation calling. Given a set of methylation arrays, the protocol steps are platform dependent. Nevertheless, all protocols share similar microarray normalization procedures (see Section 2.6.3). Briefly, the  $M$  or  $\beta$  values are calculated measuring the degree of methylation of the CpG sites followed by background normalization minimizing variation across arrays. In the case of Illumina 450k Methylation Arrays, there are several tools addressing the quality control needs including *minfi* R library [79]. For more details on common QC protocols for Bi-Seq and methylation arrays technologies please refer to [80] and [81].

#### 2.6.3. Expression Data Quality Control

Microarray expression data requires normalization procedures to ensure adequate levels of quality control thanks to minimization of technical and systematic biases. Some of these biases include different sample handling procedures imposed by different processing centers (batch effect), different microarray manufacturers, different amounts of starting material applied to microarray

chip, partial degradation of mRNA due to accidental contamination with nucleases, low probe sensitivity to poorly expressed genes coupled to high background to signal levels [82], non-specific probe binding, etc.

For example, when comparing identical microarrays with different amounts of starting sample material a scaling problem arises. The higher the amounts of total RNA, the higher are the signal intensities across all microarray probes. Fortunately, normalization and scaling procedures between and within arrays would ‘center’ expression data on a common median. One of the most popular normalization algorithms are Robust Multiarray Analysis (RMA) and MASS 5.0 implemented in *affy* R package [83]. In addition, the R package *limma* provides a complete quality control pipeline with input and normalization functions [84].

A typical RMA normalization protocol is generally characterized by the following main stages including subtraction of average probe background levels from raw probe intensity values, data transformation to log<sub>2</sub> scale, quantile normalization, and, finally, the linear model fitting leading to estimation of a normalized expression value of the probe [85]. RMA assumes a normal distribution of the global background signal. For a short review on microarray expression data processing and normalization please refer to [86].

## 2.7. References

1. Jones PA (2012) **Functions of DNA methylation: islands, start sites, gene bodies and beyond.** *Nat Rev Genet* 13: 484-492.
2. Suzuki MM, Bird A (2008) **DNA methylation landscapes: provocative insights from epigenomics.** *Nature Reviews Genetics* 9: 465-476.
3. Li E, Bestor TH, Jaenisch R (1992) **Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.** *Cell* 69: 915-926.
4. Lo K, Raftery AE, Dombek KM, Zhu J, Schadt EE, et al. (2012) **Integrating external biological knowledge in the construction of regulatory networks from time-series expression data.** *BMC systems biology* 6: 101.
5. Pertea M, Salzberg SL (2010) **Between a chicken and a grape: estimating the number of human genes.** *Genome Biol* 11: 206.
6. Li Q, Seo J-H, Stranger B, McKenna A, Pe’er I, et al. (2013) **Integrative eQTL-based analyses reveal the biology of breast cancer risk loci.** *Cell* 152: 633-641.
7. Loo L, Cheng I, Tiirikainen M, Lum-Jones A, Seifried A, et al. (2012) **cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue.** *PloS one* 7: e30477.

8. Fish A, Capra JA, Bush WS (2015) **Are Genetic Interactions Influencing Gene Expression Evidence for Biological Epistasis or Statistical Artifacts?** *bioRxiv*: 020479.
9. Kadauke S, Blobel GA (2009) **Chromatin loops in gene regulation.** *Biochim Biophys Acta* 1789: 17-25.
10. Kapur K, Schüpbach T, Xenarios I, Kotalik Z, Bergmann S (2011) **Comparison of strategies to detect epistasis from eQTL data.** *PloS one* 6: e28415.
11. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) **Five years of GWAS discovery.** *The American Journal of Human Genetics* 90: 7-24.
12. Ott J, Kamatani Y, Lathrop M (2011) **Family-based designs for genome-wide association studies.** *Nat Rev Genet* 12: 465-474.
13. Hindorff LA, Junkins HA, Hall P, Mehta J, Manolio T (2011) **A catalog of published genome-wide association studies.** *National Human Genome Research Institute*.
14. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, et al. (2010) **Data quality control in genetic case-control association studies.** *Nature protocols* 5: 1564-1573.
15. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) **Finding the missing heritability of complex diseases.** *Nature* 461: 747-753.
16. Van Steen K (2012) **Travelling the world of gene-gene interactions.** *Brief Bioinform* 13: 1-19.
17. Mackay TF (2014) **Epistasis and quantitative traits: using model organisms to study gene-gene interactions.** *Nature Reviews Genetics* 15: 22-33.
18. Bellman R, Kalaba R (1959) **A mathematical theory of adaptive control processes.** *Proceedings of the National Academy of Sciences of the United States of America* 45: 1288.
19. Van Steen K (2011) **Travelling the world of gene-gene interactions.** *Briefings in bioinformatics*: bbr012.
20. Panagiotou OA, Ioannidis JP (2012) **What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations.** *International journal of epidemiology* 41: 273-286.
21. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.** *The American Journal of Human Genetics* 87: 325-340.
22. Cordell HJ (2002) **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Human molecular genetics* 11: 2463-2468.
23. Wu J, Devlin B, Ringquist S, Trucco M, Roeder K (2010) **Screen and clean: a tool for identifying interactions in genome-wide association studies.** *Genetic epidemiology* 34: 275-285.
24. Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, et al. (2011) **Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise.** *Ann Hum Genet* 75: 78-89.
25. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, et al. (2014) **Detection and replication of epistasis influencing transcription in humans.** *Nature* 508: 249-253.
26. Zhang X, Huang S, Zou F, Wang W (2011) **Tools for efficient epistasis detection in genome-wide association study.** *Source code for biology and medicine* 6: 1.
27. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet* 133: 1343-1358.
28. Grange L (2014) **Thesis: epistasis in genetic susceptibility to infectious diseases: comparison and development of methods application to severe dengue in Asia:** Paris 7.
29. Calle ML, Urrea V, Malats i Riera N, Van Steen K (2008) **MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data.**
30. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *The American Journal of Human Genetics* 69: 138-147.

31. Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, et al. (2013) **An efficient algorithm to perform multiple testing in epistasis screening.** *BMC Bioinformatics* 14: 138.
32. Westfall PH, Young SS (1993) **Resampling-based multiple testing: Examples and methods for p-value adjustment:** John Wiley & Sons.
33. Fouladi R, Bessonov K, Van Lishout F, Van Steen K (2015) **Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis.** *Hum Hered* 79: 157-167.
34. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) **Classification and regression trees:** CRC press.
35. Strobl C, Malley J, Tutz G (2009) **An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.** *Psychological methods* 14: 323.
36. Quinlan JR (1993) **C 4.5: Programs for machine learning.** *The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann, c1993* 1.
37. Loh WY (2011) **Classification and regression trees.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1: 14-23.
38. Hill T, Lewicki P, Lewicki P (2006) **Statistics: methods and applications: a comprehensive reference for science, industry, and data mining:** StatSoft, Inc.
39. Geurts P (2002) **Contributions to decision tree induction: bias/variance tradeoff and time series classification.**
40. Breiman L (1996) **Bagging predictors.** *Machine learning* 24: 123-140.
41. Breiman L (1996) **Out-of-bag estimation.** Citeseer.
42. Dietterich TG (2000) **Ensemble methods in machine learning.** Multiple classifier systems: Springer. pp. 1-15.
43. Breiman L (2001) **Random forests.** *Machine learning* 45: 5-32.
44. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) **Conditional variable importance for random forests.** *BMC bioinformatics* 9: 307.
45. Hothorn T, Hornik K, Zeileis A (2006) **Unbiased recursive partitioning: A conditional inference framework.** *Journal of Computational and Graphical statistics* 15: 651-674.
46. Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, et al. (2012) **SNP interaction detection with Random Forests in high-dimensional genetic data.** *BMC Bioinformatics* 13: 164.
47. Boulesteix AL, Janitza S, Kruppa J, König IR (2012) **Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics.** University of Munich. 129 129.
48. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T (2007) **Bias in random forest variable importance measures: Illustrations, sources and a solution.** *BMC bioinformatics* 8: 25.
49. Hothorn T, Hornik K, Strobl C, Zeileis A (2010) **Party: A laboratory for recursive partytioning.**
50. Hothorn T, Hornik K, Zeileis A **ctree: Conditional Inference Trees.**
51. Lian-Ming Z, Xiao-Heng D, Jian-Ping Y, Xiang-Sheng W (2011) **Degree and connectivity of the Internet's scale-free topology.** *Chinese Physics B* 20: 048902.
52. Albert R (2005) **Scale-free networks in cell biology.** *Journal of cell science* 118: 4947-4957.
53. Silverman EK, Loscalzo J (2012) **Network medicine approaches to the genetics of complex diseases.** *Discov Med* 14: 143-152.
54. Barabasi AL, Oltvai ZN (2004) **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 5: 101-113.
55. Loscalzo J (2007) **Association studies in an era of too much information: clinical analysis of new biomarker and genetic data.** *Circulation* 116: 1866-1870.
56. Proulx SR, Promislow DE, Phillips PC (2005) **Network thinking in ecology and evolution.** *Trends Ecol Evol* 20: 345-353.
57. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, et al. (2011) **Characterizing genetic interactions in human disease association studies using statistical epistasis networks.** *BMC bioinformatics* 12: 364.

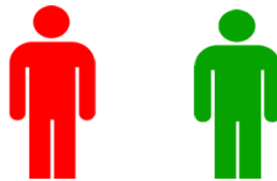
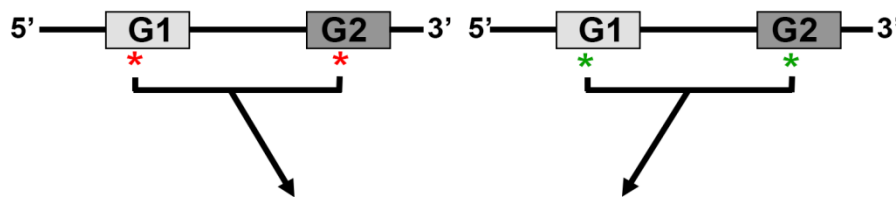
58. Teichmann SA, Babu MM (2004) **Gene regulatory network growth by duplication.** *Nature genetics* 36: 492-496.
59. Kim SY, Imoto S, Miyano S (2003) **Inferring gene networks from time series microarray data using dynamic Bayesian networks.** *Briefings in bioinformatics* 4: 228.
60. Young WC, Raftery AE, Yeung KY (2014) **Fast Bayesian inference for gene regulatory networks using ScanBMA.** *BMC Syst Biol* 8: 47.
61. Meyer PE, Lafitte F, Bontempi G (2008) **minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information.** *BMC Bioinformatics* 9: 461.
62. Liu ZP (2015) **Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data.** *Curr Genomics* 16: 3-22.
63. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. (2012) **Wisdom of crowds for robust gene network inference.** *Nat Methods* 9: 796-804.
64. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) **Inferring regulatory networks from expression data using tree-based methods.** *PLoS One* 5.
65. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) **Revealing strengths and weaknesses of methods for gene network inference.** *Proc Natl Acad Sci U S A* 107: 6286-6291.
66. Marbach D, Schaffter T, Mattiussi C, Floreano D (2009) **Generating realistic in silico gene networks for performance assessment of reverse engineering methods.** *J Comput Biol* 16: 229-239.
67. Huynh-Thu VA, Wehenkel L, Geurts P (2013) **Gene regulatory network inference from systems genetics data using tree-based methods.** *Gene Network Inference-Verification of Methods for Systems Genetics Data.*
68. Oxley ME, Thorsen SN (2004) **Fusion or Integration: What's the Difference?** : DTIC Document.
69. Van Steen K, Malats N (2014) **Perspectives on Data Integration in Human Complex Disease Analysis.** In: Wang B, Li R, Perrizo W, editors. *Big Data Analytics in Bioinformatics and Healthcare*. 1 ed: IGI Global. pp. 284-322.
70. Nibbe RK, Koyuturk M, Chance MR (2010) **An integrative -omics approach to identify functional sub-networks in human colorectal cancer.** *PLoS Comput Biol* 6: e1000639.
71. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) **Methods of integrating data to uncover genotype-phenotype interactions.** *Nature Reviews Genetics* 16: 85-97.
72. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, et al. (2009) **Data integration in genetics and genomics: methods and challenges.** *Hum Genomics Proteomics* 2009.
73. Fridley BL, Lund S, Jenkins GD, Wang L (2012) **A Bayesian integrative genomic model for pathway analysis of complex traits.** *Genetic epidemiology* 36: 352-359.
74. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) **Similarity network fusion for aggregating data types on a genomic scale.** *Nat Methods* 11: 333-337.
75. Ziegler A (2009) **Genome-wide association studies: quality control and population-based measures.** *Genetic epidemiology* 33: S45-S50.
76. König IR, Loley C, Erdmann J, Ziegler A (2014) **How to include chromosome x in your genome-wide association study.** *Genetic epidemiology* 38: 97-103.
77. Consortium IH (2010) **Integrating common and rare genetic variation in diverse human populations.** *Nature* 467: 52-58.
78. Li N, Stephens M (2003) **Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.** *Genetics* 165: 2213-2233.
79. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, et al. (2014) **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 30: 1363-1369.
80. Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M (2006) **DNA methylation: bisulphite modification and analysis.** *Nature protocols* 1: 2353-2364.
81. Laird PW (2010) **Principles and challenges of genome-wide DNA methylation analysis.** *Nature Reviews Genetics* 11: 191-203.

82. Sui Y, Zhao X, Speed TP, Wu Z (2009) **Background adjustment for DNA microarrays using a database of microarray experiments.** *J Comput Biol* 16: 1501-1515.
83. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) **affy—analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 20: 307-315.
84. Smyth GK (2005) **Limma: linear models for microarray data.** Bioinformatics and computational biology solutions using R and Bioconductor: Springer. pp. 397-420.
85. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 4: 249-264.
86. Wu Z (2009) **A review of statistical methods for preprocessing oligonucleotide microarrays.** *Statistical methods in medical research* 18: 533-541.

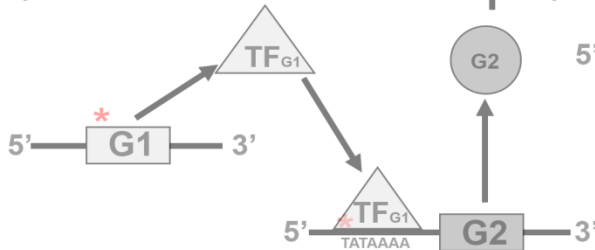
# Chapter 3: Genome-Genome Interactions

## The impact of protocol changes for genome-wide association SNP x SNP interaction

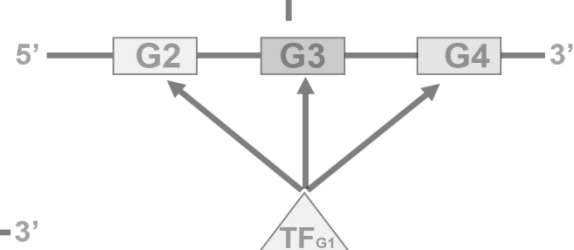
a) GWAI



b) *trans/cis* eQTL



c) GRN inference



Related publication:

Bessonov K, Gusareva ES, Van Steen K (2015) A cautionary note on the impact of protocol changes for genome-wide association SNP -SNP interaction studies: an example on ankylosing spondylitis. Hum Genet 134: 761-773.



## 3. The impact of protocol changes for genome-wide association SNP x SNP interaction

### 3.1. Chapter summary

In this chapter we empirically explore genome-wide SNP x SNP interaction (GWAIS) protocols. Specifically, we cover practical aspects of GWAIS analysis with MB-MDR and BOOST algorithms. The impact of key parameters on the final outcomes is investigated and compared across the 10 protocols focusing on MB-MDR and BOOST as the leading analytic strategies. Whereas BOOST represents a classic approach of SNP x SNP interaction discovery via logistic regression, MB-MDR represents data dimensionality reduction approaches that do not necessarily make assumption about genetic mode of inheritance. A better understanding about the effect of changing a few parameters in analysis protocols (e.g., regarding allowable LD between SNPs or regarding prior biological SNP filters) is needed to put significant findings for epistasis in the “correct” perspective. Apart from providing a contribution to fine-tune existing analysis for genome-wide association (GWA) large-scale epistasis screening, we illustrate on real-life data for WTCCC2 – ankylosing spondylitis (AS). Here, we investigate the effect of different routes in the GWAIS protocol and obtain novel insights into the pathology of AS.

**Problem:** Genome-wide association interaction studies (GWAIS) have increased in popularity. Yet to date, no standard GWAIS protocol exists. In practice, any GWAIS workflow involves making choices about quality control strategy, SNP filtering, linkage disequilibrium (LD) pruning, analytic tool to model or to test for genetic interactions. Each of these can have an impact on the final epistasis findings and may affect their reproducibility in follow-up analyses. In most cases the degree of these impacts are largely unknown. Choosing an analytic tool is not straightforward, as different such tools exist and current understanding about their performance is based on often very particular simulation settings. In the present study, we wish to create awareness for the impact of (minor) changes in a GWAIS analysis protocol can have on final epistasis findings. In particular, we investigate the influence of marker selection and marker prioritization strategies, LD pruning and the choice of epistasis detection analytics on study results, giving rise to 10 GWAIS protocols.

Discussions are made in the context of the ankylosing spondylitis (AS) data obtained via the Wellcome Trust Case Control Consortium (WTCCC2).

**Results:** As expected, the largest impact on AS epistasis findings is caused by the choice of marker selection criterion, followed by marker coding (co-dominant or additive), followed by the LD pruning. In MB-MDR, co-dominant coding of main effects is more robust to the effects of LD pruning than additive coding. We are able to reproduce previously reported epistasis involvement of *HLA-B* and *ERAP1* in AS pathology. In addition, our results suggest involvement of *MAGI3* and *PARK2*, responsible for cell adhesion and cellular trafficking. Gene Ontology (GO) biological function enrichment analysis across the 10 considered GWAIS protocols suggests a possible association of AS to Central Nervous System (CNS) malfunctions, specifically, nerve impulse propagation and in neurotransmitters metabolic processes.

**Keywords:** Genome-wide association interaction studies (GWAIS), epistasis, protocol adoption, ankylosing spondylitis

## 3.2. Introduction

High-throughput technologies give access to unprecedentedly vast amounts of data such as Single Nucleotide Polymorphisms (SNPs). In Genome Wide Association Studies (GWAS), thousands of these are scanned for their potential association with traits of interest, such as a disease status. Hard to disentangle are complex traits which assume an intricate interplay between genetic, environmental and/or many other unknown factors. For these traits added benefits can be obtained by using methods that account for biological and statistical interactions, rather than by adopting strategies that analyze each SNP at a time. This is the subject of Genome-wide association interaction studies (GWAIS), which usually focus on pairwise SNP-SNP interactions. It is believed that GWAIS can lead to novel or improved clinical and biologically relevant hypotheses.

Many strategies exist to carry out a GWAIS, such as those based on generalized linear regression models (GLM), BOOST [1], Dimensionality Reduction (MB-MDR) [2], BiForce [3], Bayesian Models (e.g., BEAM) [4] and several others [5-7]. For extensive reviews, please refer to [6,8,9]. All of these methods have their pros and cons, but the problems or hurdles encountered during the analysis are largely overlapping. Common hurdles to overcome include dealing with high dimensionality, handling a huge multiple testing problem, limiting computation time (when assessing statistical significance), and controlling false positive rates [6]. Unfortunately, often when novel GWAIS analysis methods are introduced the impact on epistasis findings of changes in the GWAIS protocol are given limited attention. Some examples of key protocol parameter changes related to marker filtering/prioritization, LD thresholds in marker pruning, *a priori* assumptions about operating two-locus inheritance models, main effects correction. It is essential to differentiate between global two-locus testing (i.e. not differentiating between main effects and interaction effects) and specific interaction testing (i.e. testing for the interaction between two loci itself, above and beyond the main effects). Specific interaction testing requires making adjustments for lower-order effects, and hence proposing a particular encoding scheme for lower-order effects. Several authors have commented upon the limitations of an additive encoding scheme for SNPs in SNP x SNP interaction studies and recommended co-dominant coding [10].

In this study, we investigated the impact on final epistasis results of changing one or more parameter settings in a GWAIS protocol, leading to 10 interesting strategies (Figure 3.1 and Table

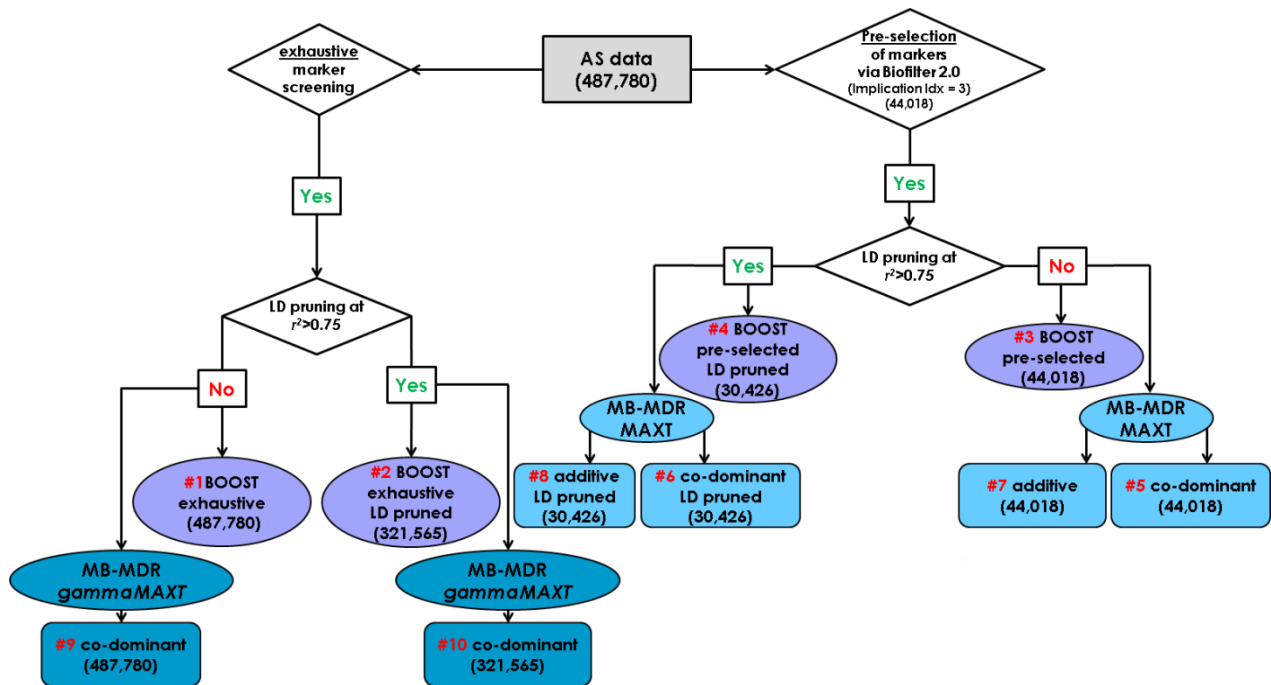
S3.1). These strategies are motivated by prior theoretical work [2,11,12] and recent epistasis tool developments. As a benchmark protocol, we took the one proposed by [8]. As analytic tools we chose BOOST [13], motivated by its popularity and computational efficiency due to a Boolean data representation, and MB-MDR (e.g., [2]), because of its non-parametric nature regarding epistasis models and its ability to correct for confounders or lower-order effects. In brief, BOOST handles binary traits and fits a full generalized linear model with main SNP effects (2 degrees of freedom (df) for each main effect) and SNP-SNP interaction effects (4 df). Significant (specific) interactions are identified via a Log-Likelihood Ratio Test (LRT) based on 4 df. The Bonferroni correction is proposed as a multiple testing corrective measure. In contrast, MB-MDR handles binary, continuous, and censored traits, and first carries out a dimensionality reduction procedure while pooling risk-alike multi-locus genotype combinations together. Its final test statistic contrasts high risk versus low risk multi-locus genotypes. While correcting for multiple testing, the significance is assessed via the resampling-based strategy proposed by [14]. For additional details about MB-MDR and BOOST, we refer to [1,2,12,15]. To achieve our goal, we used real-life ankylosing spondylitis (AS) data from the Wellcome Trust Case Control Consortium (WTCCC2). Ankylosing spondylitis (AS) is a common form of inflammatory arthritis occurring in approximately 1 to 14 out of 1,000 adults globally [16]. Apart from confirming previously known AS associated genes [17,18], we will show that combining different protocols may give new insights into disease pathology.

### **3.3. Methods**

#### **3.3.1. Data Quality Control**

Approved access to Wellcome Trust Case Control Consortium (WTCCC2) data, in particular via EBI accession no. EGAS00000000104, EGAD00010000150, EGAD00000000024 and EGAD00000000022, resulted in a dataset composed of 2005 ankylosing spondylitis (AS) cohort samples, and 3000 British 1958 Birth Cohort (BC) and 3000 National Blood Donors (NBS) Cohort samples. The 1788 cases were of British Caucasian origin recruited by Nuffield Orthopedic Centre, Oxford and Royal National Hospital for Rheumatic Diseases, Bath. The first batch of case samples were genotyped on an Illumina 670k platform; the last two batches of control samples were

genotyped on an Illumina 1.2M platform. No imputation was done for these genotypes. We used PLINK [19] to select 6,587 subjects (1788 cases plus 4799 controls), 3409 of which were male and 2864 female, and 487,780 SNPs, according to criteria described in [17]. Briefly, SNPs showing  $MAF < 0.01$ , Hardy-Weinberg  $p$ -values  $< 5 \times 10^{-20}$  and SNPTEST information measure  $< 0.975$  were excluded. The dataset inflation factor ( $\lambda$ ) was estimated as 1.02917. The QC-ed genotype data were stored in GEN format and were converted to PED and MAP files using GTOOL from Oxford University, UK [20].



**Figure 3.1:** Summary of 10 GWAS protocols included in this study and applied to AS data, the ankylosing spondylitis dataset from [17]. The number of SNPs retained at each step is shown in parenthesis. The bottom nodes refer to GWAS protocol abbreviations and chosen parameters, following protocol components as described in [21] [8]. The abbreviations *additive* and *co-dominant* refer to SNP main effects correction encodings in MB-MDR (see [10]). The abbreviation *gammaMAXT* and *MAXT* refer to the SNP x SNP interaction significance assessment strategies implemented in MB-MDR [15] (see Methods).

### 3.3.2. Additional data handling

Depending on the GWAIS protocol of choice, additional data manipulations were required, such as marker prioritization or LD pruning (Figure 3.1). We prioritized markers with the Biofilter 2.0 software developed by Ritchie *et al.* [22]. The Biofilter 2.0 uses a list of public biological databases (sources) such as KEGG, BioGRID, MINT, via the Library of Knowledge Integration (LOKI), to generate pairwise gene-gene interaction models [13]. No disease-specific information was used, but available knowledge about gene-gene interactions from different biological resources called by Biofilter 2.0 [22]. The advantage of such an approach is an 11-fold reduction of the original marker set, without selection bias introduction towards a particular disease. The disadvantage of *any* pre-filtering method is that useful information may be disregarded and biologically relevant SNPs removed from further analysis protocols. In practice, taking the 487,780 SNPs from [17] as a starting point, we applied Biofilter 2.0 with a minimum implication index threshold of 3, meaning that at least 3 data sources confirmed the associated gene-gene interaction. This resulted in the generation of 8,288 gene-gene models and a set of 44,018 unique SNPs (Figure 3.1).

To reduce the number of tests and the number of false positives based on genomic proximity (for instance, redundant epistatic SNP pairs), some GWAIS protocols involve LD filtering or pruning (Figure 3.1). As motivated and recommended by [21], we adopted a rather mild pruning threshold of  $r^2 > 0.75$ , still allowing for moderate LD but removing strong LD. Pruning at  $r^2 > 0.75$  threshold implies that every SNP pair in the pruned dataset has an  $r^2$  of at most 0.75. The proposed threshold offers a balance between power gain and false positives due to high LD. In practice, LD-pruning was performed considering the sliding windows of size 50 (i.e. 50 markers) with window increments of 1 marker. For any pair of markers under testing whose  $r^2 > 0.75$ , the first marker of the pair was discarded, as implemented in SVS Version 7.5 (Golden Helix, Inc.) [23]. After LD pruning, the original marker dataset reduced from 487,780 to 321,565 markers. After LD pruning, the biofiltered data (Biofilter 2.0) reduced from 44,018 to 30,426 markers (Figure 3.1).

### 3.3.3. Interaction testing

To test for interactions we used two software tools: BOOST [13] and MB-MDR [2]. We extended the original BOOST algorithm as it did not deal with missing genotypes and so as to properly adjust

the number of degrees of freedom (df) in case less than 3 genotypes was observed for a marker. Our implementation of BOOST was coded in C++, and was coined Epishell. For more details we refer to [11,24]. Notably, a similar adaption of BOOST was implemented in the PLINK software (PLINK version 1.9, called via “--fast-epistasis boost”). In practice, for the MB-MDR methodology, we used the algorithms implemented in MBMDR version 3.0.2 [15] that provides several advantages over classic MDR [25] or BOOST, such as the ability to analyze different trait types within the same framework, as well as non-parametric model free testing for two or three-order interactions while adjusting for lower order effects or relevant confounders. Since MBMDR versions 2.0 – 4.1.0 require significant computational resources to run on a genome-wide scale, we were not able to use these MB-MDR versions on unfiltered data, at the time of analysis. Hence, in the initial study, all MB-MDR based protocols (Figure 3.1) were implemented on a reduced dataset via Biofilter 2.0. The default main effects correction in MB-MDR is a co-dominant one. As was mentioned in [26], it is important to correct for main effects in a co-dominant way to avoid false epistasis signals. To allow for exhaustive epistasis screening with MB-MDR (protocols #9 and #10 - Figure 3.1), we used its *gammaMAXT* implementation introduced in MBMDR version 4.2.2 as in [27]. The latter has advantages over the original *MAXT* algorithm [15] when more than  $10^5$  SNP pairs need to be investigated in a large-scale epistasis study.

Results obtained from either one of the 10 GWAIS protocols included in this study were compared to results obtained in the reference study [17]. In particular, as statistical interactions may be indicative for important main effects [28], we compared SNPs derived from significant SNP pairs to the list of 49 SNPs in Supplementary Table S2 of [17] that passed quality control in their replication analysis. Also, significant SNP pairs obtained in this work were compared to the reference panel of 102 SNP-SNP pairs tabulated in Supplementary Table S5 of [17]. The latter table lists all considered SNP pairs for interaction testing, using an additive x additive term in a logistic regression model (i.e. additive encoding of SNP main effects and interaction).

### 3.3.4. Assessing consistencies between protocols

The overlap between GWAIS protocols (Figure 3.1) in identifying the same significant SNP pairs was graphically presented via the Euler diagram (Figure 3.2) with the software VennMaster 0.38 [29]. For each of the SNP pairs tested, ranks were computed, for each protocol separately, with

rank 1 assigned to the SNP pair with the smallest multiple testing corrected  $p$ -value. Then, SNP pairs that were common to each protocol were retained, in order to be able to compare exhaustive with non-exhaustive protocols. A total of 207 SNP pairs were retained. These are listed in online Table S3.4, together with their associated protocol-specific  $p$ -values, and were subsequently used to calculate “distances” between protocols. In particular, we calculated the squared Euclidean distance between 10 GWAIS protocols using 10 input vectors containing 207 ranks each. These 207 ranks for each protocol corresponded to relative positions of the common 207 SNP pairs amongst all ordered SNP pairs (from highest to lowest significance). For example, the ranks for the *rs12026423* x *rs7528311* pair in protocols 1 to 10 were 232, 2300, 97, 61, 259, 151, 59892, 43598, 15807, and 5418 respectively. We used *complete linkage* cluster agglomeration with *hclust()* to build a dendrogram (hierarchical tree) [30] (Figure 3.3). The use of SNP pair ranks coupled with hierarchical clustering allows an unbiased qualitative comparison of the top findings derived via different GWAIS protocols.

In addition, to assess the effects of MAFs on top findings in each protocol, we selected the top 1000 SNP pairs for each GWAIS protocol. We subsequently defined the following MAF classes or bins, using interval notations: 1) (0-0.05) (MAF<0.05; less common minor allele); 2) [0.05-0.10) ( $0.05 \leq \text{MAF} < 0.10$ ; moderate occurrence of the minor allele); 3) [0.10-0.50) ( $0.10 \leq \text{MAF} < 0.50$ ; rather common minor allele). Two-dimensional bins were defined by combining the aforementioned three 1-dimensional bins as follows: 1) (0-0.05)/(0-0.05); 2) [0.05-0.10)/(0-0.05); 3) [0.10-0.50)/(0-0.05); 4) [0.05-0.10)/[0.05-0.10); 5) [0.05-0.10)/[0.05-0.10); 6) [0.10-0.50)/[0.10-0.50). Note that for any SNP pair falling into one of these six 2-dimensional bins, the MAF of the first SNP in the pair will be larger or equal than the MAF of the second SNP in the pair, unless perhaps when both SNPs belong to the same one-dimensional bin.

### 3.3.5. Biological relevance

The SNP to gene symbol annotation (when possible) was done using SCAN – a SNP and CNV Annotation Database [31]. The SCAN database accepts a list of SNPs, maps them to genomic coordinates and outputs corresponding gene symbols, provided that the SNP is located within a gene coding region, which is helpful in assessing putative biological function and context. We then performed GO enrichment analyses [32] on the top 1000 most significant SNP pairs, by GWAIS

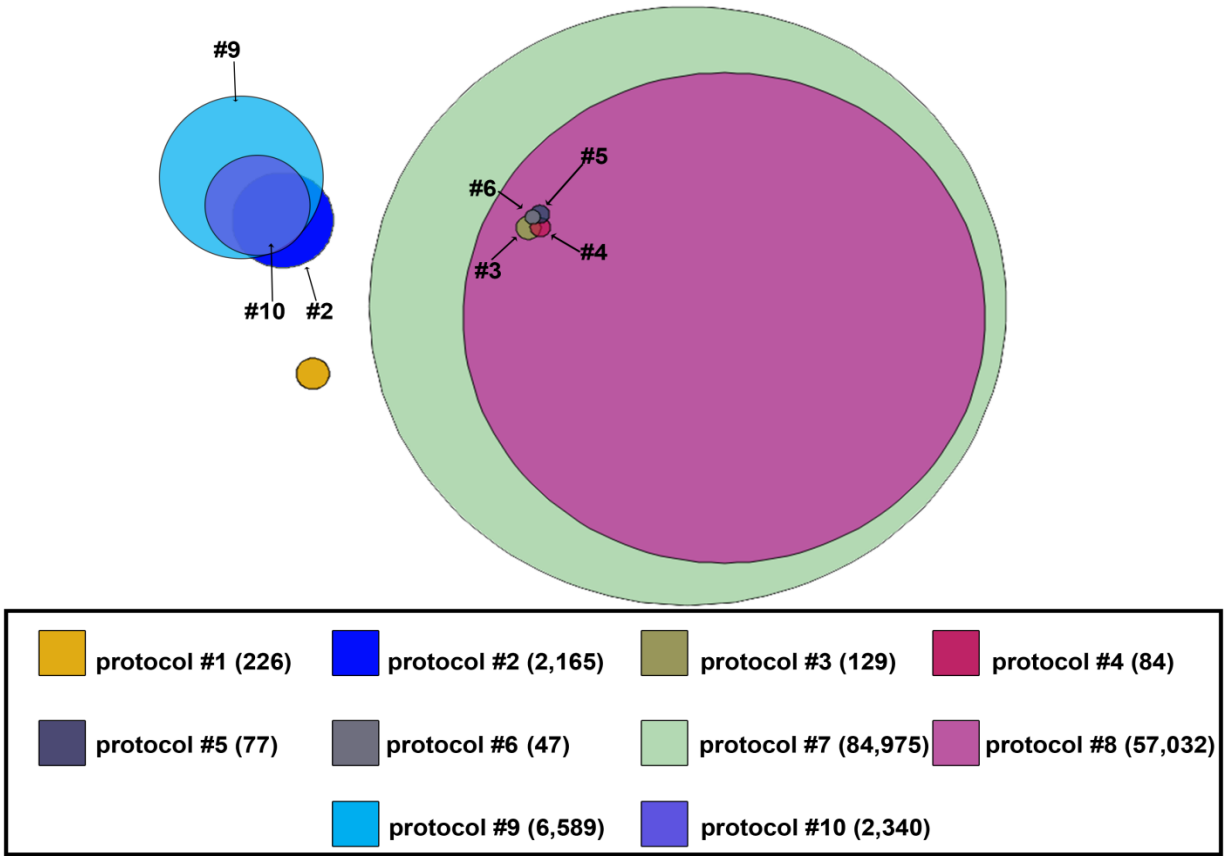
protocol. In practice, we used the *topGO* library in R that takes into account the GO graph structure and removed nodes (GO terms) that had a low number of annotated genes, i.e. less than 10 [33,34]. The *weight01* algorithm was chosen based on the author's recommendations and due to benefits incorporation from the *elim* and *weight* algorithms [33]. Significance of each GO term, per protocol, was based on Fisher's exact test. Overall significance across all protocols was assessed via Fisher's combined probability test at a significance level of 0.05.

## 3.4. Results

### 3.4.1. Consistency between interaction results derived from different GWAIS protocols

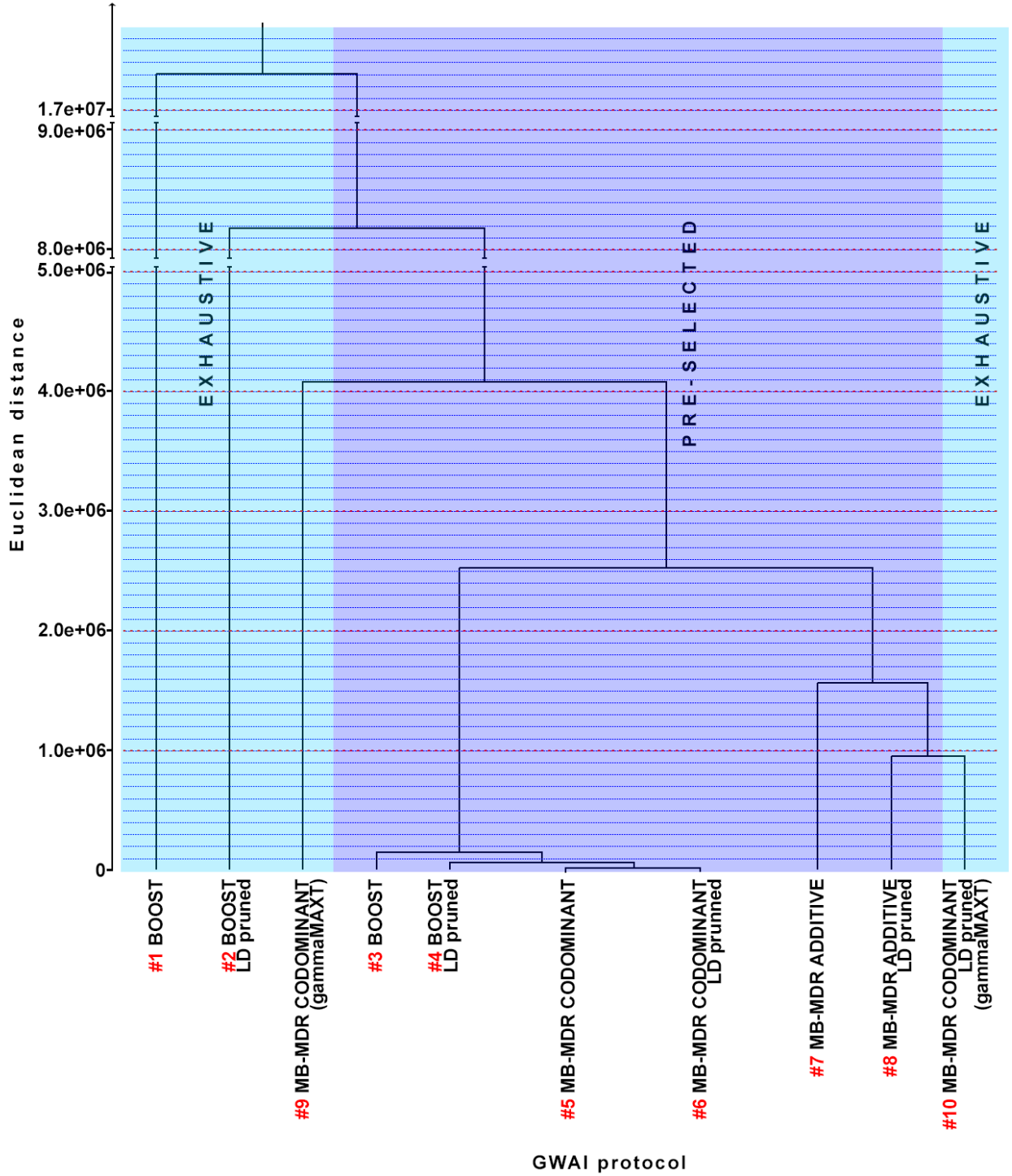
A graphical representation, showing the overlap of significant findings between considered GWAIS protocols is presented in Figure 3.2. The significant SNP pairs (multiple testing corrected) retrieved via GWAIS protocol #1-#10 (Figure 3.1) are tabulated in online Table S3.3.

The largest number of significant SNP pairs were obtained for protocols that use additive encoded corrections for main effects (protocols #7, #8). Over 2000 significant pairs were detected with an exhaustive implementation of BOOST on LD-pruned data (protocol #2). The number of significant SNP pairs reduces significantly when BOOST is used exhaustively on un-pruned data (protocol #1; 226 pairs).



**Figure 3.2:** Euler diagram capturing significant SNP pairs identified in each of the 10 GWAS protocols (Table S3.1). Each circle represents a set of the significant SNP pairs in the corresponding GWAS protocol. Protocol numbers match the protocol referencing used in Figure 3.1.

All other protocols identified less than 130 significant epistasis signals; the most liberal is protocol #3 (BOOST on filtered data), the most conservative is protocol #6 (MB-MDR on biofiltered and LD-pruned data), also using a co-dominant encoding scheme to correct the interaction testing for lower order SNP effects. Furthermore, only few of the findings obtained via exhaustive protocols (BOOST, #1-#2) were retrieved via protocols that first biofiltered the data (protocols #3-#8). With the same protocol for LD pruning on biofiltered data, both BOOST and MB-MDR in co-dominant main effects correction mode, gave partially overlapping results (Figure 3.2). In effect, over 97% of significant SNP x SNP interactions identified via MB-MDR protocols #5 and #6 were identified in BOOST protocols #3 and #4, respectively (Figure 3.2 and Table 3.3).



**Figure 3.3:** Consistency between GWAS protocols based on 207 common SNPs. Each SNP pair has a protocol-specific rank, which is stored in a protocol-specific vector. The dendrogram shows the distance between protocols, obtained via hierarchical clustering of 10 vectors (referring to the 10 GWAS protocols included in this study) of length 207 and the Euclidean distance measure. The Euclidean distances themselves are listed in Table 3.2.

Via hierarchical clustering (see Methods for details), the largest distance between protocols (i.e. the smallest overlap between top findings, not necessarily significant) was obtained for exhaustive screening protocols: protocol #1 - BOOST without pruning and protocol #2 – BOOST applied on LD-pruned data (Figure 3.3). The effect of LD in BOOST applications is less pronounced when data were first biofiltered. Actually, the smallest distance between protocols was observed between protocols #3 (BOOST without LD pruning) and #4 (BOOST applied to LD-pruned data). In general, the effect of LD on SNP pair rankings seems to be smaller in non-exhaustive protocols as compared to the exhaustive protocols considered. This is true for the exhaustive protocols #9 and #10 where LD pruning had a very significant impact on final results rankings (Figure 3.3). The second smallest distances observed between protocols was between #5 and #6 (MB-MDR with co-dominant correction of lower-order effects on the pre-selected data) and between #7 and #8 (MB-MDR with additive encoding of main SNP effects on the pre-selected data). Within non-exhaustive screening protocols (#3–#8), analyses that used an additive encoding to adjust for SNP main effects while testing for interactions stood out; all protocols involving epistasis detection analytics with co-dominant encoding schemes of some sort clustered together (Figure 3.3). The *gammaMAXT* exhaustive protocols involving additive main effects correction provided very dissimilar result to all considered protocols sharing no common SNP x SNP pairs (data not shown). Notably, MB-MDR *gammaMAXT* applied in non-exhaustive settings provided identical results to MB-MDR with *MAXT* multiple testing correction (i.e. distance zero – data not shown). A closer look at the overlapping significant SNP pairs across all 10 GWAIS protocols, reveals that only 3 out of 207 SNP pairs (*rs12026423/rs7528311*, *rs11964796/rs13194019* and *rs13194019/rs1784607*) met statistical significance at  $\alpha=0.05$ , according to at least one GWAIS protocol (Table S3.4).

We furthermore investigated whether any of the 49 main effects SNPs reported in [17] were supported by our SNP-SNP interaction results across the 10 tested GWAIS protocols (see Methods for more details). With GWAIS protocols #5, #6, #7 and #8 based on the MB-MDR framework, we were able to confirm *rs9788973* ( $p$ -value 0.49), which maps to *HLA-B* and *rs30187* ( $p$ -value  $1.1 \times 10^{-9}$ ), which maps to *ERAP1* [17]. These SNPs occurred in the pairs *rs2523608 x rs9788973* and *rs30187 x rs284498* (see Table 3.2).

**Table 3.1:** Most significant SNP pairs across 10 adopted GWAIS analysis protocols. All  $p$ -values are multiple testing corrected, either Bonferroni-based (BOOST protocols) or re-sampling based (MB-MDR protocols).

### 3. GENOME-GENOME INTERACTIONS

SNP A	SNP B	GWAIS protocols										Gene A	Gene B
		BOOST				MB-MDR							
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10		
rs12026423	rs7528311 <sup>+</sup>	0.009	0.004	7.72E-05	3.69E-05	0.401	1	0.001	0.004	1	1	MAGI3	MAGI3
rs11964796	rs13194019 <sup>++</sup>	1	1	0.024	0.012	0.401	1	1	0.995	1	1	PARK2	PARK2
rs13194019	rs1784607 <sup>+++</sup>	1	1	0.144	0.069	0.401	1	1	0.995	1	1	PARK2	PARK2

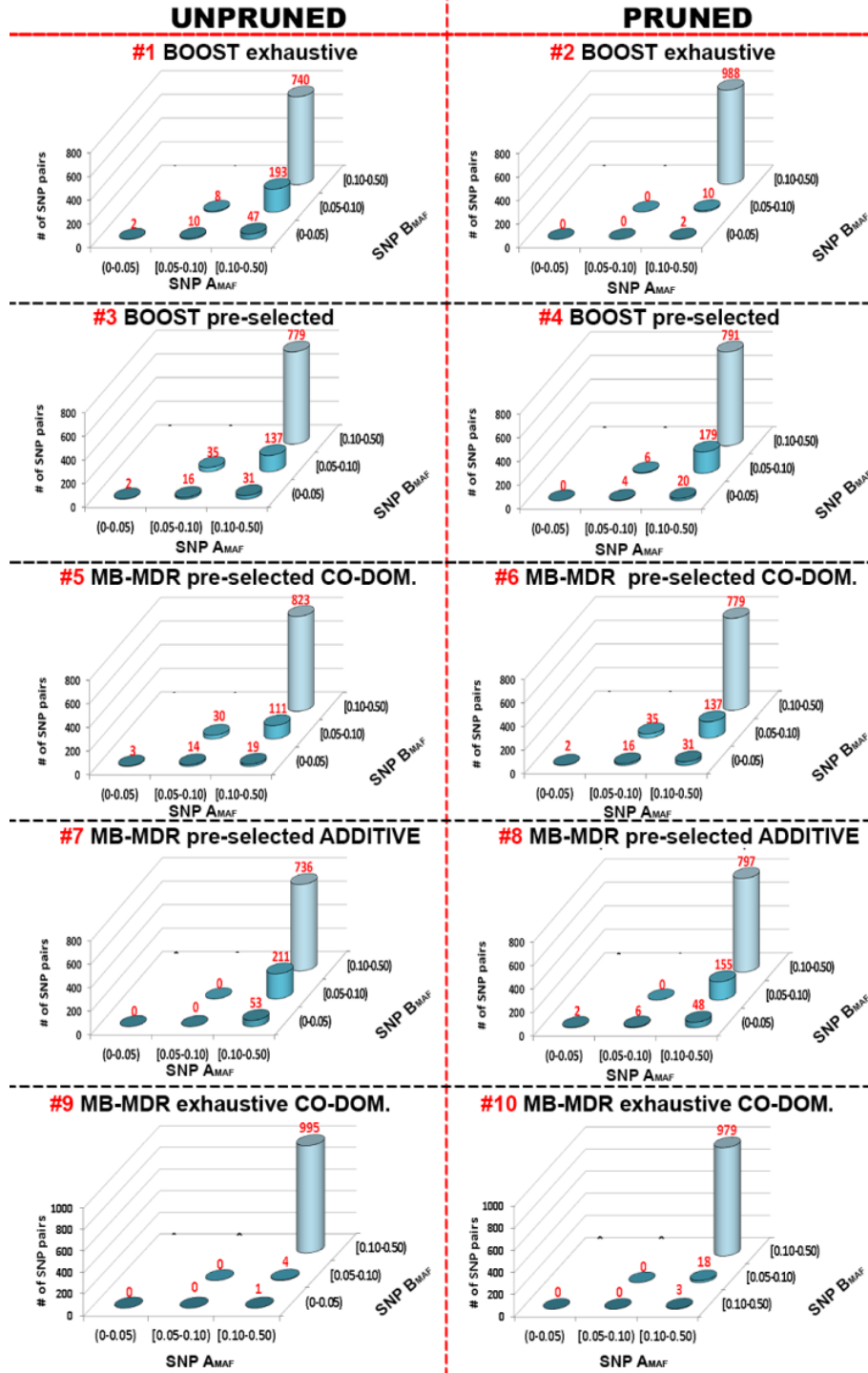
<sup>+</sup> rs12026423/rs7528311 are separated by 13833 bp,  $r^2 = 0.0178$ ; <sup>++</sup> rs11964796/ rs13194019 are separated by 9824 bp and  $r^2 = 0.0309$ ; <sup>+++</sup> rs13194019/rs1784607 are separated by 3127 bp and  $r^2 = 0.0610$

Only GWAIS protocols #7 and #8 coined the aforementioned two pairs as being statistically significant. None of the 102 SNP pairs listed in [17] were found to be statistically significant in our re-analysis, regardless of the protocol used. Relaxing the conditions, we determined the number of SNP pairs with a SNP that occurred in at least one of the 102 SNP pairs reported by [17]. A total of 38 such SNP pairs were detected. These are listed in online Table S3.5. From these, only 8 significant SNP pairs were highlighted by at least one of our GWAIS protocols (in particular, protocol #7 and #8 - Table 3.3)

**Table 3.2:** Significant pairs containing one of the 49 SNPs associated to main effects [17], obtained via the 10 GWAIS protocols.

SNP A	SNP B	GWAIS protocols										Gene A	Gene B
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10		
		multiple testing adjusted <i>p</i> -values											
rs2523608	rs9788973*	1	1	1	1	1	1	0.001	0.001	1	1	<i>HLA-B</i>	<i>MAP2K4</i>
rs30187*	rs2844498	1	1	1	1	1	1	0.001	0.002	1	1	<i>ERAPI</i>	<i>NA</i>

\*SNPs that occurring as main effects SNPs in Supplementary Table 2 of [17] are highlighted in bold.



**Figure 3.4:** Effect of SNP MAFs on ranked epistasis results. For each protocol, the top 1000 epistasis results are presented. Each SNP pair was ordered such that the SNP with the largest MAF was assigned to locus A, and the SNP with the lowest MAF to locus B. The numbers in red refer to the # of SNP pairs that were assigned to each 2-dimensional MAF bin. Protocol numbers match the protocol referencing used in Figure 3.1.

To investigate the influence of MAFs on epistasis findings using different protocols, we defined six 2-dimensional bins (see Methods for more information). The allocation of top 1000 epistasis findings (significant or not) to either of these bins is presented in Figure 3.4. Hence, adding up the number of allocated SNP pairs to each bin (red numbers in Figure 3.4), within the same protocol, gives 1000. Within the exhaustive protocols (#1 and #2, respectively, BOOST applied to unpruned and LD-pruned data; #9 and #10, respectively, MB-MDR applied to unpruned and LD-pruned data), there is a tendency for SNP pairs each having  $MAF \geq 0.05$  to occur in the top 1000. The same is observed for non-exhaustive protocols, in particular those that rely on additive encodings when adjusting for main effects (protocols #7 and #8, MB-MDR applied to unpruned and LD-pruned data, respectively) and those that rely on codominant main effects encoding schemes (protocols #3-#6). The highest number of SNP pairs (out of 1000) with  $MAFs < 0.05$  were obtained with exhaustive BOOST screening on unfiltered and unpruned data (protocol #1), followed by MB-MDR applied unpruned pre-selected data (protocol #5). In general, all protocols give rather similar results, apart from protocols for which all of the top 1000 SNP pairs involved at least one SNP with  $MAF \geq 0.10$ . For protocols #1-#6, the percentage of SNP pairs appearing in the top 1000 list with at least one  $MAF < 0.05$  ranged from 0.2% (protocol #2) to 5.9% (protocol #1).

### 3.4.2. AS pathology relevance

To provide a biological context, we performed a GO functional enrichment analysis on the top 1000 SNP pairs identified within each individual GWAIS protocol. Each SNP was mapped to a gene, when possible (see Methods for additional details). A GO term was considered when at least 10 of these genes were annotated to them. This led to a total of 1326 common GO terms across all 10 GWAIS protocols with combined  $p$ -values  $< 0.05$  (online Table S3.6). Top 10 GO terms are shown in Table 3.2. Using a significance level of 0.05, significant combined  $p$ -values were obtained for GO terms related to the central nervous system (CNS). In particular, links between AS pathology and nervous system signal transmission via synapses biological processes was observed via e.g., GO:0007411 (combined  $p$ -value:  $7.86 \times 10^{-77}$ ), GO:0007268 (combined  $p$ -value:  $2.00 \times 10^{-36}$ ), and GO:0043524 (combined  $p$ -value:  $2.91 \times 10^{-17}$ ). To a lesser degree, we also observed a link between AS and immune system processes that involve antigen processing and presentation via MHC complex: combined  $p$ -value for GO:0002479 of  $1.77 \times 10^{-8}$  (not corrected for multiple

testing). Other overall significant GO terms were linked to biological processes such as membrane transport (GO:0055085, combined  $p$ -value:  $3.04 \times 10^{-50}$ ) and sudden response to stimuli (GO:0001964, combined  $p$ -value:  $1.48 \times 10^{-10}$ ) without a clear association to AS. In addition, we detected an involvement of the *Notch* pathway responsible for the proliferation of neurons (GO:0007219, combined  $p$ -value of  $1.02 \times 10^{-5}$ ), again linking AS to CNS processes.

**Table 3.3:** Statistically significant SNP x SNP interactions that contain a SNP occurring in at least one of 102 SNP pairs listed in Supplementary Table 5 in Evans *et al.* [17]\*.

GWAIS protocol	SNP A	SNP B	Chr A	Chr B	$p$ -value	Gene A	Gene B
#8	<b>rs30187*</b>	rs2844498	5	6	0.002	<i>ERAPI</i>	<i>MICB</i>
	<b>rs30187*</b>	rs2523608	5	6	0.038	<i>ERAPI</i>	<i>HLA-B</i>
#7	<b>rs10050860*</b>	rs2844498	5	6	0.001	<i>ERAPI</i>	<i>MICB</i>
	<b>rs10050860*</b>	rs2523608	5	6	0.001	<i>ERAPI</i>	<i>HLA-B</i>
	<b>rs30187*</b>	rs2844498	5	6	0.001	<i>ERAPI</i>	<i>MICB</i>
	<b>rs30187*</b>	rs2523608	5	6	0.001	<i>ERAPI</i>	<i>HLA-B</i>
	rs2523608	<b>rs10781500*</b>	6	9	0.001	<i>HLA-B</i>	<i>SNAPC4</i>
	rs2844498	<b>rs10781500*</b>	6	9	0.001	<i>MICB</i>	<i>SNAPC4</i>

\* - SNPs that were analyzed in Supplementary Table 5 by [17] are highlighted.

**Table 3.4:** Top 10 Significant GO terms related to top 1000 SNP pairs per GWAIS protocol, based on Fisher's combined *p*-value at a significance level of 0.05. Protocol-specific *p*-values are also reported.

GO ID	GO Term Description	GWAIS protocols										combined*
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
<b>GO:0007411</b>	<b>axon guidance</b>	<b>5.2E-02</b>	<b>1</b>	<b>4E-16</b>	<b>4.4E-18</b>	<b>1.9E-12</b>	<b>2.2E-15</b>	<b>1.2E-13</b>	<b>5.7E-16</b>	<b>5.7E-16</b>	<b>1</b>	<b>7.9E-77</b>
GO:0030168	platelet activation	5.8E-01	1	2.9E-15	2.3E-15	3.2E-11	1.2E-10	4.1E-09	1.2E-11	1.2E-11	1	3.9E-58
<b>GO:0055085</b>	<b>transmembrane transport</b>	<b>4.7E-02</b>	<b>1.6E-01</b>	<b>1.8E-09</b>	<b>1E-09</b>	<b>3.2E-11</b>	<b>5.4E-11</b>	<b>6E-09</b>	<b>1E-12</b>	<b>1E-12</b>	<b>5.5E-01</b>	<b>3.0E-50</b>
<b>GO:0007268</b>	<b>synaptic transmission</b>	<b>2.2E-02</b>	<b>1</b>	<b>8E-10</b>	<b>3.1E-08</b>	<b>1.5E-06</b>	<b>2.4E-09</b>	<b>6.3E-07</b>	<b>5E-08</b>	<b>5E-08</b>	<b>1</b>	<b>2.0E-36</b>
GO:0007173	epidermal growth factor receptor signaling pathway	2.1E-02	1	7.8E-10	1.4E-11	2.4E-07	6.8E-07	2.4E-05	7.2E-06	7.2E-06	1	1.6E-34
GO:0008543	fibroblast growth factor receptor signaling pathway	9.8E-02	1	5.4E-08	6.9E-11	5.1E-07	1.8E-08	2.2E-04	3.6E-04	3.6E-04	1	3.0E-30
GO:0007202	activation of phospholipase C activity	1.0E-02	1	2.6E-08	9.4E-09	1.8E-06	6.8E-06	5.1E-06	3.9E-06	3.9E-06	1	6.4E-30
GO:0006112	energy reserve metabolic process	1.8E-01	1	9.9E-07	3.4E-09	1.2E-04	1.8E-07	5.9E-06	3.6E-05	3.6E-05	1	1.5E-26
GO:0042493	response to drug	1.3E-01	5.6E-01	2.7E-05	1.4E-09	5.1E-03	9.8E-05	1.9E-07	6.6E-08	6.6E-08	6.1E-01	7.9E-26
GO:0006198	cAMP catabolic process	5.2E-03	1	5.1E-04	2.5E-05	2.9E-06	5.6E-08	1.0E-05	1.5E-06	1.5E-06	1	6.0E-25

\* - Combined *p*-values summarize information across the 8 considered protocols. The most relevant GO terms for AS are indicated in bold, as well as, GWAIS-specific *p*-values when < 0.05. The exhaustive list of significant GO terms is shown in online Table S3.6

### 3.5. Discussion

In our study, we demonstrated that choices about data filtering, pruning and lower order effects adjustment may cause substantial variation in epistasis findings. We showed this by making changes to the reference GWAIS protocol we published earlier [8], giving rise to 10 GWAIS protocols under investigation in this work (Figure 3.1). The reference GWAIS protocol consists of a set of guidelines designed to address problems of epistasis reproducibility in the context of genome-wide epistasis screening with thousands of SNP markers. It contains recommendations on rigorous data quality control steps, exhaustive or non-exhaustive marker screening, LD pruning thresholds and the selection of a suitable analytic epistasis detection tool.

Based on our results (for instance Figure 3.3) the major cause of heterogeneity in findings is the choice about which markers to retain in the analysis. We referred to it as “pre-selection of markers”. We used filtering based on biological knowledge to make educated pre-selections, using a compendium of biological databases via Biofilter 2.0 [22]. The effects of pre-selections on the number of SNPs can be huge, as was exemplified on AS: before selection, 487,780 SNPs; after selection, 44,018 SNPs. This has huge consequences for subsequent analyses. In a negative sense, there is a risk of removing pairs of SNPs that may lead to interesting new hypotheses, for which no reported evidence exists in existing biological data repositories. In a positive sense, less multiple tests are need to be performed, hereby reducing computation time and potentially also the number of false positives. Seeking a balance between potentially improving the power of the GWAIS by relying on prior knowledge versus decreasing the chance of missing important findings remains a challenging task. When inspecting the overlap between significant results for each protocol, it is therefore not surprising that little overlap may exist between significant results obtained via exhaustive protocols and significant results obtained via non-exhaustive protocols. In fact, for the AS data we re-analyzed, no overlap was found at the SNP level (see Figure 3.2 and Figure 3.3 protocols #1-#2 versus #3-#8). Furthermore, the protocol adopted by [17] makes a heavy pre-selection of markers. Only those SNPs showing a significant association with AS via main effects GWAs were considered. This involved 15 SNPs, half of which were also included in the 487,780 SNPs that served as input to our own GWAIS protocols (#1-#10): *rs30187*, *rs10781500*, *rs10865331*, *rs11209026*, *rs2297909*, *rs378108*, *rs11209032*. The likelihood ratio interaction tests

adopted in their work were similar to the ones implemented in BOOST. However, whereas in BOOST tests are based on 4 df, interaction tests in [17] were based on 1 df (testing departure from additivity on the log-odds scale). Hence, it is not surprising that none of the significant SNP pairs reported in [17] can be reproduced in our study. Notably, neither BOOST nor MB-MDR in our protocols adjusted for population stratification. In contrast, [17] did correct for potential population stratification using a two-stage approach involving Bayesian clustering and Hidden Markov models. In theory, this may explain additional differences between our analyses and the ones performed in the reference study [17]. However, given that the inflation factor based on median  $X^2$  for the AS data is 1.029, we believe that no adjustments were necessary and hence no spurious results were generated as a result of not correcting for population stratification in our adopted protocols.

Our results, visualized in Figure 3.3, suggest that the second largest cause for heterogeneity in significant findings, derived from different protocols, is the adopted encoding scheme for genetic variants. This is clear for the non-exhaustive protocols included in our study (#5-#8). It is less clear for exhaustive protocols, since the ones included in our study only considered co-dominant encoding schemes (#1-#2). However, our experience with other real-life applications seems to support our suggestion also for exhaustive protocols (data not shown). Previous theoretical work also showed that additive encodings for lower order effects may increase false positives rates in interaction studies [10]. This is in line with the large number of significant interactions identified via protocols #7 and #8 (Figure 3.3). It is very unlikely that over 50,000 significant interactions highlighted by these protocols are genuine, and are caused by the (strong) main effects blurring the epistasis signal [10].

The third largest cause for heterogeneity is attributed to differences in employed LD-pruning approaches. Here, the effect of LD-pruning (i.e. pruning at  $r^2 > 0.75$  or not) was more pronounced under additive encoding schemes (protocols #7 versus #8) as opposed to co-dominant encoding strategies (protocols #3 versus #4, and protocols #5 versus #6). Therefore, it is important to discuss the primary interaction study performed in [17], targeting *additive x additive* interactions, with caution, and in the light of the adopted pruning protocol. Figure 3.3 shows that the effects of LD pruning are more severe for exhaustive protocols compared to non-exhaustive protocols. This is

not surprising, given that the LD pruning in the first implies a reduction of about 150,000 SNPs, compared to less than 15,000 SNPs in the second. Hence, although potentially more significant SNP pairs can be revealed in protocol #1 (exhaustive, BOOST, unpruned), less significant pairs are highlighted as compared to protocol #2 (exhaustive, BOOST, LD-pruned; Figure 3.3). This can be explained by the reduced number of tests to account for Bonferroni corrections. The reverse is observed for protocols #3 (BOOST, pre-selected) and #4 (BOOST, pre-selected and LD-pruned). Here, protocol #4 gives rise to less significant SNP pairs compared to protocol #3 (Figure 3.3). There is still a reduction of the multiple testing burden in protocol #4 is true, but this cannot explain the phenomenon. More likely, an increased number of redundant epistasis signals (due to high LD between some marker pairs) are an explanatory factor. The same can be observed for MB-MDR-based protocols #5 and #6. In particular, again LD pruning as part of protocol #6 gives rise to a smaller number of significant SNP x SNP interactions (47 – see Figure 3.2) compared to protocol #5 (no LD pruning; 77 – see Figure 3.2). Note that MB-MDR and BOOST use quite different multiple testing correction strategies. In case of BOOST, a conservative Bonferroni correction is advocated. In MB-MDR, a permutation-based *maxT* strategy is implemented, which relies on subset pivotality to guarantee strong FWER control at  $\alpha = 0.05$ .

Less common and rare variants tend to increase false positive rates, when inappropriate tests are used, as reported in [35,36]. According to [35] rare SNPs with  $MAF < 0.05$  showed a significantly higher likelihood of being classified as false positives in the logistic regression based GWAS [35]. For BOOST-based protocols (#1 - #4), the percentage of top 1000 SNP pairs with at least one  $MAF < 0.05$  that were statistically significant (multiple testing corrected), was respectively 5.9%, 0.2%, 4.9 % and 2.4% (data not shown). For MB-MDR based protocols (protocols #5-#6) the percentage of such SNP pairs was respectively 0.1% and 0.2%, smaller than with BOOST-based protocols. However, for MB-MDR based protocols #7 and #8 (using additive encoding schemes for main effects adjustment), the percentages were higher (4.8% and 5.3%, respectively). This is in line with earlier findings about MB-MDR performance [10,12,26]. When MB-MDR is applied to rare variants, three factors are at play. First, FWER can be elevated due to violations of the subset pivotality assumption in the built-in *maxT* multiple-testing correction procedure [37]. Second, when marker frequencies are rare, less than 10 individuals may contribute to a multi-locus genotype combination, in which case there is no power to assess whether this combination is related to a

significantly higher or lower disease risk. As a consequence, the power to detect an interaction with such a combination may be hampered. Third, additive coding will always give rise to increased false positives, irrespective of whether rare or common variants are considered.

The fact that protocols #7 and #8 were the only ones that were able to highlight significant interactions, with either one of the 49 main effects SNPs listed in Evans *et al.* 2011, namely *rs2523608 x rs9788973* and *rs310787 x rs2844498* (Table 3.2), is not surprising. MB-MDR with additive encodings has a tendency towards generating more liberal test results than MB-MDR with co-dominant encodings [10,12]. The SNPs *rs9788973* and *rs2523608* map to the genes *MAP2K4* and *HLA-B*. The *HLA-B* gene showed very strong association to AS (*rs4349859* *p*-value  $<10^{-200}$ ) in [17] and was also related to AS in other studies [38,39]. In addition, the *rs2523608 x rs9788973* pair resides in the coding regions of the *HLA-B x MAP2K4* genes (Table 3.2), suggesting that AS pathology is not only linked to irregularities in peptide presentation to immune cells via major histocompatibility complex (MHC), but also to dysfunctions in intra-cellular signaling pathways.

Focusing on the common SNP pairs between GWAIS protocols in our study (207 pairs), only 3 showed a significant interaction in at least one protocol (Table 3.1), pointing towards the genes *MAGI3* and *PARK2*. The gene *MAGI3* controls intracellular signaling cell-cell adhesion and communication [40]. In the context of AS, *MAGI3* potentially regulates cell-cell communication and adhesion of the cells in the inflamed joint areas between spinal discs and vertebra. *PARK2* was suggested before as a candidate gene for AS in [41]. Mutations in the *PARK2* gene can cause alteration in cellular trafficking and protein degradation [42]. In [43], alterations incorrect antigenic peptide presentation by major histocompatibility complex (MHC) class I molecules to CD8<sup>+</sup> T lymphocytes were linked with an early onset of chronic inflammation and AS. Further alteration in protein degradation, partially controlled by *PARK2*, may also suggest an alteration in the proper disposal of antigens. The aberrations in this process may potentially contribute to chronic inflammation of the spine followed by AS onset.

Only 20 pairs were common between our 10 protocols and the list of the 102 SNP x SNP interactions investigated in [17]. Clearly, several interesting pairs are missed by only looking at SNP pairs that are tested by all considered protocols (i.e. common SNP pairs). Imputation, to make

the SNP x SNP pool more alike between protocols, may not only over-rule removal of SNPs after biofiltering (for which one may have had good reasons), it may also induce additional LD between SNPs, which may hugely increase false positives, depending on the analytic tool used. Interestingly, 8 significant SNP x SNP interactions were detected for which at least one SNP was present in the 102 SNP pairs of [17] (Table 3.3). These 8 pairs involved the SNPs *rs30187*, *rs10050860* and *rs10781500*, and allowed to reproduce the statistically interacting gene pair *ERAPI* x *HLA-B* reported in [17] via the interactions *rs3018* x *rs2523608*, *rs10050860* x *rs2523608* and *rs30187* x *rs2523608* (Table 3.3). Notably, these findings were obtained with the only protocols using additive main effects encodings (protocols #7 and #8). Evans and colleagues also primarily based their interaction testing on additive encodings.

However, by allowing more SNPs for interaction testing than in [17], we identified gene pairs not previously associated to AS: *ERAPI* x *MICB*, *MICB* x *SNAPC4* and *HLA-B* x *SNAPC4* (Table 3.3), pointing towards interacting loci or regions between chromosome 5 and 6, and between 6 and 9. *MICB* is MHC Class I Mic-B Antigen linked to cell immune response and is functionally similar to MHC Class I encoded by the *HLA-B* gene. *MICB* is implicated in rheumatoid arthritis [44]. *SNAPC4* encodes small Nuclear RNA Activating Complex important for proper functioning of RNA Polymerase II and III. *ERAPI* encodes for endoplasmic reticulum aminopeptidase that trims peptides.

One of the top 1326 common GO terms across GWAIS protocol #1-#10 was GO:0002479 (online Table S3.6). This term is functionally related to antigen processing and exogenous antigen presentation via MHC class I, TAP-dependent. It may suggest that that AS pathology is partially caused by the inability of *ERAPI* aminopeptidase to correctly trim *HLA* class I-binding peptides and subsequently to present them to MHC complexes [18]. This possibly causes deregulation of the innate immunity and chronic inflammation of spine tissues that are typical symptoms displayed by AS patients [45]. Also appearing in the top 10 are GO terms linked to neural transmission processes (Table 3.4). This agrees with AS known disease pathology characterized by consistent pain and inflammation in the spine – part of the central nervous system (CNS). In particular, the GO terms highlighted in bold in Table 3.4 and online Table S3.6 (column 1), even though based on the top 1000 SNP x SNP interactions (not necessarily statistically significant) may suggest a

link between AS and mutations in genes involved in nerve impulse transmission and propagation (GO:0007411, GO:0007268, etc.). Furthermore, GO:0007219 (online Table S3.6), linked to genes of the *Notch* signaling pathway (e.g., *RBP-J*, *PSEN1*, *ADAM10*), suggests AS interference with the correct development and growth of nerve tissue [46]. It was shown by [47] that the *Notch* pathway also controls angiogenesis and that Vascular Endothelial Growth Factor (*VEGF*) and Angiopoietin (*Ang*) are both over-expressed in synovial tissues of Psoriatic Arthritis and Rheumatoid Arthritis patients.

### 3.6. Conclusions

Any GWAI analysis involves making choices about the input data (e.g., filtering using candidate genes or using prior biological knowledge), about LD-pruning thresholds, about adjusting for lower order effects (and how to encode these), and about the selection of the analytical tool (e.g., non-parametric, semi-parametric or fully parametric), as well as, the corrective method for multiple testing [8]. We have shown that even slight differences in protocols to perform a Genome-Wide Association Interaction (GWAi) study may hamper the results reproducibility. We did so by applying the 10 GWAI protocols to real-life genome-wide SNP data on ankylosing spondylitis (AS) and controls.

Choices about marker selection (for instance filtering based on prior knowledge) are the most severe, as it may give rise to a dramatic reduction in SNPs for further GWAI analysis [6,8,48]. Although biofiltering may reduce the ability to generate novel hypotheses about interactions [48], when doing so the effects of LD pruning and other protocol parameters seem to be less impactful on the final analysis results. More work is needed though to fully understand the interplay between LD-pruning and filtering strategies commonly adopted in GWAIS and to derive operational guidelines. In general, the second largest cause for heterogeneity in GWAI results is the adopted encoding scheme to adjust the interaction analysis for the lower-order effects [21]. The third largest cause is the adopted LD-pruning strategy. To date, no published work exists that comprehensively investigates the effect of LD on epistasis findings derived from several analytic tools. In order not to waste carefully acquired data, researchers are often tempted to adopt exhaustive screening tools whenever computationally feasible. As suggested in [8], we nevertheless advocate LD-pruning at

an  $r^2$  of 0.75, to increase power, yet to reduce the generation of redundant (significant) SNP x SNP interactions. Exhaustively applying BOOST to LD-pruned AS data at an  $r^2$  of 0.75 generated over 2,000 significantly interacting SNP pairs. The existence of moderate LD may induce multicollinearity in regression models and may increase the number of false positives (even when using a conservative multiple testing correction method such as Bonferroni). It shows that when applying a GWAI protocol, the results should be interpreted and discussed under the appropriate context, which includes the limitations and strengths of the adopted protocol, hereby addressing its different components.

Finally, with so many tools for GWAI analysis around, truly comparing these remains a challenging task in the absence of reference synthetic data sets that are rich enough to capture real-life complexities. Care has to be taken when “replicating” interactions with analytic tools that have a tendency to generate false positives: Can one be sure that one is not replicating a false positive? Clearly, no single tool will fit all. Tools are heterogeneous in their ability to recognize specific active epistasis modes and several such modes are likely to occur throughout the genome. This observation puts limitations to strategies that use agreement between different GWAS approaches as evidence for an interaction. It also favors the development of a hybrid SNP x SNP interaction detection tool, combining the best of several worlds when screening the genome.

### **3.7. Acknowledgments**

The research was funded by the Fonds de la Recherche Scientifique (FNRS) (incl. FNRS F.R.F.C. project convention n° 2.4609.11). We thank François Van Lishout and Elena Gusareva from the Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium for their support and advice. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from [www.wtccc.org.uk](http://www.wtccc.org.uk). Funding for the project was provided by the Wellcome Trust under award n° 076113. The authors declare that they have no competing interests.

### 3.8. Chapter highlights

In this chapter, we compared different methodologies in GWAIS. We assessed how genetic information is implicated in complex diseases such as AS. We successfully validated SNP x SNP  $\rightarrow$  phenotype epistasis methodologies on the genome-wide and smaller scales via MB-MDR and EpiShell tools. Importantly, we brought more awareness of careful parameter selection during QC steps of any GWAS and GWAI protocol. In the next chapter, we will incorporate gene expression as an additional data source to the interactions mining task (i.e. *trans/cis* epistatic eQTLs identification).

### 3.9. Appendix

**Table S3.1:** Parameters used to run ten GWAI protocols

<b>GWAI protocol</b>	<b>#1*</b>	<b>#2*</b>	<b>#3*</b>	<b>#4*</b>
LD pruning	NO	YES	NO	YES
LD pruning threshold	NA	R2>0.75	NA	R2>0.75
Analytical tool (epistasis detection)	BOOST	BOOST	BOOST	BOOST
Running parameters	permutations=0, min test stat=54	permutations=0, min test stat=54	permutations=0, min test stat=54	permutations=0, min test stat=54
Input dataset size (number of SNPs)	487,780	321,565	44,018	30,426
<b>GWAI protocol</b>	<b>#5*</b>	<b>#6*</b>	<b>#7*</b>	<b>#8*</b>
LD pruning	NO	YES	NO	YES
LD pruning threshold	NA	R2>0.75	NA	R2>0.75
Analytical tool (epistasis detection)	MB-MDR	MB-MDR	MB-MDR	MB-MDR
Running parameters	<b>-mt MAXT,</b> permutations=999, -a CODOMINANT, -m 10, -x 0.1	<b>-mt MAXT,</b> permutations=999, -a CODOMINANT, -m 10, -x 0.1	<b>-mt MAXT,</b> permutations=999, -a ADDITIVE, -m 10, -x 0.1	<b>-mt MAXT,</b> permutations=999, -a ADDITIVE, -m 10, -x 0.1
Main effect correction	co-dominant	co-dominant	additive	additive
Input dataset size (number of SNPs)	44,018	30,426	44,018	30,426
<b>GWAI protocol</b>	<b>#9*</b>	<b>#10*</b>		
LD pruning	NO	YES		
LD pruning threshold	NA	R2>0.75		
Analytical tool (epistasis detection)	MB-MDR	MB-MDR		
Running parameters	<b>-mt gammaMAXT,</b> permutations=999, -a CODOMINANT, -m 10, -x 0.1	<b>-mt gammaMAXT,</b> permutations=999, -a CODOMINANT, -m 10, -x 0.1		
Main effect correction	co-dominant	co-dominant		
Input dataset size (number of SNPs)	487,780	321,565		

(\*) Legend:

protocol #1 - BOOST exhaustive - BOOST entire data (487,780 SNPs);

protocol #2- BOOST exhaustive LD pruned (321,565 SNPs);

protocol #3 - BOOST pre-selected (44,018 SNPs);

### 3. GENOME-GENOME INTERACTIONS

protocol #4 - BOOST pre-selected LD pruned (30,426 SNPs);  
 protocol #5 - MB-MDR pre-selected CODOMINANT (44,018 SNPs);  
 protocol #6 - MB-MDR pre-selected CODOMINANT LD pruned (30,426 SNPs);  
 protocol #7 - MB-MDR pre-selected ADDITIVE (44,018 SNPs);  
 protocol #8 - MB-MDR pre-selected ADDITIVE LD pruned (30,426 SNPs).  
 protocol #9 - MB-MDR exhaustive CODOMINANT - *gammaMAXT* (487,780 SNPs);  
 protocol #10 - MB-MDR exhaustive CODOMINANT LD pruned - *gammaMAXT* (321,565 SNPs);

**Table S3.2:** Euclidean distances amongst GWAI protocols (ref. to Figure 3.1)

			Gwai protocol									
			BOOST				MB-MDR					
			#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Gwai protocol	BOOST	#1	0									
		#2	9,123,905	0								
		#3	17,155,486	8,031,864	0							
		#4	17,241,213	8,117,600	85,745	0						
	MB-MDR	#5	17,290,386	8,166,857	136,725	53,636	0					
		#6	17,304,827	8,181,254	149,798	64,508	18,013	0				
		#7	15,884,721	6,882,567	2,434,639	2,487,225	2,508,128	2,522,599	0			
		#8	16,343,189	7,268,044	1,378,617	1,432,511	1,454,519	1,469,530	1,065,717	0		
		#9	14,253,250	5,506,100	3,966,195	4,033,412	4,059,666	4,077,675	2,591,591	2,975,847	0	.
		#10	15,991,829	6,914,928	1,572,583	1,638,264	1,664,079	1,682,085	1,568,040	950,823	2,395,651	0

**Note:** highest distance is highlighted in **red**; the lowest distance in **green**;

(\*) Legend:

protocol #1 - BOOST exhaustive - BOOST entire data (487,780 SNPs);  
 protocol #2- BOOST exhaustive LD pruned (321,565 SNPs);  
 protocol #3 - BOOST pre-selected (44,018 SNPs);  
 protocol #4 - BOOST pre-selected LD pruned (30,426 SNPs);  
 protocol #5 - MB-MDR pre-selected CODOMINANT (44,018 SNPs);  
 protocol #6 - MB-MDR pre-selected CODOMINANT LD pruned (30,426 SNPs);  
 protocol #7 - MB-MDR pre-selected ADDITIVE (44,018 SNPs);  
 protocol #8 - MB-MDR pre-selected ADDITIVE LD pruned (30,426 SNPs).  
 protocol #9 - MB-MDR exhaustive CODOMINANT - *gammaMAXT* (487,780 SNPs);  
 protocol #10 - MB-MDR exhaustive CODOMINANT LD pruned - *gammaMAXT* (321,565 SNPs);

**Table S3.3:** Significant SNP pairs with multiple testing adjusted *p*-values (<0.05). (See the online supplement)

**Table S3.4:** List of common 207 SNP pairs amongst 10 GWAI protocols findings (including significant and non-significant SNP pairs). (See the online supplement).

**Table S3.5:** Annotated Evans' *et al.* (2011) 38 SNP pairs out of 102 listed in Supplementary Table 5 of [17]. These pairs contain one SNP (in bold) that was present amongst the significant findings of the 10 GWAI protocols. (See the online supplement)

**Table S3.6:** Significant GO terms related to top 1000 SNP pairs per GWAI protocol, based on Fisher's combined  $p$ -value at a significance level of 0.05. Protocol-specific  $p$ -values are also reported. (See the online supplement)

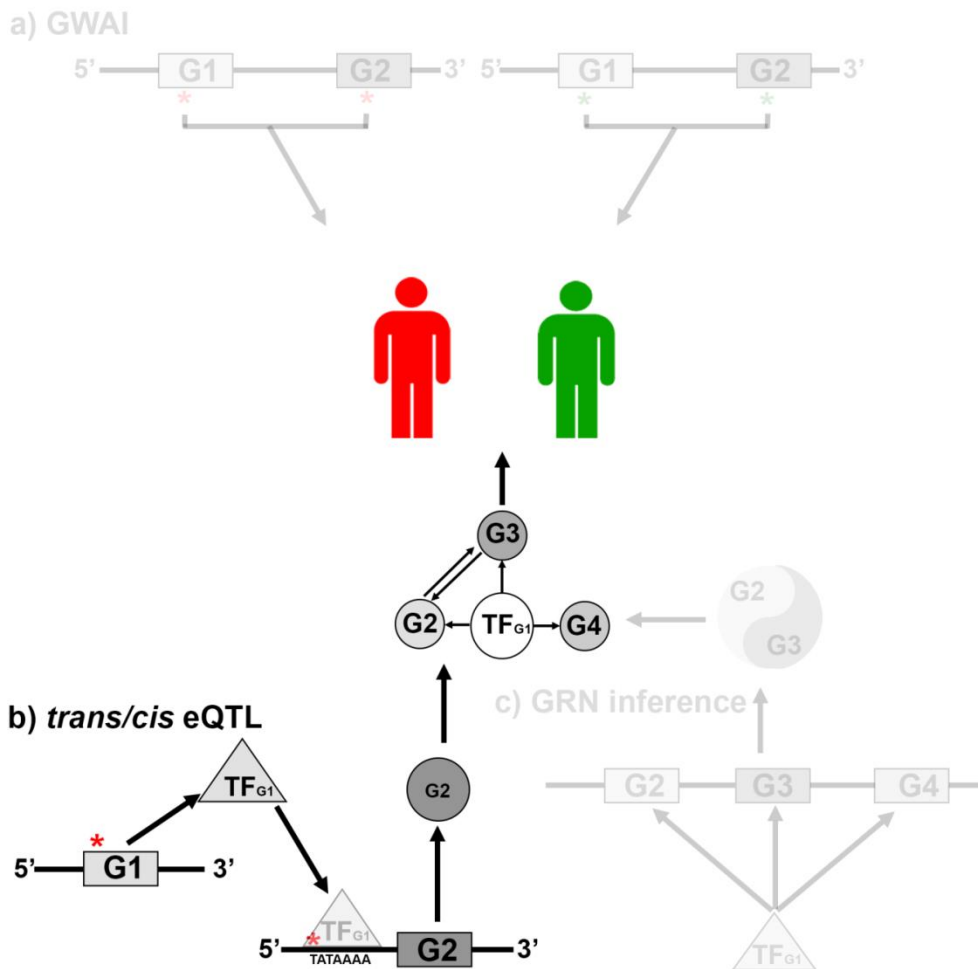
### 3.10. References

1. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.** *The American Journal of Human Genetics* 87: 325-340.
2. Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, et al. (2011) **Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise.** *Ann Hum Genet* 75: 78-89.
3. Gyenesei A, Moody J, Semple CA, Haley CS, Wei WH (2012) **High-throughput analysis of epistasis in genome-wide association studies with BiForce.** *Bioinformatics* 28: 1957-1964.
4. Zhang Y, Jiang B, Zhu J, Liu JS (2011) **Bayesian models for detecting epistatic interactions from genetic data.** *Ann Hum Genet* 75: 183-193.
5. Pang X, Wang Z, Yap JS, Wang J, Zhu J, et al. (2013) **A statistical procedure to map high-order epistasis for complex traits.** *Brief Bioinform* 14: 302-314.
6. Van Steen K (2012) **Travelling the world of gene-gene interactions.** *Brief Bioinform* 13: 1-19.
7. Wei WH, Hemani G, Haley CS (2014) **Detecting epistasis in human complex traits.** *Nat Rev Genet.*
8. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet.*
9. Wei W-H, Hemani G, Haley CS (2014) **Detecting epistasis in human complex traits.** *Nature Reviews Genetics.*
10. Mahachie John JM, Cattaert T, Lishout FV, Gusareva ES, Steen KV (2012) **Lower-order effects adjustment in quantitative traits model-based multifactor dimensionality reduction.** *PLoS One* 7: e29594.
11. Grange L (2014) **Thesis: epistasis in genetic susceptibility to infectious diseases: comparison and development of methods application to severe dengue in Asia:** Paris 7.
12. Mahachie J (2012) **Thesis: Genomic Association Screening Methodology for High-Dimensional and Complex Data Structures:** University of Liege.
13. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.** *Am J Hum Genet* 87: 325-340.
14. Westfall PH, Young SS (1993) **Resampling-based multiple testing: Examples and methods for  $p$ -value adjustment:** John Wiley & Sons.
15. Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, et al. (2013) **An efficient algorithm to perform multiple testing in epistasis screening.** *BMC Bioinformatics* 14: 138.
16. Dean LE, Jones GT, MacDonald AG, Downham C, Sturrock RD, et al. (2014) **Global prevalence of ankylosing spondylitis.** *Rheumatology* 53: 650-657.
17. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, et al. (2011) **Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility.** *Nat Genet* 43: 761-767.
18. Alvarez-Navarro C, Lopez de Castro JA (2013) **ERAP1 structure, function and pathogenetic role in ankylosing spondylitis and other MHC-associated diseases.** *Mol Immunol.*
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 81: 559-575.
20. Colin Freeman JM (2012) **GTOOL.** Oxford University.
21. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet* 133: 1343-1358.

22. Bush WS, Dudek SM, Ritchie MD (2009) **Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies.** *Pac Symp Biocomput*: 368-379.
23. Bozeman M (2015) **Golden Helix, Inc. SNP & Variation Suite (Version 7. x)[Software].**
24. Laura Grange KB, Tom Cattaert, Iryna Nikolayeva, Jestinah M Mahachie John, Benno Schwikowski, Jean-François Bureau, Anavaj Sakuntabhai, Kristel Van Steen (2016) **Finding the tree for the forest: which epistasis analysis method to choose.** Department of Genomes and Genetics, Institut Pasteur, Functional Genetics of Infectious Diseases, Systems and Modeling Unit – BIO3, Quartier Polytech 1, University of Liège, Liège, Belgium, Systems Biology and Chemical Biology, GIGA-R, University of Liège, Liège, Belgium. pp. 21.
25. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, et al. (2001) **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 69: 138-147.
26. Mahachie John JM, Van Lishout F, Van Steen K (2011) **Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data.** *Eur J Hum Genet* 19: 696-703.
27. Lishout FV, Gadaleta F, Moore JH, Wehenkel L, Steen KV (2015) **gammaMAXT: a fast multiple-testing correction algorithm.** *BioData Min* 8: 36.
28. Greene CS, Penrod NM, Williams SM, Moore JH (2009) **Failure to replicate a genetic association may provide important clues about genetic architecture.** *PLoS One* 4: e5639.
29. Kestler HA, Muller A, Gress TM, Buchholz M (2005) **Generalized Venn diagrams: a new method of visualizing complex genetic set relations.** *Bioinformatics* 21: 1592-1595.
30. RCoreTeam (2013) **R: A Language and Environment for Statistical Computing**. Vienna, Austria.
31. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, et al. (2010) **SCAN: SNP and copy number annotation.** *Bioinformatics* 26: 259-262.
32. Huang da W, Sherman BT, Lempicki RA (2009) **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 37: 1-13.
33. Ackermann M, Strimmer K (2009) **A general modular framework for gene set enrichment analysis.** *BMC Bioinformatics* 10: 47.
34. Alexa A, Rahnenfuhrer J, Lengauer T (2006) **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 22: 1600-1607.
35. Tabangin ME, Woo JG, Martin LJ (2009) **The effect of minor allele frequency on the likelihood of obtaining false positives.** *BMC Proc* 3 Suppl 7: S41.
36. Mahachie John JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, et al. (2011) **Comparison of genetic association strategies in the presence of rare alleles.** *BMC Proc* 5 Suppl 9: S32.
37. Mahachie John JM, Van Lishout F, Gusareva ES, Van Steen K (2013) **A robustness study of parametric and non-parametric tests in model-based multifactor dimensionality reduction for epistasis detection.** *BioData Min* 6: 9.
38. Jenisch S, Henseler T, Nair RP, Guo SW, Westphal E, et al. (1998) **Linkage analysis of human leukocyte antigen (HLA) markers in familial psoriasis: strong disequilibrium effects provide evidence for a major determinant in the HLA-B/-C region.** *Am J Hum Genet* 63: 191-199.
39. Nischwitz S, Cepok S, Kroner A, Wolf C, Knop M, et al. (2010) **Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis.** *J Neuroimmunol* 227: 162-166.
40. Adamsky K, Arnold K, Sabanay H, Peles E (2003) **Junctional protein MAGI-3 interacts with receptor tyrosine phosphatase beta (RPTP beta) and tyrosine-phosphorylated proteins.** *J Cell Sci* 116: 1279-1289.
41. Claushuis D, Cortes A, Bradbury LA, Martin TM, Rosenbaum JT, et al. **A genomewide association study of anterior uveiti**; 2012; Washington, DC, United States. John Wiley & Sons. pp. S259-S259.

42. Verdecia MA, Joazeiro CA, Wells NJ, Ferrer JL, Bowman ME, et al. (2003) **Conformational flexibility underlies ubiquitin ligation mediated by the WWP1 HECT domain E3 ligase.** *Mol Cell* 11: 249-259.
43. Boisgerault F, Mounier J, Tieng V, Stolzenberg MC, Khalil-Daher I, et al. (1998) **Alteration of HLA-B27 peptide presentation after infection of transfected murine L cells by Shigella flexneri.** *Infect Immun* 66: 4484-4490.
44. Lopez-Arbesu R, Ballina-Garcia FJ, Alperi-Lopez M, Lopez-Soto A, Rodriguez-Rodero S, et al. (2007) **MHC class I chain-related gene B (MICB) is associated with rheumatoid arthritis susceptibility.** *Rheumatology (Oxford)* 46: 426-430.
45. Chaudhary SB, Hullinger H, Vives MJ (2011) **Management of acute spinal fractures in ankylosing spondylitis.** *ISRN Rheumatol* 2011: 150484.
46. Housden BE, Fu AQ, Krejci A, Bernard F, Fischer B, et al. (2013) **Transcriptional dynamics elicited by a short pulse of notch activation involves feed-forward regulation by E (spl)/Hes genes.** *PLoS genetics* 9: e1003162.
47. Gao W, Sweeney C, Walsh C, Rooney P, McCormick J, et al. (2013) **Notch signalling pathways mediate synovial angiogenesis in response to vascular endothelial growth factor and angiopoietin 2.** *Ann Rheum Dis* 72: 1080-1088.
48. Sun X, Lu Q, Mukheerjee S, Crane PK, Elston R, et al. (2014) **Analysis pipeline for the epistasis search—statistical versus biological filtering.** *Frontiers in genetics* 5.

# Integrative network-based analysis of *cis* and *trans* regulatory effects in asthma



Related publication:

Bessonov K, Croteau-Chonka D, Qi W, Carey VJ, Raby BA, Van Steen K (2015) **Integrative network-based analysis of *cis* and *trans* regulatory effects in asthma.** (*circulating among co-authors*)



## 4. Integrative network-based analysis of *cis* and *trans* regulatory effects in asthma

### 4.1. Chapter summary

The previous chapter main topic was the two-way (SNP x SNP) interactions in relation to a complex trait. We highlighted the importance of rigorous data pre-processing prior to or as part of a GWA analysis. In this chapter, we extend the application of the MB-MDR methodology to gene expressions as traits. In contrast to Chapter 3, as many GWAIS are performed as there are gene probes in the available transcriptome data (one per each gene probe).

Presence or absence of a genetic marker at a given locus may lead to different mRNA expression levels which constitutes an example of DNA-RNA biological interaction. In addition, the genetic marker can be either located nearby of the target gene (i.e. *cis*) transcription start site (TSS) or further away (i.e. *trans*). These *trans* and *cis* DNA-RNA interactions commonly referred to as *trans* and *cis* expression quantitative trait loci (eQTLs) can be detected via classical statistical approaches (e.g., linear regression). As highlighted in Section 1.3, there is a difference between statistical and biological interactions which not always correspond to one another and are not necessarily directly translatable. Since false positives is a problem especially in datasets with  $p \gg n$  datasets, where  $p$  is the number of variables and  $n$  are observations. The inferred DNA-RNA interactions ideally should be experimentally validated and rigorous multiple-testing correction applied. In addition, the most of the *trans* eQTL and *cis* eQTL methods do not capture the complexity of DNA-RNA transcriptional processes as they analyze each eQTL locus in isolation without consideration of entire genomic context. Taking into consideration interactions between loci with respect to expression trait by the statistical eQTL model is a step forward towards holistic biologically relevant models. In this chapter, we introduce *trans/cis* eQTL model which takes into account interaction effects between *trans* and *cis* loci with respect to an expression trait.

Identification of interactions within and between omics data layers is an active area of research [1-3]. Glass *et al.* developed an innovative message parsing method PANDA [3] accounting for a ‘cross-talk’ between multiple omics data layers resulting in an aggregated consensus gene regulatory network (gene - gene interaction network). Also inspired by the work on integrative

omics by several other authors, such as [2,3], we developed an MB-MDR based *trans/cis* epistatic eQTL detection methodology, designed to better account for inter- and intra- cross-talk between DNA and RNA omics information layers.

**Problem:** We are interested in the role of DNA-RNA interactions and of DNA effect modifiers of genetic markers on gene expression levels. Specifically, using MB-MDR and a given pool of *cis* eQTLs we develop and validate a *trans* eQTL epistasis protocol to detect effect modifiers to *cis* SNP  $\rightarrow$  gene-expression relationships. These effect modifiers are assumed to be outside the boundaries of the target gene (i.e. in *trans*). We use the identified epistasis signals (for *trans* SNP x *cis* SNP  $\rightarrow$  gene expression level) to build a gene regulatory network. The network can be seen as a gene-based statistical epistasis network. The complex disease context is asthma. Our epistatic *trans/cis* eQTL method needs to deal with false positives in the presence of a huge number of tests performed between *trans* and *cis* loci pairs and the expression trait. Adequate control of false positives is still an open question in epistatic eQTL analyses. Our goal is to explore different multiple testing corrections within the context of the chosen real-life dataset.

**Results:** Valuable information can be retrieved from the inclusion of *trans* modifier effects to *cis* eQTL  $\rightarrow$  gene expression associations, as was shown on available data for asthmatic children [4]. Multiple testing is an issue and exclusion of SNPs with  $MAF < 0.20$  from the analysis results in a permutation-based False Discovery Rate (FDR) and Familywise Error Rate (FWER) of less than 0.05. In our real-life application to asthma we observe a pathway overlap between *trans* and *cis* gene sets mapped from *trans/cis* and *cis* eQTLs of 18.7%. The *trans* gene set is defined by mapping the *trans* locus of the *trans/cis* eQTL loci pair to the nearest gene. Moreover, the *trans/cis* eQTL network is rather sparse with maximal cliques only reaching the size of 2 genes. In the context of asthma an important transcription regulation pathway (REACTOME: R-HSA-212436) is highlighted by both the *trans* and *cis* gene sets included in the above-mentioned 18.7% overlap. In addition, the epithelial cell adhesion pathway (REACTOME: R-HSA-418990) is enriched in the *trans* gene set and is known to be strongly linked to asthma disease etiology [5]. A differential network analysis between smokers and non-smokers suggests strong environmental impact of smoking on topological and biological properties of the derived *trans* x *cis* eQTL gene networks.

Among the most impacted biological functions are those related to the immune system and DNA repair.

**Keywords:** *cis*-regulation, *trans*-regulation, eQTLs, gene regulatory networks, MB-MDR, asthma, CAMP, smoking

## 4.2. Introduction

Asthma is a complex disease characterized by the interplay of genetic and environmental components. It is diagnosed in 43 out of 1000 individuals worldwide according to the World Health Organization [6]. Its prevalence is increasing, especially in children. In the USA, the disease prevalence almost doubled over a 20 years period (1984-2004), reaching 8 to 10% [7]. The exact causes and triggers of this complex disease are to a great extent unknown. It is believed that causes are linked to the Western culture aspects including increased exposure to allergens, immunization shots, and cleaner living conditions [7]. The pathological aspect of asthma is characterized by chronic inflammation of airways with episodes of airway obstruction. The airway tissue is infiltrated with CD4<sup>+</sup> T-helper, eosinophils and other cells controlling the inflammation processes resulting in airway remodeling. Specifically, the airway wall thickens from 10 to 300% leading to a reduction of air flow causing “breathlessness”. There are different types of asthma but allergic asthma is the most common with ~80% prevalence rates [7]. A more detailed description of asthma subtypes can be found in Section 6.2.2.

A given locus associated with expression of a gene is commonly referred to as expression Quantitative Trait Locus (eQTL). Expression QTL studies are often performed as a functional follow-up to GWAS and GWAIS. However, these treat each genetic locus independently from others in assessing genome – transcriptome associations. A vast amount of publications exist on eQTL analyses [8-11]. Classical one-way legacy *cis* eQTL methodologies include sparse partial least squares (SPLS) [12], Haley-Knott regression (HK) [13], and composite interval mapping (CIM) [12,14]. Briefly, eQTL studies involve the identification of genetic variants (i.e. causal loci) that either affect expression of nearby (*cis* eQTLs) or distant genes (*trans* eQTLs), respectively. In general, *cis* eQTLs are found to be more common (16.9%) compared to *trans* eQTLs (0.2%) in complex disease data [15]. While *cis* eQTLs provide a better understanding of direct genotype effects, *trans* eQTLs are valuable in the identification of indirect effects caused by downstream affected genes (Figure 2.2), thus potentially providing a deeper understanding of disease pathology. The power of eQTL studies lies in their ability to combine genetic and expression data [16]. Yet, only a limited number of studies exist that investigate the influence of genetic interactions on gene expression [8,9,17,18]. In this Chapter 4 we consider *trans/cis* epistatic eQTLs that measure

association of *trans* and *cis* loci pairs with respect to expression trait. These eQTLs are schematically depicted by Figure 2.2C.

In this chapter, eQTL loci in *cis*, may be impacted by other distant loci (not mapped to the targeted *cis* gene), here called loci in *trans*. (Figure 2.2 and Figure 4.1). This novel integrated view, considering epistatic interactions between *trans* and *cis* loci affecting gene expression, may shed new light on gene - gene interaction networks and may increase our understanding about biological mechanisms underlying complex traits under investigation. As mentioned before, most of the current studies separately study *trans* and *cis*-regulation (i.e. *trans* and *cis* eQTL mappings) without consideration of their potential synergistic interaction effects. In contrast, our approach is specifically designed to detect the *trans/cis* gene regulation synergies: effects of *trans* genetic components modifying *cis*-eQTL associations. We identify statistically significant *trans* x *cis* epistatic eQTL interactions using MB-MDR [19] introduced in Section 2.3.2. Significant interactions are subsequently translated to gene - gene networks and analysed via standard network analysis tools. Application of our methodology to genome and transcriptome data from the childhood asthma management program (CAMP) [4], shows that there is up to 18.7% overlap in biological functions between *trans* and *cis* transcription regulatory components corresponding to the *trans* x *cis* eQTL interactions. The main overlapping functions include immune system and signaling pathways, amongst others. Our integrative *trans/cis* eQTL methodology reaches acceptable FWER levels: less than or equal to 0.05 extensively studied in Section 4.4.2. The proposed methodology is a step forward towards integrative *trans/cis* eQTL analysis, harnessing the power of both statistical and network-based approaches. In the sequel, we explain the proposed methodology in a greater detail.

From statistical point of view identification of associations between SNPs (predictors) and gene expression (trait) data led to the development of multiple methods based on logic regression [20], simulated annealing [21], tree-based to search for possible genome-transcriptome interactions [22]. Validation of statistical significance of the identified eQTLs is a delicate issue. Most studies employ cross-validation error or model size reduction via variable pre-selection. Given the complex nature of the epistatic *trans/cis* eQTL analysis and a large number of interaction hypotheses, a careful selection of multiple testing correction strategy is required. Compared to

classical single locus eQTL analysis testing a single locus at a time, the multiple testing issue is particularly severe in epistatic eQTL analysis due to significantly larger number of hypotheses. This prompted development of novel multiple testing correction solutions for epistatic eQTL analysis that can broadly be divided into ‘both significant’ or ‘either significant’ requiring either one or both loci to meet a significance threshold [9]. Storey *et al.* applied a step-wise ‘either significant’ epistatic eQTL method [23] where for a given *trans/cis* eQTL pair they first selected a single *cis* locus, followed by the selection of the secondary *trans* locus provided the largest improvement in statistical power over the first single locus model. As multiple test control they applied FDR cut-off at 0.05 and showed that epistatic loci affected 14% of the *S.cerevisiae* genes [23]. Another example of ‘either significant’ method is given by Fish *et al.* [24] which first identified *cis* eQTLs with marginal  $p$ -value  $< 0.05$  via classical regression model with top 3 principal components followed by likelihood ratio test (LRT) comparing full model to a reduced model lacking the interaction term. FDR at 0.05 ( $p$ -value  $\leq 1.328 \times 10^{-5}$ ) was selected as multiple testing correction. The ‘both significant’ solution selects *trans/cis* eQTL loci such that at both *trans* and *cis* loci have strong marginal association effects to the expression trait. Carlborg *et al.* [25] applied genome-wide regression-based scan considering all potential *cis* and *trans* loci with significant marginal effects using randomization test based on genetic algorithm followed by epistatic eQTL scan amongst the selected loci. Their epistatic model was regression-based and includes both the marginal genetic effects and the four pairwise interaction terms to account for additive and dominant effects between loci pairs. The type I errors were controlled via empirically derived population-based, genome-wide significance thresholds from randomization testing based on permutations of the predictor variables (i.e. loci) with regard to expression trait as discussed in [26]. Other solutions in order to alleviate stringent multiple testing burden reduce the number of markers (i.e. predictors) and, therefore, decrease the number of hypotheses to test. Similar to Chapter 3, complementary information such as knowledge of disease etiology, known list of marker genes, physical protein interaction data can improve statistical power, lower stringency of thresholds, and alleviate computational requirements [27]. Suthram *et al.* [28] applied a prior information from protein-protein interaction networks to fine-map eQTLs. Finally, Boulesteix *et al.* proposes a novel approach for calculation of multiplicity adjusted significance of the SNP x SNP eQTL pair via estimation of the maximally selected chi-square statistic assuming its

multivariate normal distribution under the null hypothesis of no association between the SNP pair [29].

In our epistatic *trans/cis* eQTL MB-MDR-based methodology, the multiple testing corrections are done at both *cis* eQTL and *trans/cis* eQTL loci selection stages similar to ‘both significant’ approach of [9]. At *cis* eQTL stage the multiple testing correction is done via classical Benjamini, and Hochberg FDR correction method [30] at the 0.05 threshold. At *trans/cis* eQTL search stage with fixed *cis* eQTL locus the multiple testing corrections are done for each expression trait. These corrections are based on the built-in step-down *MAXT* correction implemented in MB-MDR [31].

## 4.3. Methods

### 4.3.1. Data

The childhood asthma management program (CAMP) is composed of 1348 subjects (728 males and 620 females) [4]. We chose a subset with 177 subjects of Caucasian origin with age ranging between 16 and 25 years. For each selected subject peripheral blood CD4+ lymphocytes were used to extract 19,451 gene expression values and 528,890 SNPs via the Illumina HumanRef-8 v2 Expression BeadChip platform and Human550-Quad and Human610-Quad Illumina platform [32]. Excellent concordance rates of minimum 99.89% between the two genotype platforms were observed based on the 4 subjects genotyped on both platforms [32].

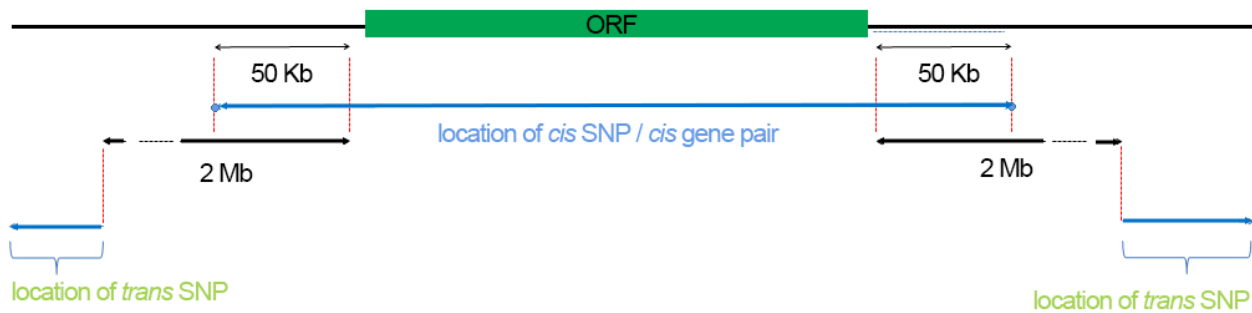
To assess the impact of smoking (i.e. environmental variable), the 177 subjects were further subdivided into 87 smokers (*S*) and 90 non-smokers (*NS*) groups. The “TobaccoSmoke” and “EnvironmentSmoke” were used as environmental selection variables. The *NS* group was formed by non-smoker subjects (TobaccoSmoke=0) that were also not exposed to environmental effects of second-hand smoke (EnvironmentSmoke=0). Specifically, the 177 subjects part of either *NS* or *S* groups were either smokers not exposed to secondhand smoke (TobaccoSmoke=1, EnvironmentSmoke=0) or smokers also exposed to secondhand smoke (TobaccoSmoke=1 and EnvironmentSmoke=1), or non-smokers exposed to secondhand smoke (TobaccoSmoke=0, EnvironmentSmoke=1).

In our analyses, non-imputed genotype data were used. After data filtering and general quality control (QC) steps in PLINK 1.9 [33] and *GenABEL* library in R [34], the final genomic data consisted of 189,969 SNPs for 177 subjects. The QC steps involved filtering out SNPs with  $HWE < 1e-05$ ,  $MAF < 0.20$ , call rate  $< 0.98$  and  $r^2 > 0.75$  (LD pruning). Gene expression data were normalized using quantile-quantile normalization [35] as implemented in the *lumi* R library [36]. Gene expression values for 16,020 genes were available for all 177 subjects, but only 1763 selected (one expression probe per gene) corresponding to the 1763 *cis* eQTLs.

#### 4.3.2. eQTLs epistasis mapping

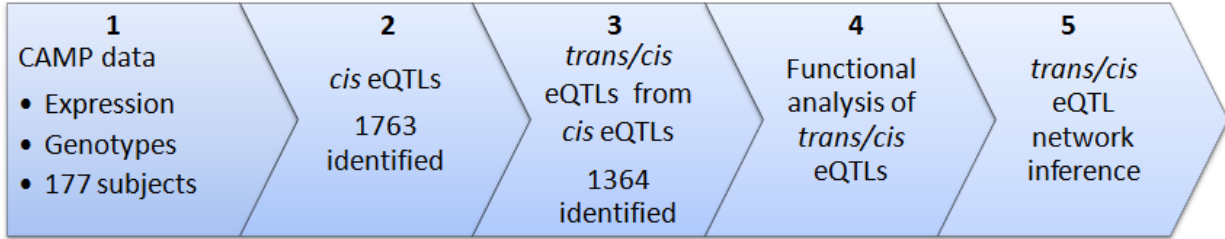
The proposed *trans*  $\times$  *cis* eQTL hybrid analysis pipeline couples a traditional *a priori* eQTL search with a posteriori *trans*  $\times$  *cis* eQTL epistasis analysis, allowing the identification of novel *trans/cis* eQTL regulators. A *trans*  $\times$  *cis* eQTL interaction is a statistical interaction between a *trans* SNP and a *cis* eQTL where *trans* SNP modifies the effect of the *cis* eQTL SNP. Next, the identified *trans/cis* eQTL interactions are visualized as a gene - gene network and analyzed as such.

The *cis* gene mapping was defined by *cis* SNPs located 50Kb upstream and downstream from the *cis* eQTL open reading frame (ORF). The *trans* genes were defined by *trans* SNPs located at least 2 Mb away of the *cis* eQTL ORF. These *trans* SNPs, although interacting with the *cis* eQTL SNP, were physically mapped to the nearest ORF.



**Figure 4.1:** Definition of *cis* and *trans* SNPs with respect to the open reading frame (ORF) of an eQTL gene. ORF is the region of DNA that codes for protein and includes both intron and exon sequences. A *trans*  $\times$  *cis* eQTL pair refers to a SNP pair involved in a *trans*  $\times$  *cis* eQTL interaction and is defined by *trans* and *cis* SNPs defined schematically by this diagram.

Starting point of the analysis work flow (Figure 4.2) is a *cis* eQTL analysis. Such an analysis, using linear regression and least squares parameter estimation, revealed 1763 significant *cis* hits corresponding to 1585 unique genes [32]. The linear model had the form  $Y = \beta_0 + \beta_1 X$ , where  $X$  represents a locus (i.e. SNP, additive encoding) and  $Y$  gene expression intensity for its corresponding gene. The regression coefficients  $\beta_0$  and  $\beta_1$  were estimated via least squares estimation.



**Figure 4.2:** General workflow diagram of the *trans/cis* eQTL methodology. A total number of 1763 *cis* eQTLs identified in step 2 were used as “seeds” for subsequent *trans/cis* eQTL analysis. Thus, the 1364 *trans/cis* eQTLs SNP pairs contained one of the previously identified *cis* eQTLs.

In order to identify genes with *trans*-regulatory (modifying) effects, we used MB-MDR which can test for 2-way and 3-way interactions [19,37] with categorical input predictor variables. Briefly, MB-MDR performs dimensionality reduction procedure by pooling multi-locus genotype combinations into high, low and no-risk categories followed by association tests between trait and genotypes. For additional details about MB-MDR, we refer to [19,31,37] and Chapter 3. *Trans* loci were defined as follows: we used 2 Mb windows upstream and downstream of a gene’s ORF to delimit start and end locations of the *cis* region. Any SNP outside this region was considered to be in *trans* with respect to this gene. MB-MDR runs were, thus, able to identify those genes whose expression is regulated by the interaction of *trans* and *cis* eQTL SNPs (step 3 of Figure 4.2). To deal with multiple testing of 1763 genes (gene expressions), larger effect sizes with gene expression as trait values, and in particular to reduce the computational burden related to permutation-based significance assessment within MB-MDR, a two-stage interaction approach was developed. In the first stage, we identified potential *trans/cis* SNP pairs with significance assessment based on  $10^3$  permutations. In the second stage run, only the significant SNP pairs from the first run were used to produce final results based on  $10^7$  permutations. The MB-MDR run settings included continuous trait (`--continuous`), the *MAXT* algorithm to adjust for multiple testing errors (`-mt MAXT`),

co-dominant main effect correction ( $-a$  CODOMINANT). *MAXT* [31] was used, rather than the *gammaMAXT* algorithm [38] as only *trans* x *cis* eQTL interactions with a fixed list of 1763 *cis* eQTLs were considered reducing significantly the search space, and, hence, computational requirements.

#### 4.3.3. Controlling false positives

Statistical interactions can be produced by process other than true biological epistatic interactions including type I errors, presence of population structure in the data (i.e. population stratification), technological artifacts (e.g., batch effects, dye bias), linkage disequilibria between loci pairs, etc. Stringent experimental designed are required in order to account for all potentially false non-epistatic interactions, especially at the genome-wide scale with upper limit of 528,890 x 1763 interactions. The multiple-testing corrections both done at the *cis* eQTL and *trans/cis* eQTL inference stages. The raw *cis* eQTL *p*-values were FDR corrected at the 0.05 threshold following Benjamini and Hochberg method [30,32] The false positives in *trans/cis* eQTL runs were controlled via the *MAXT* algorithm [31] implemented in MB-MDR [19] .

Given the complex nature of analysis rigorous estimation of false positive rates of the *trans/cis* eQTL analysis were done on the permuted data. Using the 1763 *cis* eQTLs as initial seeds, for each *cis* eQTL a total of 100 *trans* x *cis* eQTL permutation-based epistasis screenings were performed, on data naively derived from the original data by only permuting the target gene's expression levels. Since the expression levels for each gene were permuted individually 100 times, the correlation structure between the *cis* genes was broken. Note that the correlation structure between all *cis* and *trans* eQTL SNPs was preserved. The permuted data was used to calculate various types of the family-wise error (FWER) defined as the probability of making one or more false discoveries, or type I errors, among all the hypotheses. Different FWERs can be computed highlighting possibility of having a false interaction within each *trans/cis* eQTL run ( $FWER_{within}$ ), between *trans/cis* eQTL runs ( $FWER_{between}$ ), or globally across all *trans/cis* eQTL runs ( $FWER_{global}$ ). In  $FWER_{within}$  we are limiting our attention to either all ~528,890 interaction hypotheses within each *trans/cis* eQTL run in isolation. For each *cis* gene the  $FWER_{within}$  was computed as the number of permutations with at least one significant *trans* x *cis* SNP pair at *p*-

value  $< 0.05$  out of 100 replica runs. The  $\text{FWER}_{\text{between}}$  took a wider context by accounting for all hypotheses of the 1763 *trans/cis* eQTL runs ( $\sim 528,890 \times 1763$ ). The  $\text{FWER}_{\text{between}}$  is defined as the number of *cis* eQTL genes with at least one significant *trans* x *cis* SNP pair at  $p$ -value  $< 0.05$  out of 1763. The  $\text{FWER}_{\text{between}}$  represents more closely the network inference context where both *trans-cis* and *cis-cis* interactions take place and where a given node can potentially participate in both types of interactions. Finally, the  $\text{FWER}_{\text{global}}$  represents the entire procedure and is the most stringent case since for each permutation (i.e. replica run with  $\sim 528,890 \times 1763$  hypotheses) a possibility of having at least one false positive across all 1763 *trans/cis* eQTL MB-MDR runs is tested.

The false discovery rates (FDR) within each *trans/cis* eQTL MB-MDR run and within each of the 100 replica runs was also computed. This provided a more refined picture on the error rates of our *trans/cis* eQTL procedure given the complex context associated to real-life data as described above. The FDR within each replica run was computed by counting the number of false positives *trans/cis* loci with  $p$ -value  $< 0.05$  over the total number of loci within a given replica. Each replica run with 1763 *trans/cis* MB-MDR eQTL runs contained approximately 877,615 loci pairs.

In addition to the classical permutation strategy with only the response variables being permuted, more refined strategies to generate replicates could have been adopted, such as obtaining a replicate by keeping a “*cis* region” – gene expression “pair” intact and permuting such pairs in the presence of all other SNPs. This ensures that correlation between *cis* eQTL gene and SNP is maintained, but not between the remaining *trans* SNPs. While other strategies are interesting venues to take, the permutation-based FWER and FDR results shown in Chapter 4 are based on the classical permutation scheme affecting only the expression values corresponding to the *cis* eQTL genes.

#### 4.3.4. SNP to gene mapping and pathway enrichment

The significant *trans* x *cis* eQTL SNP pairs ( $p$ -value  $< 0.05$ ) identified from 1763 selected *cis* eQTLs were pooled together. Next, the *trans/cis* eQTL SNP pairs were mapped to the nearest genes using *biomaRt* [39,40] and *GenomicFeatures* R libraries [41], leading to 1078 *trans* and 411 *cis* eQTL genes, respectively. Gene and pathway Venn diagrams were built with the help of *venneuler*

R library [42], to facilitate the overlap estimation between *trans* and *cis*-regulation components. The overlap is composed of 30 genes which were common to both *trans* and *cis* eQTL significant gene sets ( $p$ -value  $< 0.05$ ). These are all listed in Table S4.2. These genes both can act as *trans* and *cis* transcriptional regulators.

Pathway enrichment was performed by first mapping genes to Reactome pathways (human) extracted from MsigDb [43]. In particular, the aforementioned 1078 *trans* and 411 *cis* genes were mapped to 674 Reactome pathways. Fisher's exact test [44] was used to assess pathway enrichment of *cis* and *trans* eQTL gene sets against the entire set of 16,020 unique genes. The  $p$ -values thus obtained from 2x2 contingency tables, storing gene counts associated with a given pathway, were Bonferroni corrected for 674 Reactome pathways.

#### 4.3.5. Network analysis

We built a weighted directed gene-gene network  $G$  using the list of 1364 statistically significant *trans* x *cis* eQTL interaction SNP pairs (Figure S4.1 and Figure 4.8). The *igraph* package [45] was used to visualize and analyze the resulting network. The 1364  $p$ -values associated with those significant pairs were used as edge weights. This led to a network  $G$  with 1459 nodes and 1347 edges.

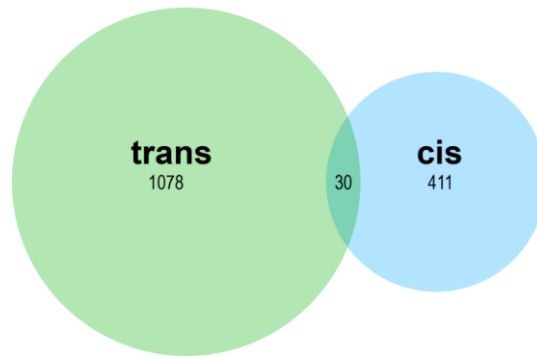
#### 4.3.6. Differential network analysis

We used differential network analysis (DNA) to investigate the impact of smoking. Such an analysis enables the identification of topological changes between gene networks derived from smokers and non-smokers. For  $NS$  and  $S$  groups, a weighted gene network  $G_{NS}$  and  $G_S$  was built using 1552 and 707 significant *trans* x *cis* eQTL SNP pairs, respectively. Again, we used the *igraph* package [45]. For differential network analysis (DNA) the  $G_{NS}$  and  $G_S$  networks were converted into unweighted networks and subsequently merged into one differential network  $G_D$  using the *XOR* rule. The *XOR* rule merging concentrates on edge differences between the  $G_{NS}$  and  $G_S$  networks. For example, if 1 denotes edge presence and 0 its absence, the following  $G_D$  will be obtained after application of the following rules to  $G_{NS}$  and  $G_S$ :  $\{0,1\}=1$  or  $\{1,0\}=1$ , but  $\{1,1\}=0$  and  $\{0,0\}=0$ . The resulting *XOR* differential network,  $G_D$ , highlights differences between  $NS$  and

*S* groups, potentially induced by condition-specific gene-gene interaction patterns and regulatory mechanisms.

## 4.4. Results

Using an initial pool of 1763 *cis* eQTLs, subsequent *trans*  $\times$  *cis* eQTL epistasis analyses identified a total of 1364 statistically significant *trans/cis* eQTL SNP pairs ( $p$ -value  $< 0.05$ ). These are listed in the Supplementary Table S4.1. These *trans*  $\times$  *cis* SNP pairs were mapped to the nearest genes as described in the Section 4.3.4.



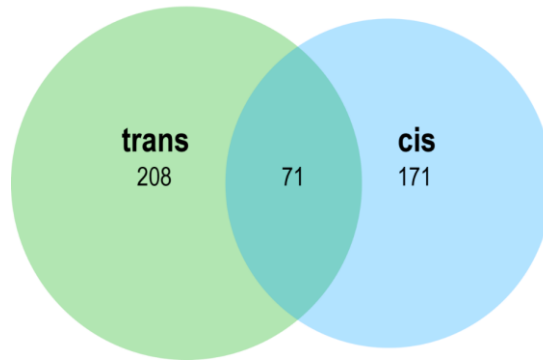
**Figure 4.3:** Overlap between the “*cis*” and “*trans*” eQTL gene sets of the significant 1364 *trans/cis* eQTL SNP pairs. The “*trans*” and “*cis*” gene set refers to genes associated with the significant *trans* SNPs of the 1364 *trans/cis* eQTL SNP pairs. The numbers represent the unique gene counts. The overlap area was 2.014 % (30 genes) of the total combined *trans* and *cis* areas. Common genes are listed in Table S4.2.

This resulted in *cis* and *trans* separate gene sets. The overlap reached 2.014 % of the total area (Figure 4.3). Annotated 30 genes common to both *trans* and *cis* gene sets are listed in supplementary Table S4.2.

### 4.4.1. Pathway enrichment

From the 1364 *trans/cis* eQTL SNP pairs, pathway enrichment analysis identified significantly enriched pathways partially listed in Table 4.1 (top 20 are shown). Pathway enrichment analysis showed strong involvement of the immune system and signaling components in both *trans* and *cis* gene sets. The *trans* gene set was dominated by central nervous and cell-cell communication pathways while the *cis* gene set was dominated by cell cycle and expression control pathways. The

overlap between *cis* and *trans* pathways is shown by the Venn diagram for both groups in Figure 4.4. These 71 overlapping *cis* and *trans* pathways are listed in Table S4.3. Alternatively, the 5 most significant pathways enriched in the gene set common to both *trans* and *cis* gene sets are immune system (R-HSA-168256), adaptive immune system (R-HSA-1280218), class I MHC mediated antigen processing presentation (R-HSA-983169), DNA repair (R-HSA-73894), insulin receptor signalling cascade (R-HSA-74751) (see Table 4.1 and Table S4.3).



**Figure 4.4:** Overlap between “*cis*” and “*trans*” significantly enriched pathways obtained from the list of 1364 *trans/cis* eQTLs. The numbers refer to significantly enriched pathway counts. The overlap area is 18.7% (71 genes) of total combined *trans* and *cis* areas. Pathways are listed in Table S4.3.

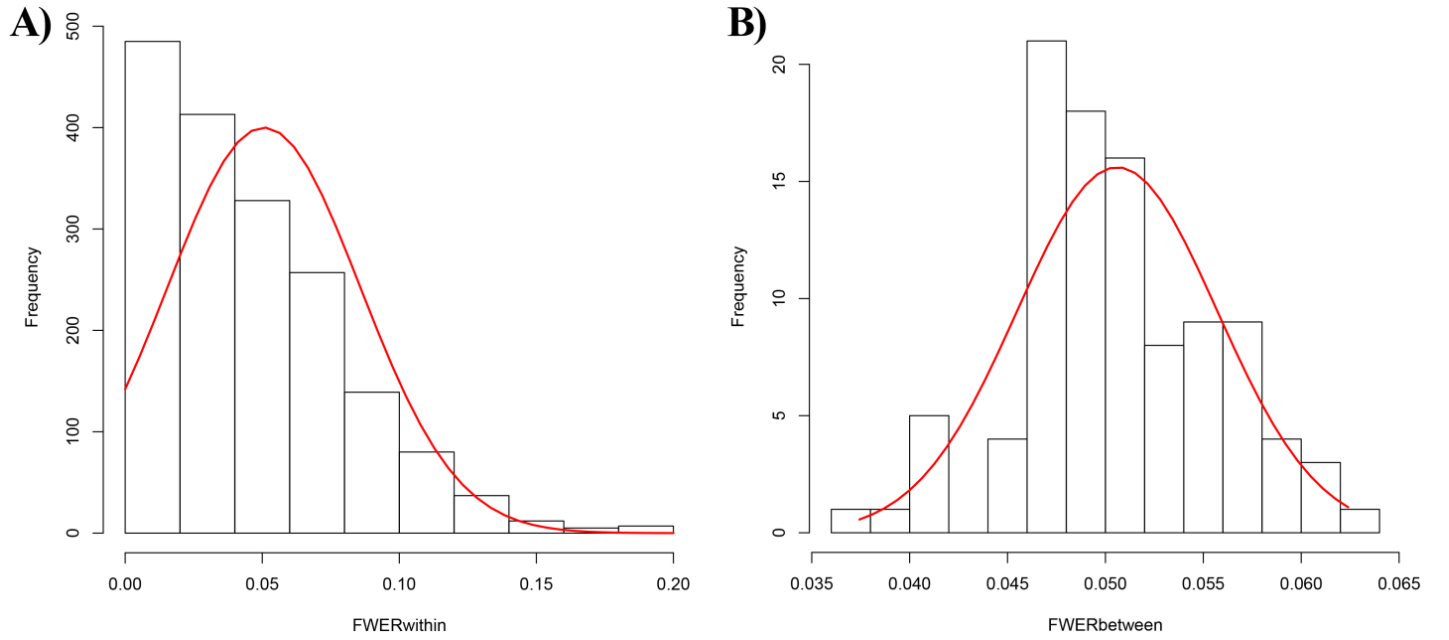
#### 4.4.2. Control of type I error rates

Three types of FWER were computed ( $\text{FWER}_{\text{within}}$ ,  $\text{FWER}_{\text{between}}$  and  $\text{FWER}_{\text{global}}$ ) with progressively stringent conditions as described in Section 4.3.3. The  $\text{FWER}_{\text{within}}$  estimates error rates within each *trans/cis* eQTL run ignoring existence of other epistatic runs. The  $\text{FWER}_{\text{within}}$  is calculated as the number of permutation runs with at least one type I error hypothesis out of 100 permutation replica runs. The  $\text{FWER}_{\text{between}}$  estimates type I error rates considering all 1763 *trans/cis* eQTL runs and is calculated as the number of eQTL runs with at least one false positive signal out of 1763. The  $\text{FWER}_{\text{global}}$  estimated type I errors globally across all 1763 *trans/cis* eQTL epistatic and 100 permutation replica runs (176,300). Thus, the  $\text{FWER}_{\text{global}}$  is the ratio out of 100. In addition, the false discovery rate (FDR) per replica was also computed to provide a more refined detail on the number of false positives per each replica. Compared to FWER, FDR procedures are less stringent providing a greater number of epistatic signals albeit at potentially greater type I error rates.

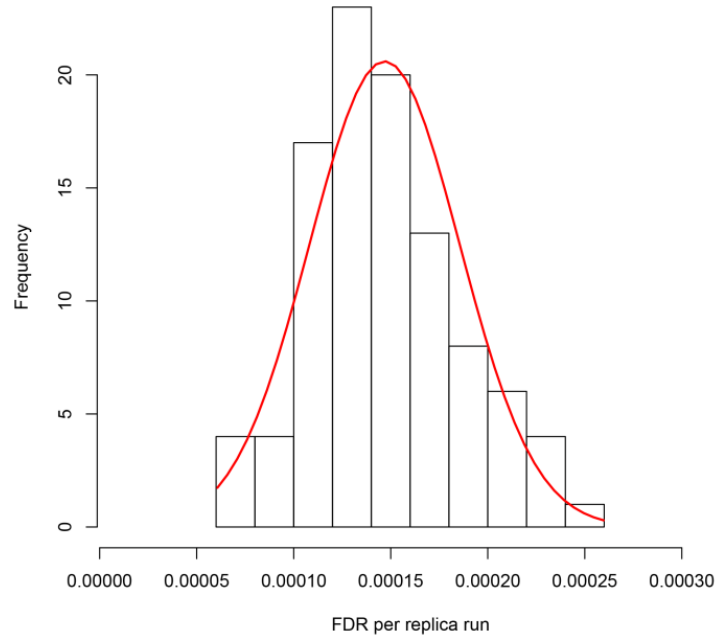
The median of 1763  $\text{FWER}_{\text{within}}$  values were 0.04 at 0.05 significance threshold ( $p\text{-value} < 0.05$ ), meanwhile the median on 100  $\text{FWER}_{\text{between}}$  was 0.05. The  $\text{FWER}_{\text{within}}$  was within the MB-MDR advertised  $\text{FWER} < 0.05$  under co-dominant main effects correction mode [46]. The most stringent  $\text{FWER}_{\text{global}}$  taking into account the whole procedure including 100 replicates with 1763 runs each was 1. This means that within each permuted replica there was at least one *trans/cis* eQTL SNP pair with  $p\text{-value} < 0.05$ . The distribution of  $\text{FWER}_{\text{within}}$  and  $\text{FWER}_{\text{between}}$  together with their corresponding density functions are shown in Figure 4.5.

The *trans/cis* eQTL analysis median per replica false discovery rate (FDR) based on 100 permutations across 1763 *trans/cis* eQTLs was 0.000143. The median FDR per replica run shows approximate number of false positive signals that one might expect from the entire procedure across the 1763 *trans/cis* eQTL runs. The histograms showing the distribution of FDR per replica run are presented in Figure 4.6. The FDR median and mean considering individually each *trans/cis* eQTL MB-MDR run over the 100 permutations (a total of 1763x100 FDR values) was 0 and 0.0001543062, respectively.

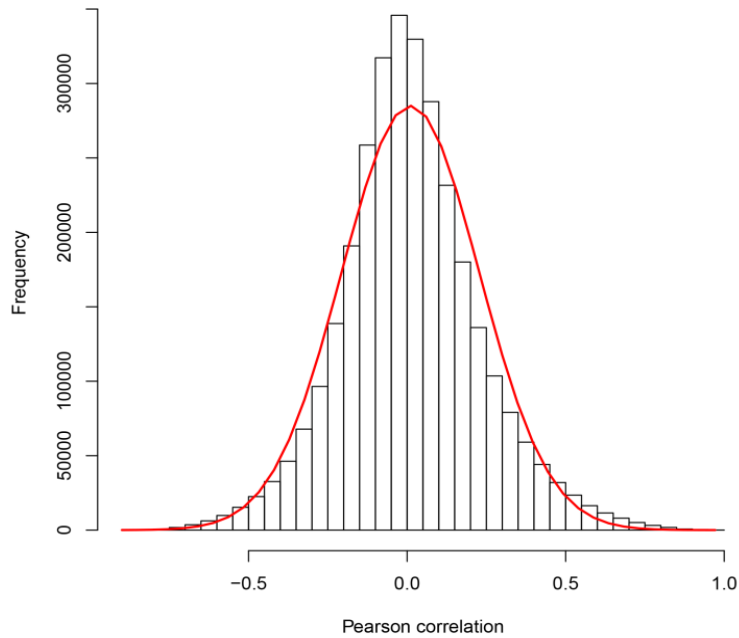
The correlation between the expression traits of the 1763 eQTL genes was rather low. The computation was done amongst the 1763 x 1763 possible pairs (1,554,084). The 25<sup>th</sup> and 75<sup>th</sup> quantile reached values of -0.12 and 0.13. The respective distribution of Pearson correlation values across all pairs is shown in Figure 4.7 with median at  $7.201 \times 10^{-5}$ , mean at 0.01044, and standard deviation at 0.217.



**Figure 4.5:** Distribution of  $\text{FWER}_{\text{within}}$  and  $\text{FWER}_{\text{between}}$  in subplots A) and B), respectively. A) the  $\text{FWER}_{\text{within}}$  mean and medians are 0.0506 and 0.04; B) the  $\text{FWER}_{\text{between}}$  mean and medians are 0.0506 and 0.0501; The density function is shown in red.  $\text{FWER}_{\text{within}}$  and  $\text{FWER}_{\text{between}}$  see Section 4.3.3. Note we consider a false positive result any *trans/cis* loci pair with  $p\text{-value} < 0.05$ . The FWER values are computed based on complete 100 permutation-based *trans/cis* eQTL replica runs on the null data where only response variable was permuted as described in Section 4.3.3.



**Figure 4.6:** Distribution of false discovery rate (FDR) per 100 permutation replicas each containing 1763 *trans/cis* eQTL runs. The FDR was defined as number of false positives within each replica permutation run containing approximately 877,615 *trans/cis* loci pairs. FDR per replica run is defined in Section 4.3.3. The false positive result is any *trans/cis* loci pair with  $p$ -value  $< 0.05$ . The red line represents the density function.



**Figure 4.7:** Distribution of the Pearson correlation values of all unique *cis* eQTL gene pairs (1,554,084). The mean and median was 0.0104 and 0 with standard deviation of 0.21, respectively.

#### 4.4.3. Network analysis

The *trans/cis* eQTL graph  $G$ , derived from the earlier obtained 1364 significant *trans* x *cis* SNP x SNP interactions, contained a total of 1459 nodes and 1347 edges (see Figure S4.1 and Figure 4.8). Only nodes with 2 or more edges are shown. Nodes with degree  $> 8$  are shown in red while nodes with degree  $> 2$  are shown in orange. The obtained network did not exhibit cliques with  $> 2$  nodes. The network  $G$  reached a graph density of  $6.33 \times 10^{-4}$  (graph density is the ratio of the number of edges in a given graph over the total number of possible edges) and can be assumed to have a scale-free topology (Figure 4.9). This was confirmed by fitting power law to the 1459 total degree values of the graph ( $p(k) \sim k^{-\gamma}$ ). The Kolmogorov-Smirnov test statistic was 0.0968 with the corresponding  $p$ -value 0.9998 and exponent  $\gamma = 1.6$ . The typical value of scale-free networks the  $2 < \gamma < 3$  [47]. Based on this data, the null hypothesis was not rejected and *trans/cis* eQTL graph  $G$  scale-free topology confirmed.

#### 4.4.4. Differential network analysis

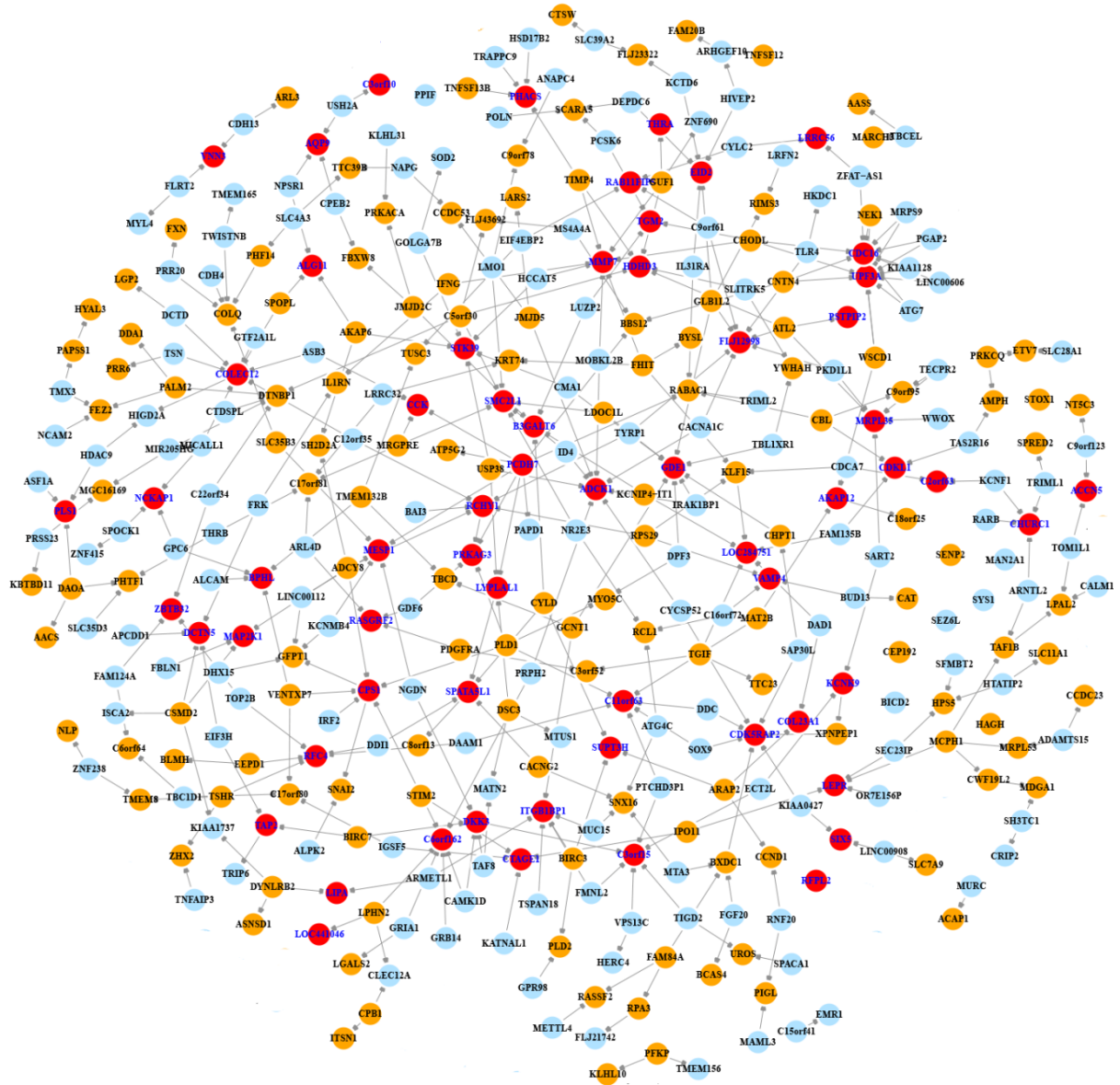
Directed gene networks,  $G_{NS}$  for non-smoker  $NS$  and  $G_S$  for smoker  $S$  separately, were based on 1552 and 707 significant *trans* x *cis* SNP interaction pairs, respectively (Figures S4.2 and S4.3). The network  $G_{NS}$  contained a total of 1492 nodes and 1474 edges. The  $G_S$  network included a total of 907 nodes and 689 edges (Figures S4.2 and S4.3). As can also be observed from their graphical presentations,  $G_{NS}$  was more sparsely connected than  $G_S$  with respective  $6.62 \times 10^{-4}$  and  $8.38 \times 10^{-4}$  graph densities. A total of 322 nodes with degree  $\geq 1$  were common to both  $G_{NS}$  and  $G_S$ . These common nodes are listed in Table S4.5.

In order to highlight condition-specific differences between non-smokers and smokers a differential network  $G_D$  was built from  $G_{NS}$  and  $G_S$ . The resulting  $G_D$  network contained a total of 331 nodes and 291 edges (Figure 4.10). The  $G_D$  average graph density was estimated at  $5.32 \times 10^{-3}$  and was 6 times higher than that of the  $G_{NS}$  and  $G_S$ . The largest maximally interconnected sub-graph (i.e. clique) was the only one and contained the following genes *BCL11A*, *LEPR* and *PPP1R13L*. The highest total degree nodes included *BCL11A*, *HDCC2*, *SERAC1* and *TMEM136*.

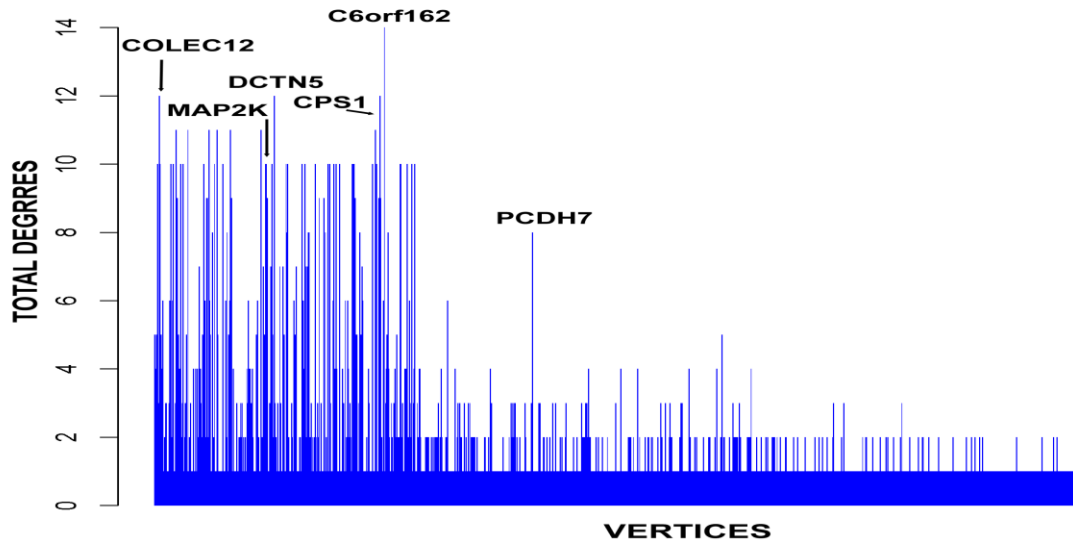
**Table 4.1:** Top 20 significantly enriched Reactome pathways in *trans* and *cis* sets from 1364 *trans/cis* eQTLs

<i>trans</i> pathways			<i>cis</i> pathways		
ID	pathway name	<i>p</i> -value*	ID	pathway name	<i>p</i> -value*
<b>R-HSA-168256</b>	<b>IMMUNE SYSTEM</b>	1.22E-50	R-HSA-69278	CELL CYCLE MITOTIC	3.78E-57
R-HSA-372790	SIGNALING BY GPCR	4.82E-46	<b>R-HSA-168256</b>	<b>IMMUNE SYSTEM</b>	8.97E-53
R-HSA-112316	NEURONAL SYSTEM	5.16E-45	<b>R-HSA-1280218</b>	<b>ADAPTIVE IMMUNE SYSTEM</b>	2.09E-47
R-HSA-500792	GPCR LIGAND BINDING	9.73E-35	R-HSA-69278	CELL CYCLE	8.25E-46
R-HSA-382551	TRANSMEMBRANE TRANSPORT OF SMALL MOLECULES	3.59E-33	R-HSA-983169	CLASS I MHC MEDIATED ANTIGEN PROCESSING PRESENTATION	7.19E-41
R-HSA-1266738	DEVELOPMENTAL BIOLOGY	9.39E-33	R-HSA-380259	LOSS OF NLP FROM MITOTIC CENTROSOMES	8.62E-32
R-HSA-388396	GPCR DOWNSTREAM SIGNALING	4.54E-32	R-HSA-380270	RECRUITMENT OF MITOTIC CENTROSOME PROTEINS AND COMPLEXES	9.43E-30
R-HSA-373076	CLASS A1 RHODOPSIN LIKE RECEPTORS	1.39E-26	R-HSA-73894	DNA REPAIR	2.41E-28
R-HSA-416476	G ALPHA Q SIGNALLING EVENTS	1.44E-26	R-HSA-74751	INSULIN RECEPTOR SIGNALLING CASCADE	1.21E-27
R-HSA-112315	TRANSMISSION ACROSS CHEMICAL SYNAPSES	2.80E-26	R-HSA-983168	ANTIGEN PROCESSING UBIQUITINATION PROTEASOME DEGRADATION	2.21E-27
<b>R-HSA-1280218</b>	<b>ADAPTIVE IMMUNE SYSTEM</b>	2.19E-24	R-HSA-392499	METABOLISM OF PROTEINS	7.37E-27
R-HSA-881907	GASTRIN CREB SIGNALLING PATHWAY VIA PKC AND MAPK	7.82E-24	R-HSA-166520	SIGNALLING BY NGF	1.48E-26
R-HSA-112314	NEUROTRANSMITTER RECEPTOR BINDING AND DOWNSTREAM TRANSMISSION IN THE POSTSYNAPTIC CELL	4.52E-22	R-HSA-453274	MITOTIC G2 M PHASES	5.53E-26
R-HSA-1296071	POTASSIUM CHANNELS	5.37E-21	R-HSA-1474244	EXTRACELLULAR MATRIX ORGANIZATION	4.32E-25
R-HSA-392499	METABOLISM OF PROTEINS	6.01E-20	R-HSA-69620	CELL CYCLE CHECKPOINTS	1.45E-24
R-HSA-109582	HEMOSTASIS	2.53E-19	R-HSA-6782210	GAP-FILLING DNA REPAIR SYNTHESIS AND LIGATION IN TC-NER	4.58E-24
(R-HSA-373752	NETRIN1 SIGNALING	6.25E-19	R-HSA-446652	<b>IL1 SIGNALING</b>	1.62E-23
R-HSA-212436	GENERIC TRANSCRIPTION PATHWAY	1.07E-18	R-HSA-74752	SIGNALING BY INSULIN RECEPTOR	1.66E-23
R-HSA-418990	ADHERENS JUNCTIONS INTERACTIONS	2.89E-18	R-HSA-69183	LAGGING STRAND SYNTHESIS	4.76E-23
R-HSA-168638	NOD1 2 SIGNALING PATHWAY	2.89E-18	R-HSA-6783310	REGULATION OF THE FANCONI ANEMIA PATHWAY	5.15E-23

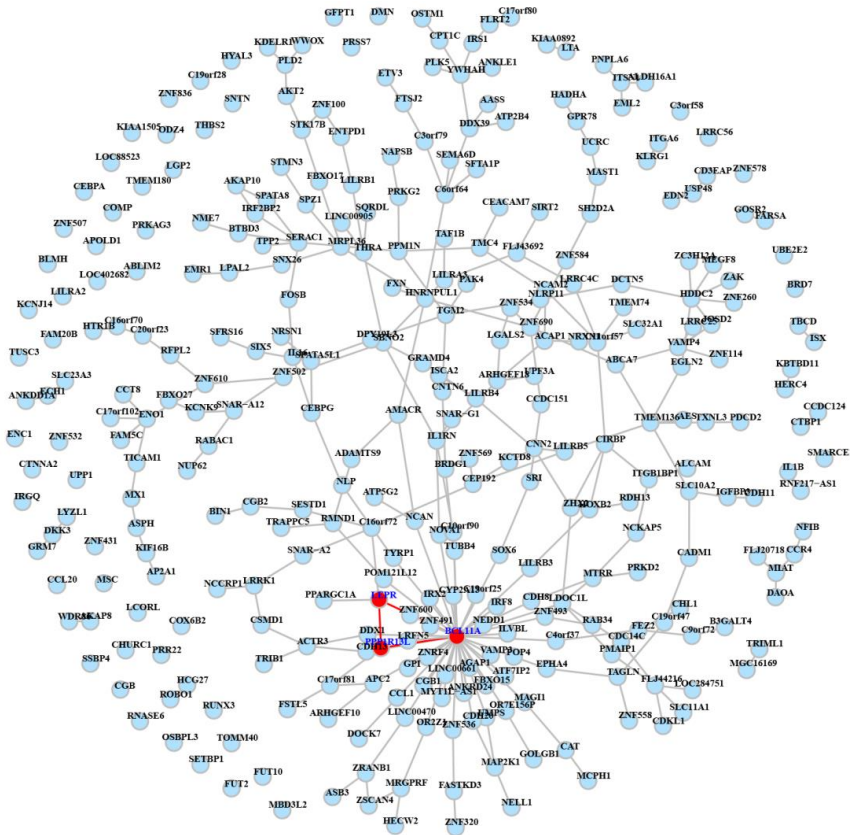
**Note:** \* Fisher's exact *p*-values were Bonferroni adjusted to the total number of Reactome pathways (674). The pathways in **bold** are related to immune system.



**Figure 4.8:** Directed *trans/cis* eQTL network  $G$  composed of 1459 nodes. Nodes with degree  $\geq 2$  are shown. No cliques containing more than 2 nodes were found. Nodes with degree  $\geq 8$  are shown in red while nodes with degree  $\geq 2$  but  $< 8$  are shown in orange. The node names correspond to gene symbols representing SNPs mapped to the nearest genes (see Methods).



**Figure 4.9:** Total degree distribution for the *trans/cis* eQTL network  $G$ . The total node degrees represent the sum of the “in” and “out” degrees. The associated genes of the selected highest degree nodes are indicated. The total degree distribution is also available in online Table S4.4.



**Figure 4.10:** Differential undirected network  $G_D$  built from  $G_{NS}$  and  $G_S$  networks. The  $G_D$  highlights the three nodes *LEPR*, *BCL11A*, *PPP1R13L*, members of the unique largest clique of size 3. No other cliques were present in the  $G_D$ . The node names correspond to gene symbols.

## 4.5. Discussion

One of the main bottlenecks in large-scale eQTL epistasis analysis is the huge number of tests that need to be performed. This causes computational challenges and demands on IT-infrastructures, as well as statistical challenges in that adequate corrections for multiple testing need to be made. As indicated in the introduction section, different routes can be travelled by to deal with multiple testing challenges. Some solutions are pragmatic such as arbitrary FDR threshold selection at 0.05 in ‘either significant’ and ‘both significant’ [9] eQTL mapping solutions [23,24], while others are more elaborate including integration of prior knowledge (e.g., protein-protein interaction networks)[28]. The problems of adequate multiple testing procedures application and selection of sensible methods to detect biologically relevant epistatic eQTLs are not entirely independent from each other. In addition to classical the regression-based tests discussed in Section 4.2 to detect eQTLs, there are several others such as the median test (MED) and tests based on mutual information (MI) (see [48]). The MB-MDR was not used to detect marginal effects (1D) of the *cis* eQTLs, as it was not intended for 1D screening of SNPs and, hence, also does not perform well in this context. MB-MDR was used during 2D screening of *trans/cis* epistatic eQTLs as described in Section 4.3.2. As discussed in Section 4.2, the majority of eQTL epistasis runs are performed in a regression context using additive encoding or a combination of additive and dominant encodings as in [25]. As shown in Chapter 3, additive encoding schemes may elevate the number of false positives, regardless of the adequacy of the multiple testing correction [49,50].

A typical eQTL and epistatic eQTL study is faced with a huge number of hypotheses tests between genetic loci and expression traits requiring sensible multiple testing correction strategy. Classical Bonferroni correction are often overly conservative. In addition, transcriptome and genome data presents complicated correlations, population stratification, linkage disequilibrium between loci pairs, batch effects and others existing further complicating multiple testing problem. These underlying sources of error can lead to statistical artefacts (i.e. false positives) [51]. In addition to FDR, permutation-based and marker selection methods extensively used in eQTL studies introduced in Section 4.2, other approaches are also used to simulate the null distribution of either responses or predictors via resampling-based permutation or bootstrap procedures [51]. Nevertheless, the later multiple-testing procedures are computationally prohibitive due to large-

scale eQTL data sets. Several heuristic approaches were introduced to alleviate computation burden and to provide approximate  $p$ -values. One of them relies on Monte Carlo to approximate the distribution of the test statistic [52], while others use a Bayesian approach to estimate model parameters such as in [53]. As demonstrated by [54], a more liberal FDR approach adopted by many eQTL studies as compared to FWER. Nevertheless, FDR, used as a measure of global error, often fails to control false discoveries at the set threshold (e.g., 0.05) due to the improper definition of true discovery [54]. Peterson *et al.* suggests to divide hypotheses into “families” and estimate FDR within each family via the proposed hierarchical testing procedure [54].

Since our epistatic eQTL MB-MDR based methods uses *MAXT* [31] for multiple-testing control and significance assessment, the permutation methods merit further discussion. As defined in [55] permutation runs fall in three categories: (1) a direct permutation scheme that relies on a fixed number of permutations; (2) an adaptive permutation scheme which maintains a reasonable computational load by adjusting the number of permutations to the significance level of the eQTL pairs; (3) a beta approximation which models the permutation outcome via a beta distribution. These permutation suggestions are made within a linear regression context. In a framework as provided by MB-MDR, the approaches implemented in the MB-MDR software [31] are (1) and (3), with possibility of using a gamma distribution instead of a beta distribution [38]. The reason for not using the permutation scheme (2) the latter is that our order statistics from an epistasis MB-MDR screen with *MAXT* correction cannot be considered to be coming from independent loci (multilocus seen as a new variable) and, hence, can no longer be assumed to follow beta-distribution.

Despite a large number of multiple testing solutions, there is not a readily a way to translate multiple testing correction methods from the single locus main effects eQTL scene (1D) to the epistasis eQTL one (2D). In the future we expect development of more tailored multiple testing correction approaches for epistatic eQTL studies. In addition there exist approaches [56,57] that explicitly take into account the linkage disequilibrium structure among variants which is also one of the potential causes of elevated false positives. Therefore, the impact of LD on “epistasis” findings needs further investigation in eQTL settings. In Chapter 3 we have shown that MB-MDR is only modestly impacted by LD provided prior biological knowledge is used. More work is

needed to investigate optimal approaches to incorporate prior knowledge into eQTL epistasis screenings.

Our epistatic eQTL MB-MDR method implements adaptive permutation scheme (2) [55] via the step-wise approach based on  $10^3$  and  $10^7$  permutation steps. The adapted adaptive permutation scheme uses the same number of samples at each stage while decreasing the total number of predictor loci. Contrary to [55] we adopted a pragmatic strategy with fixed the number of permutations in order to reach as high as possible resolution. The resulting approach invests more time over significant hits and discards insignificant ones allowing to reach a high number of permutations (up to  $10^7$ ) at reasonable computational cost. The adopted epistatic eQTL MB-MDR based analysis used a large number of permutations and rigorous FWER checks since the *MAXT* permutation based  $p$ -values are dependent on the total number of predictor variables. This was also observed in Chapter 3 on unpruned and LD pruned datasets (see Table 3.1). Thus, one needs to allow for fluctuation of  $p$ -values and semi-empirically estimate possibility of having false positive interaction at the significance threshold of 0.05 which was extensively done in Section 4.4.2.

As first rigorous measure the Bonferroni correction for the number of *cis* expression traits considered (here 1763), though computationally trivial, is overly conservative ( $p$ -value  $< 2.83 * 10^{-5}$ ) and fails to account for correlations between genes at the transcriptome level, which do exist as shown in subsequent Chapters 5 and by numerous studies [2,3,47,58]. For instance, when supplementing our strategy with the Bonferroni correction at the significance threshold of  $2.83 * 10^{-5}$ , only 3 SNP pairs were identified shown in Table S4.6. As stated earlier, Bonferroni correction especially in epistatic scenarios is very conservative. In order to decrease the number of effective tests and, thus, increase the significance threshold, several groups have proposed the use of the effective number of markers ( $M_{\text{eff}}$ ) for the adjustment of multiple testing. We adopted the  $M_{\text{eff}}$  method by [59] that takes into account the correlation patterns amongst the expression traits which, in our case, are *cis* eQTLs. As demonstrated by the  $\text{FWER}_{\text{within}}$  results and previous studies [46,60], the *trans/cis* epistatic eQTL MB-MDR runs have already been adequately corrected for multiple testing. Thus, the only multiplicity not adjusted for is the correlatedness between the gene expressions. Due to low correlation between the *cis* eQTL genes as shown in Figure 4.7 there was no significant drop in the threshold ( $p$ -value  $< 3.05 * 10^{-5}$ , compare to  $p$ -value of  $2.83 * 10^{-5}$ )

resulting in the identification of the same 3 epistatic SNP pairs (Table S4.6). The  $\text{FWER}_{\text{global}}$  at  $p$ -value  $< 3.05 * 10^{-5}$  and  $< 0.05$  cut-offs were 0.02 (2/100) and 1 (100/100). The  $\text{FWER}_{\text{global}}$  as described in Section 4.3.3 was obtained for the entire procedure including 1763 *trans/cis* eQTL MB-MDR and 100 permutation-based replica runs. Considering the extremely large number of tests, the  $\text{FWER}_{\text{global}}$  is a very stringent condition causing likely to miss important biological epistatic interactions. Thus, the  $\text{FWER}_{\text{between}}$  controlling for error rates within and between the 1763 *trans/cis* epistatic eQTL runs is a more realistic and adequate quality control check. The  $\text{FWER}_{\text{between}}$  even at threshold of 0.05 reached median of 0.04 (88.5/1763) as indicated in Figure 4.5. Thus our findings at relaxed threshold condition set at  $p$ -value  $< 0.05$  are valid and adequately controlled against multiplicity and correlation effects as discussed above. The  $\text{FWER}_{\text{within}}$  at 0.05 confirms MB-MDR ability to control false positives at 0.05 level.

In addition to statistical complexities of eQTL studies one also needs to assess biological relevance of results in the dataset which, in our case, is asthma. The final result of our *trans/cis* epistatic eQTL method is transcriptional gene regulatory network further explored in Chapters 5 and 6. Transcriptional regulation of genes is complex and assumes interactions within transcriptome and other omics information layers. Genes can be regulated by nearby gene regulatory elements such as promoters, but also by enhancers located thousands of kilobases (kb) away from a given gene [61]. Hence, both *cis*-regulatory eQTL [62,63] and *trans*-regulatory eQTL analyses [64] are useful. In general, processes of gene expression can be rather complex and involve both epigenetic and interaction components. For instance, epigenetic components such as methylation and histone modification are also known to play important roles in gene expression regulation. Not as thoroughly investigated are large-scale genetic epistasis screenings for transcriptomes. One of the issues is related to epistasis detection analytics itself. A multitude of epistasis analytics are available, each of them potentially highlighting differential genetic architectures underlying interaction mechanisms. The performance of parametric regression-based analytics seems to be more depending on the underlying genetic models, as compared to semi- or non-parametric methods [65,66]. Also, the most heavily used epistasis detection tools (often regression-based) target linear interactions only. Moreover, performance may highly depend on the adopted multiple comparison adjustment procedures, which is already substantial in genome-wide epistasis screening, but becomes even more elevated when gene expressions are alternatively considered as

“traits” in the epistasis screening. Therefore, in this work, we have used the MB-MDR framework as an epistasis screening tool, hereby overcoming several of the aforementioned shortcomings and allowing for confounder adjustments. A second issue encapsulates analytics and concerns the differential use of epistasis screening protocols. We have shown that minor changes in such protocols (e.g., Gusareva *et al.* [49]), combined with regression-based (BOOST)[67] or non-parametric dimensionality reduction methods (MB-MDR), can have a dramatic impact on findings (see also Chapter 3 and [50]). One of these changes is data reduction. Though the more reliable biological prior information is used in the process, the more stable the results are obtained via different protocols. Therefore, we propose to first screen for *cis* eQTL effects, define *trans* regions, and to look for interaction between *trans* x *cis* eQTL SNPs in relation to the quantitative trait of the *cis* eQTL. This implies that the all of the considered interactions will have at least one significant main effect. Our *trans* x *cis* eQTL epistasis screening relies on a previously published GWAIS protocol which advocates mild LD pruning at a threshold of  $r^2 > 0.75$  [49]. This protocol also advocates restricting attention to  $MAF \geq 0.05$ . However, we took a more stringent decision and only included SNPs with  $MAF \geq 0.20$ . Our motivation was two-fold: 1) to increase the utility of results in clinical practice and 2) to avoid an abundance of false positives (data not shown). Notably, by only pairing SNPs to *cis* eQTLs that are outside a sufficiently large *cis*-region (2 Mb upstream and downstream of the gene ORF), we dramatically reduce the emergence of spurious interactions (redundant interactions caused by LD between markers). Since thousands of interactions need to be evaluated, adequate assessment of significance is a major concern. For the step-down *MAXT* approach as implemented in MB-MDR, FWER is strongly controlled provided the assumption of subset pivotality holds [68]. However, several *cis* genes are considered interchangeably as traits, imposing as many GWAIS as there are *cis* genes. This implies that all MB-MDR *MAXT* corrected *p*-values, still need to be adjusted for the number of GWAIS, so as to keep FWER under control. We did so using Bonferroni correction. Concerning the *MAXT* corrected *p*-values, we observed the need to implement a large number of permutations. This need is caused by the large number of interaction hypothesis tested ( $\sim 528,890 \times 1763$ ) across all 1763 *trans/cis* eQTL runs. Although no exhaustive epistasis screening is performed, the level of accuracy needed in MB-MDR *p*-value estimation and associated computational burden is still cumbersome. Therefore, we first assessed interaction significance on the basis of  $10^3$  permutations and followed up significant findings with  $10^7$  permutations. Note that  $0.05/1763 = 2.8 \cdot 10^{-5}$ . Even though the

effect sizes for main gene expression traits (i.e. *cis* eQTLs) are larger than those for complex disease traits explored in Chapter 3, due to the fewer number of contributing factors, in case of the *trans/cis* eQTLs the threshold was too strict resulting in only 3 significant pairs (Table S4.6). We acknowledge that this procedure, which involves testing a subset of pairs, may give rise to overly optimistic *p*-values. However, when reducing the SNP space via LD pruning and thus testing a smaller number of SNP pairs, MB-MDR typically gives rise to a smaller number of significant findings (See Chapter 3 - Figure 3.2). Regardless, to test FWER control of our implemented *trans/cis* eQTL epistasis protocol, we estimated FWER on 100 null data replicates.

The overlap between *trans* and *cis* gene sets generated for the significant *trans/cis* SNP pairs was quite low 2.014% (Figure 4.3) but was considerably higher at the pathways level reaching 18.7% figure (Figure 4.4). A closer look at the enriched pathways in *trans* and *cis* gene sets highlighted common functionalities. In particular, the immune and signaling components were strongly shared between *cis* and *trans* gene sets: antigen processing (REACTOME: R-HSA-983168) and adaptive immunity (REACTOME: R-HSA-1280218) (Table 4.1). Interestingly, the *trans* gene set was the most significantly enriched for transcription regulation pathways (REACTOME: R-HSA-212436) (Table 4.1), and may point towards significant modulation effects on transcription regulation exerted by *trans* components. In addition, the *trans* gene set was also enriched for cell adhesion pathways (REACTOME: R-HSA-418990). The involvement of cell adhesion pathways in asthma was highlighted by previous studies: Bentley *et al.* found elevated expression of *ICAM-1* and *VCAM-1* adhesion molecules by endothelial cells in asthma patients [5]. Analysis of the *cis* gene set showed significant enrichment in signaling pathways (Table 4.1). In particular, the involvement of G-protein coupled receptor (GPCR) via signaling via GPCR pathway (REACTOME: R-HSA-372790) was identified and confirmed by previous GWAS study in relation to asthma [69] (Table 4.1). In addition, *cis* eQTLs showed particularly strong enrichment in cell cycle and DNA damage repair pathways including G1 phase (REACTOME: R-HSA-69236), and inhibition of replication initiation of damaged DNA (REACTOME: R-HSA-113501) (Table 4.1).

In the context of asthma, 0.0055 % of the *trans* genes (CMA1, CRB1, SETDB2, IFNG, HLA-DRA, CCL2) and 0.0073 % of the *cis* genes (SOD2, GSTT1, IL1RN) are known asthma genes, none of the 30 *cis* genes that were also active as *trans* modifiers are known asthma genes as per

123 disease-causing genes identified in admixed USA population listed in the interactive Asthma Gene Browser [70]. Cadherins are important for asthma in the maintenance of epithelial tissue integrity [71]. This was supported by significant enrichment of genes of the cell-cell junction organization pathway (REACTOME: R-HSA-421270) in *trans* genes ( $p$ -value of  $1.22 \times 10^{-14}$ ). Cadherins *CDH13*, *CDH4* and *CDH6* obtained degrees 2, 2 and 1 in  $G$  respectively (Figure 4.9 and Table S4.4). Nodes *MAP2K1*, *DDA1* and *DCTN5* in  $G$  received comparatively high total degrees of 10, 3 and 12, respectively. *MAP2K1* encodes mitogen-activated protein kinase 1 and is one of the key signaling enzymes highlighting importance of the signaling component in asthma [72]. Other highly interconnected genes included *DDA1* - DET1 and DDB1 Associated 1 and *DCTN5* - Dynactin 5. *DDA1* is involved in degradation while *DCTN5* participates in cellular cargo movements thanks to the cytoplasmic dynein motor machinery. The largest maximally interconnected sub-graph (i.e. maximum clique) in  $G$  consisted of no more than 2 genes indicating rather the strong sparsity of the *trans/cis* eQTL network  $G$  (Figure 4.8). In general, 0.0431% (63/1459) of the genes in network  $G$  with degree  $\geq 8$  are reported known asthma genes.

Smoking stratified analysis revealed both topological and biological differences. The non-smoker network  $G_{NS}$  has a larger sparsity compared to  $G_S$  for smokers, although both networks were built on similar numbers of individuals. Interestingly, the size of the  $G_{NS}$  network is larger than  $G_S$ , but the degree distribution for  $G_{NS}$  (Figure S4.2) shows higher degrees than for  $G_S$ . The lower sparsity of the  $G_S$  is partially due to the *MAP2K1* which has an extremely high total degree of 160 (Figure S4.3). The largest maximally interconnected sub-graph (i.e. maximum clique) in the differential network  $G_D$  derived from  $G_{NS}$  and  $G_S$  consisted of *LEPR*, *BCL11A*, *PPP1R13L* genes (Figure 4.10). Interestingly all 3 genes are linked to the immune system, specifically, lymphomas are characterized by abnormalities in T and B cells. *LEPR* is leptin receptor involved in the regulation of fat metabolism and normal generation of lymphocytes [73]. There is a very strong relationship between obesity, leptin and asthma. Obesity is a major asthma susceptibility risk factor and a modifier of asthma control and severity. Obese asthmatics also demonstrate greater resistance to glucocorticoid therapy. Leptin levels are directly correlated with degree of obesity, and there is considerable data demonstrating an important role for Leptin in asthma and allergic inflammation: leptin levels are increased in allergic airway inflammation, and increasing leptin levels augment the allergic response in mice [74]. In humans, leptin levels in obese asthmatics are higher than in

obese non-asthmatics, and compared to controls, and leptin is a potent stimulator of macrophage-induced inflammation [75]. The leptin receptor is expressed on bronchial epithelium, but this expression decreases with increasing asthma severity [76]. *BCL11A* is B-cell CLL/lymphoma 11A zinc finger protein (i.e. transcription factor) modulating responses of B cells through *IL4* and, thus, is vital for normal T-cell, B-cell and dendritic cell development [77]. *BCL11A* is often hyper-mutated and translocated within B-cell heavy chain. This marker has a strong association with the B cell malignancies [78]. In addition, *BCL11A* is a critical activator of RAG1 and RAG2 in B cells [79], and *BCL11A* deficiency in experimental models results in a marked reduction in the ability of the immune system to generate an antiviral response [80].

*PPP1R13L* codes protein phosphatase 1, regulatory subunit 13 like that interacts with B-cell lymphoma (*BCL*) family of genes [81]. Similar to *BCL11A*, its translocations are strongly associated with B-cell malignancies. In addition, *PPP1R13L* is associated with obesity and lipid metabolism [82]. The gene, also known as RAI, is a major inhibitor of both RELA and nuclear factor kappa-B (NFKB), two central modulators of inflammation. RAI binds to the p65 subunit of NFKB, and also inhibits tumor necrosis factor-alpha-induced activation of NFKB [83]. Thus *LEPR*, *BCL11A*, *PPP1R13L* genes, members of the largest clique, were not just topologically, but also biologically strongly linked. They suggest that the major driver of the differential network  $G_D$  is an immune component coupled to the damaging effects of smoking.

## 4.6. Conclusions

In the present study, we showed that *cis* eQTLs for asthma are often co-regulated jointly by *trans* SNPs. As a tool we used MB-MDR screening methodology to detect *trans* x *cis* eQTL interactions for a pre-determined set of *cis* eQTLs. The developed protocol maintained adequate control over FWER. We identified significant functional overlap between *trans* and *cis* gene regulatory components that included immune and signaling pathways, amongst others. The network-based approaches were proven to be powerful in confirming results of the joint *trans/cis* gene set enrichment analyses. Although further work is needed to fully understand the impact of our findings, the network-based MB-MDR hybrid approach, as introduced in this work, seems to be

useful in highlighting meaningful gene-gene interactions and biological mechanisms, while integrating transcriptome and genome data.

#### 4.7. Chapter highlights

In this chapter we investigated DNA-RNA interplays via a novel hybrid approach, combining genome-wide epistasis screening with MB-MDR and network theory. The MB-MDR method was proven to be flexible enough to identify meaningful gene - gene interactions, as obtained from networks derived from statistically significant *trans* x *cis* eQTL interactions. In order to guarantee adequate FWER control of our approach, and to enhance clinical applicability of findings, we restricted attention to genetic markers with  $MAF \geq 0.20$ . The FWER of our adaptive step-wise *trans/cis* eQTL method calculated on the permuted null data at the  $p$ -value  $< 0.05$  threshold was within the acceptable range of 0.04 - 0.05 validating our epistatic findings. The 9 genes were known asthma markers according to the Asthma Genome Browser [70]. The *LEPR*, *BCL11A*, *PPP1R13L* genes are members of the largest clique of our gene-gene interaction network have strong biological links to asthma and can be considered as a new set of the disease markers. Gene-way and pathway-level analyses pointed towards previously reported asthma-relevant mechanisms, but also pointed towards novel routes for further investigation.

#### 4.8. Acknowledgements and funding

The research was funded by the Fonds de la Recherche Scientifique (FNRS) (incl. FNRS 428 F.R.F.C. project convention n° 2.4609.11). We thank Channing lab, specifically Benjamin Ravi and Scott Tillman Weiss, in providing access to CAMP data and computational facilities.

## 4.9. Appendix

**Figure S4.1:** The complete weighted directed gene network  $G$  built using the list of 1364 significant *trans/cis* eQTLs complimenting Figure 4.8. (See the online supplement)

**Figure S4.2:** The complete weighted directed gene network  $G_{NS}$  built using the list of 1552 significant *trans/cis* eQTLs complimenting Figure 4.10. (See the online supplement)

**Figure S4.3:** The complete weighted directed gene network  $G_S$  built using the list of 707 significant *trans/cis* eQTLs complimenting the differential network  $G_D$  shown in Figure 4.10 (See the online supplement)

**Table S4.1:** The complete list of significant *trans/cis* eQTLs. (See the online supplement)

**Table S4.2:** The list of 30 common genes of the *trans/cis* and *cis* eQTL genes sets. See online supplement. (See the online supplement)

**Table S4.3:** The 71 common significantly enriched pathways between *trans/cis* and *cis* eQTL genes sets. (See the online supplement)

**Table S4.4:** Total degree distribution of nodes of the *trans/cis* eQTL network  $G$  (see Figure 4.5). (See the online supplement)

**Table S4.5:** The common list of nodes between  $G_{NS}$  and  $G_S$  networks with the corresponding total degrees. (See the online supplement)

**Table S4.6:** The 3 significant epistatic *trans/cis* eQTL pairs both at Bonferroni and  $M_{\text{eff}}$  method [59] thresholds

<i>trans</i> SNP	<i>cis</i> SNP	<i>p</i> -value*	<i>trans</i> gene	<i>cis</i> gene	<i>trans</i> gene name	<i>cis</i> gene name
rs12060945	rs4711338	4.3e-06	RTCD1	ITPR3	RNA 3'-Terminal Phosphate Cyclase	Inositol 1,4,5-Trisphosphate Receptor, Type 3
rs12582824	rs12420868	2.68e-05	CLEC4D	LRRC56	C-Type Lectin Domain Family 4, Member D	Leucine Rich Repeat Containing 56
rs954639	rs1045895	2.83e-05	SEC23IP	LEPR	SEC23 Interacting Protein	Leptin Receptor

\* per *cis* gene multiple-testing adjusted via *MAXT* as implemented in MB-MDR [31]

## 4.10. References

1. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS (2004) **A statistical framework for genomic data fusion.** *Bioinformatics* 20: 2626-2635.
2. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) **Similarity network fusion for aggregating data types on a genomic scale.** *Nat Methods* 11: 333-337.
3. Glass K, Huttenhower C, Quackenbush J, Yuan GC (2013) **Passing messages between biological networks to refine predicted interactions.** *PLoS One* 8: e64832.
4. (1999) **The Childhood Asthma Management Program (CAMP): design, rationale, and methods.** *Childhood Asthma Management Program Research Group. Control Clin Trials* 20: 91-120.
5. Bentley AM, Durham SR, Robinson DS, Menz G, Storz C, et al. (1993) **Expression of endothelial and leukocyte adhesion molecules interacellular adhesion molecule-1, E-selectin, and vascular cell adhesion molecule-1 in the bronchial mucosa in steady-state and allergen-induced asthma.** *J Allergy Clin Immunol* 92: 857-868.
6. Organization TWH (2014) **The Global Asthma Report**
7. Cohn L, Elias JA, Chupp GL (2004) **Asthma: mechanisms of disease persistence and progression.** *Annu Rev Immunol* 22: 789-815.
8. Becker J, Wendland JR, Haenisch B, Nothen MM, Schumacher J (2012) **A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals.** *Eur J Hum Genet* 20: 97-101.
9. Huang Y, Wuchty S, Przytycka TM (2013) **eQTL Epistasis - Challenges and Computational Approaches.** *Front Genet* 4: 51.
10. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) **Systematic identification of trans eQTLs as putative drivers of known disease associations.** *Nat Genet* 45: 1238-1243.
11. Albert FW, Kruglyak L (2015) **The role of regulatory variation in complex traits and disease.** *Nat Rev Genet* 16: 197-212.
12. Michaelson JJ, Alberts R, Schughart K, Beyer A (2010) **Data-driven assessment of eQTL mapping methods.** *BMC Genomics* 11: 502.
13. Haley CS, Knott SA (1992) **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.** *Heredity (Edinb)* 69: 315-324.
14. Chun H, Keles S (2009) **Expression quantitative trait loci mapping with multivariate sparse partial least squares regression.** *Genetics* 182: 79-90.
15. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, et al. (2011) **Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.** *PLoS Genet* 7: e1002197.
16. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) **A genome-wide association study of global gene expression.** *Nat Genet* 39: 1202-1207.
17. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, et al. (2014) **Detection and replication of epistasis influencing transcription in humans.** *Nature* 508: 249-253.
18. Fitzpatrick DJ, Ryan CJ, Shah N, Greene D, Molony C, et al. (2015) **Genome-wide epistatic expression quantitative trait loci discovery in four human tissues reveals the importance of local chromosomal interactions governing gene expression.** *BMC Genomics* 16: 109.
19. Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, et al. (2011) **Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise.** *Ann Hum Genet* 75: 78-89.
20. Ruczinski I, Kooperberg C, LeBlanc ML (2004) **Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications.** *Journal of Multivariate Analysis* 90: 178-195.
21. Kooperberg C, Ruczinski I (2005) **Identifying interacting SNPs using Monte Carlo logic regression.** *Genetic epidemiology* 28: 157-170.

22. Boulesteix AL, Tutz G, Strimmer K (2003) **A CART-based approach to discover emerging patterns in microarray data.** *Bioinformatics* 19: 2465-2472.
23. Storey JD, Akey JM, Kruglyak L (2005) **Multiple locus linkage analysis of genomewide expression in yeast.** *PLoS Biol* 3: e267.
24. Fish A, Capra JA, Bush WS (2015) **Are Genetic Interactions Influencing Gene Expression Evidence for Biological Epistasis or Statistical Artifacts?** *bioRxiv*: 020479.
25. Carlborg O, Brockmann GA, Haley CS (2005) **Simultaneous mapping of epistatic QTL in DU6i x DBA/2 mice.** *Mamm Genome* 16: 481-494.
26. Carlborg O, Andersson L (2002) **Use of randomization testing to detect multiple epistatic QTLs.** *Genet Res* 79: 175-184.
27. Michaelson JJ, Loguercio S, Beyer A (2009) **Detection and interpretation of expression quantitative trait loci (eQTL).** *Methods* 48: 265-276.
28. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) **eQED: an efficient method for interpreting eQTL associations using protein networks.** *Mol Syst Biol* 4: 162.
29. Boulesteix AL, Strobl C, Weidinger S, Wichmann HE, Wagenpfeil S (2007) **Multiple testing for SNP-SNP interactions.** *Stat Appl Genet Mol Biol* 6: Article37.
30. Hochberg Y, Benjamini Y (1990) **More powerful procedures for multiple significance testing.** *Statistics in medicine* 9: 811-818.
31. Van Lishout F, Mahachie John JM, Gusareva ES, Urrea V, Cleynen I, et al. (2013) **An efficient algorithm to perform multiple testing in epistasis screening.** *BMC Bioinformatics* 14: 138.
32. Murphy A, Chu JH, Xu M, Carey VJ, Lazarus R, et al. (2010) **Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes.** *Hum Mol Genet* 19: 4745-4757.
33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 81: 559-575.
34. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 23: 1294-1296.
35. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 19: 185-193.
36. Du P, Kibbe WA, Lin SM (2008) **lumi: a pipeline for processing Illumina microarray.** *Bioinformatics* 24: 1547-1548.
37. Mahachie John JM, Van Lishout F, Van Steen K (2011) **Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data.** *Eur J Hum Genet* 19: 696-703.
38. Lishout FV, Gadaleta F, Moore JH, Wehenkel L, Steen KV (2015) **gammaMAXT: a fast multiple-testing correction algorithm.** *BioData Min* 8: 36.
39. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2005) **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 21: 3439-3440.
40. Durinck S, Spellman PT, Birney E, Huber W (2009) **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 4: 1184-1191.
41. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013) **Software for computing and annotating genomic ranges.** *PLoS Comput Biol* 9: e1003118.
42. Wilkinson L (2012) **Exact and approximate area-proportional circular Venn and Euler diagrams.** *IEEE Trans Vis Comput Graph* 18: 321-331.
43. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, et al. (2011) **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics* 27: 1739-1740.
44. Kost JT, McDermott MP (2002) **Combining dependent P-values.** *Statistics & Probability Letters* 60: 183-190.

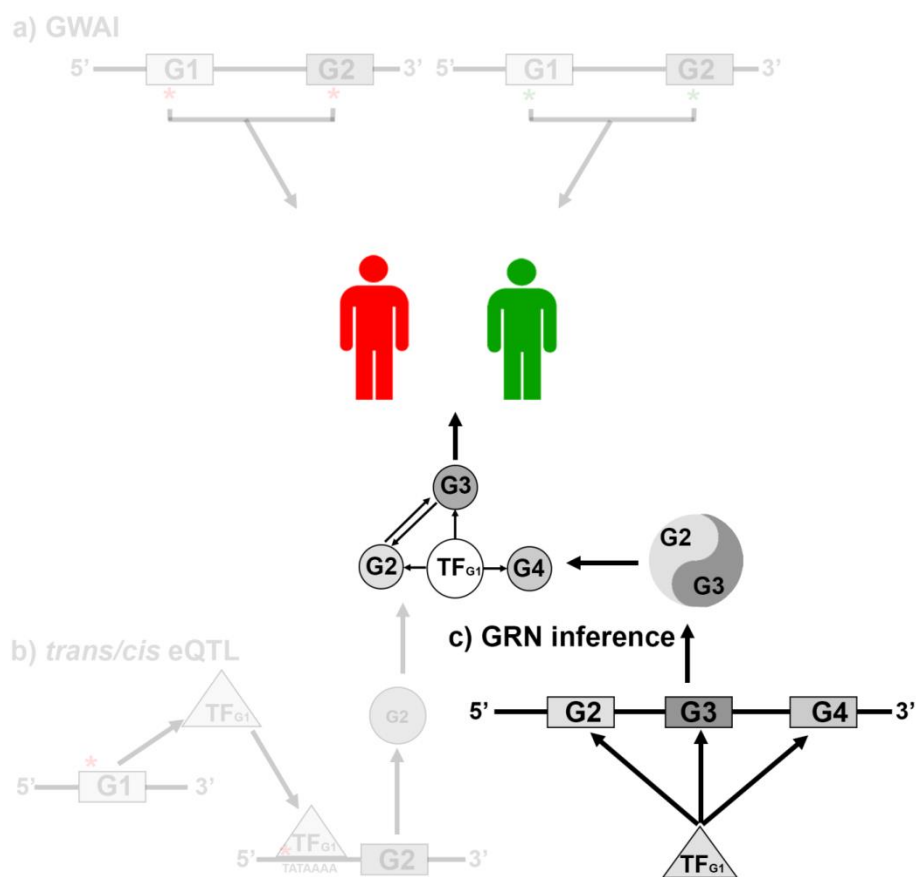
45. Csardi G, Nepusz T (2006) **The igraph software package for complex network research.** *InterJournal, Complex Systems* 1695: 1-9.
46. Mahachie John JM, Van Lishout F, Gusareva ES, Van Steen K (2013) **A robustness study of parametric and non-parametric tests in model-based multifactor dimensionality reduction for epistasis detection.** *BioData Min* 6: 9.
47. Barabasi AL, Oltvai ZN (2004) **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 5: 101-113.
48. Szymczak S, Igl BW, Ziegler A (2009) **Detecting SNP-expression associations: a comparison of mutual information and median test with standard statistical approaches.** *Stat Med* 28: 3581-3596.
49. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet* 133: 1343-1358.
50. Bessonov K, Gusareva ES, Van Steen K (2015) **A cautionary note on the impact of protocol changes for genome-wide association SNP x SNP interaction studies: an example on ankylosing spondylitis.** *Hum Genet* 134: 761-773.
51. Zhang X, Huang S, Sun W, Wang W (2012) **Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study.** *Genetics* 190: 1511-1520.
52. Lin DY (2005) **An efficient Monte Carlo approach to assessing statistical significance in genomic studies.** *Bioinformatics* 21: 781-787.
53. Li G, Shabalin AA, Rusyn I, Wright FA, Nobel AB (2013) **An empirical Bayes approach for multiple tissue eQTL analysis.** *arXiv preprint arXiv:13112948*.
54. Peterson CB, Bogomolov M, Benjamini Y, Sabatti C (2016) **Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies.** *Genetic epidemiology* 40: 45-56.
55. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O (2015) **Fast and efficient QTL mapper for thousands of molecular phenotypes.** *Bioinformatics*: btv722.
56. Sul JH, Raj T, de Jong S, de Bakker PI, Raychaudhuri S, et al. (2015) **Accurate and fast multiple-testing correction in eQTL studies.** *Am J Hum Genet* 96: 857-868.
57. Davis JR, Fresard L, Knowles DA, Pala M, Bustamante CD, et al. (2016) **An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants.** *Am J Hum Genet* 98: 216-224.
58. Mutation C, Pathway Analysis working group of the International Cancer Genome C (2015) **Pathway and network analysis of cancer genomes.** *Nat Methods* 12: 615-621.
59. Li J, Ji L (2005) **Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix.** *Heredity (Edinb)* 95: 221-227.
60. Mahachie J (2012) **Thesis: Genomic Association Screening Methodology for High-Dimensional and Complex Data Structures:** University of Liege.
61. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G (2013) **Enhancers: five essential questions.** *Nat Rev Genet* 14: 288-295.
62. Nica AC, Dermitzakis ET (2013) **Expression quantitative trait loci: present and future.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 368: 20120362.
63. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, et al. (2012) **Patterns of cis regulatory variation in diverse human populations.** *PLoS Genet* 8: e1002639.
64. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, et al. (2012) **Mapping cis- and trans-regulatory effects across multiple tissues in twins.** *Nat Genet* 44: 1084-1089.
65. Grange L (2014) **Thesis: epistasis in genetic susceptibility to infectious diseases: comparison and development of methods application to severe dengue in Asia:** Paris 7.
66. Laura Grange KB, Tom Cattaert, Iryna Nikolayeva, Jestinah M Mahachie John, Benno Schwikowski, Jean-François Bureau, Anavaj Sakuntabhai, Kristel Van Steen (2016) **Finding the tree for the**

- forest: which epistasis analysis method to choose.** Department of Genomes and Genetics, Institut Pasteur, Functional Genetics of Infectious Diseases, Systems and Modeling Unit – BIO3, Quartier Polytech 1, University of Liège, Liège, Belgium, Systems Biology and Chemical Biology, GIGA-R, University of Liège, Liège, Belgium. pp. 21.
67. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) **BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.** *The American Journal of Human Genetics* 87: 325-340.
  68. Westfall PH (1993) **Resampling-based multiple testing: Examples and methods for p-value adjustment:** John Wiley & Sons.
  69. Melen E, Bruce S, Doekes G, Kabesch M, Laitinen T, et al. (2005) **Haplotypes of G protein-coupled receptor 154 are associated with childhood allergy and asthma.** *Am J Respir Crit Care Med* 171: 1089-1095.
  70. Mersha TB (2015) **Mapping asthma-associated variants in admixed populations.** *Frontiers in genetics* 6.
  71. Ierodiakonou D, Postma DS, Koppelman GH, Boezen HM, Gerritsen J, et al. (2011) **E-cadherin gene polymorphisms in asthma patients using inhaled corticosteroids.** *Eur Respir J* 38: 1044-1052.
  72. Liang Q, Guo L, Gogate S, Karim Z, Hanifi A, et al. (2010) **IL-2 and IL-4 stimulate MEK1 expression and contribute to T cell resistance against suppression by TGF-beta and IL-10 in asthma.** *J Immunol* 185: 5704-5713.
  73. Mammes O, Aubert R, Betoulle D, Pean F, Herbeth B, et al. (2001) **LEPR gene polymorphisms: associations with overweight, fat mass and response to diet in women.** *Eur J Clin Invest* 31: 398-404.
  74. Shore SA, Schwartzman IN, Mellema MS, Flynt L, Imrich A, et al. (2005) **Effect of leptin on allergic airway responses in mice.** *J Allergy Clin Immunol* 115: 103-109.
  75. Lugogo NL, Hollingsworth JW, Howell DL, Que LG, Francisco D, et al. (2012) **Alveolar macrophages from overweight/obese subjects with asthma demonstrate a proinflammatory phenotype.** *Am J Respir Crit Care Med* 186: 404-411.
  76. Bruno A, Pace E, Chanez P, Gras D, Vachier I, et al. (2009) **Leptin and leptin receptor expression in asthma.** *J Allergy Clin Immunol* 124: 230-237, 237 e231-234.
  77. Wen X, Liu H, Xiao G, Liu X (2011) **Downregulation of the transcription factor KLF4 is required for the lineage commitment of T cells.** *Cell Res* 21: 1701-1710.
  78. Weniger MA, Pulford K, Gesk S, Ehrlich S, Banham AH, et al. (2006) **Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11A(XL) protein are frequent in primary mediastinal B-cell lymphoma.** *Leukemia* 20: 1880-1882.
  79. Lee B-s, Dekker JD, Lee B-k, Iyer VR, Sleckman BP, et al. (2013) **The BCL11A transcription factor directly activates RAG gene expression and V (D) J recombination.** *Molecular and cellular biology* 33: 1768-1781.
  80. Ippolito GC, Dekker JD, Wang YH, Lee BK, Shaffer AL, 3rd, et al. (2014) **Dendritic cell fate is determined by BCL11A.** *Proc Natl Acad Sci U S A* 111: E998-1006.
  81. Ayllon V, Cayla X, Garcia A, Roncal F, Fernandez R, et al. (2001) **Bcl-2 targets protein phosphatase 1 alpha to Bad.** *J Immunol* 166: 7345-7352.
  82. Song M-A, Lim U, Ernst T, Tiirikainen M, Wilkens LR, et al. (2013) **Genome-wide blood leukocyte DNA methylation in relation to visceral, subcutaneous, and hepatic adiposity in postmenopausal women.** *Cancer Research* 73: 4253-4253.
  83. Yang JP, Hori M, Sanda T, Okamoto T (1999) **Identification of a novel inhibitor of nuclear factor-kappaB, RelA-associated inhibitor.** *J Biol Chem* 274: 15662-15670.



## Chapter 5: Gene expression networks

### Practical aspects of gene regulatory network inference via conditional inference forests from expression data



Related publication:

Bessonov K, Van Steen K (2016) **Practical aspects of gene regulatory inference via conditional inference forests from expression data.** (*Genetic Epidemiology –accepted for publication*)



## 5. Practical aspects of gene regulatory network inference via conditional inference forests from expression data

### 5.1. Chapter summary

In previous chapters, interactions always involved a genetic (SNP-based) layer of information in relation to a complex disease trait, in this chapter we restrict attention to transcriptome layers of information and the inference of gene-expression networks in a well-defined population (here, a group of individuals exhibiting the same phenotype). Only one transcriptome per subject is considered, either derived from synthetic or real-life expression microarray data.

In particular, in this work we propose a novel framework to create GRNs, based on Conditional Inference Forests (*CIFs*) as proposed by Strobl *et al.* [1] and gene expression data. Our framework consists of using ensembles of Conditional Inference Trees (CITs) prior to network construction. We show on synthetic microarray data from the DREAM challenges that the original implementation of *CIFs* with conditional permutation scheme leads to improved performance compared to Breiman's implementation of Random Forest (*RF*). Although more work is needed to improve on speed, especially when fully exploiting the advantages of conditional inference trees in the context of heterogeneous and correlated data via a conditional permutation scheme, we show that the *CIF* methodology can be flexibly inserted in the GRN inference framework to mine for biologically meaningful interactions. In contrast, networks derived from well-tuned *CIFs*, obtained by simply averaging  $p$ -values over tree ensembles ( $CIF_{mean}$ ) are a particularly attractive less computationally intensive alternative: adequate performance is combined with computational efficiency. Moreover, thresholds for variable selection are based on significance levels for  $p$ -values and hence do not need to be tuned. The latter is important when working with real-life biological data, for which the truth is largely unknown. Finally, the *CIF*'s theoretical advantage in the presence of multiple omics data, measured on different scales, makes it a promising tool for integrative omics data analyses.

**Problem:** GRN inference using real-life expression data is not novel, but to our knowledge available algorithms and methodologies fail to provide adequate performance on real-life data with genome-wide scales. This has to do with the complexities linked to expression data, including complex correlation patterns of the expression profiles, systematic measurement biases due to different binding efficiencies per microarray probe, variable amounts of RNA per sample, etc. [2]. In addition, most of the classical statistical analysis methods used to assess differential gene expression (t-test, ANOVA, F-test) ignore dependencies of genes (i.e. correlation structure) which are omnipresent in real-life gene-expression data [3]. Therefore, it is essential to have a tool available that can reliably derive gene - gene interactions from gene expression data. In this chapter, we employ and compare the GRN inference potential of non-linear tree-based algorithms such as *RFs* and *CIFs*. We investigate whether *CIFs* can provide suitable GRN inference performance as compared to *RF* in small-scale and genome-wide scenarios. In addition, we assess performance impact of main tuning parameters such as the number of variables to pick from at a given tree node (*mtry*), and of different multiple-testing correction options (*Bonferroni*, *Monte-Carlo*), as well as performance sacrifices after omission of the permutation step during estimation of variable importance measure (*VIM*), etc. In addition, we explore the benefits of the conditional permutation scheme that makes *CIFs* so unique. Finally, testing theory to practice, we build a GRN from type 1 diabetes (T1D) expression data and provide biologically plausible gene interaction hypotheses.

**Results:** *CIF* based methods that utilize both test statistic and *p*-value as  $VIM_{node}$  show similar and at times better than *RF* GRN inference performance. For example in DREAM4 data [4] *CIF<sub>cond</sub>*, *CIF* and *RF* reach performance DREAM scores of 34.24, 33.92 and 33.50, respectively. Due to significant computational runtime requirements the classical *CIF* methods [1,5,6] cannot be applied to GRN contexts with >100 genes requiring the introduction of alternative heuristic methods – *CIF<sub>mean</sub>*. This method is based on averaging of node-specific *p*-values of a conditional inference tree. The highest performance measured via AUROC and AUPR across all datasets was for *CIF<sub>mean</sub>* test-statistic (Uncorrected) and *CIF<sub>mean</sub>* *p*-value (Monte-Carlo) at *mtry* values equal to 1/3 of input variables. We discover and confirm highly statistically and biologically relevant interactions between *IL2RA* and *FOXP1* members of IL-2 signaling pathway linked to type 1 diabetes.

**Keywords:** gene regulatory networks (GRN), conditional inference forests (*CIFs*), DREAM datasets, diabetes, random forest (*RF*), microarrays, expression data, transcriptome.

## 5.2. Introduction

Real-life biological systems display interactions and regulation schemes that are part of complex pathways or networks. Understanding these networks is important to unravel gene regulatory mechanisms or the genetic basis of complex disease traits. The availability of genome-wide transcriptome data offers opportunities and challenges for data analysts to extract gene regulation information directly from gene expression profiles: genes regulate each other's expression and activity.

One of the challenges when dealing with data derived from high-throughput technologies (i.e. omics data) involves the curse of dimensionality. This refers to the fact that number of variables  $p$  is usually much larger than the number of samples  $n$  for these data and, hence, model parameter estimation becomes unstable. Ignoring the  $p \gg n$  issue and adhering to classical statistics, is bound to generate singularities in matrix algebra (e.g., singular matrices) [7]. The curse of dimensionality particularly applies to transcriptome data derived via *RNA-seq*, but also holds true for microarray-based data that typically considers between 10,000 and 57,000 transcripts, depending on the platform and organism [8]. One way to circumvent this problem is to reduce the number of variables. This can be done by using prior biological knowledge leading to biologically motivated constraints, or via mathematical/statistical variable selection algorithms. Alternatively, novel representations of the data are looked for, such as principal components in a lower-dimensional linear space [9] or kernels for non-linear data dimensionality reduction [10]. Graph structures are easy to interpret and naturally represent biological networks [11]. These networks may refer to genes and gene products or to networks between macro- and micro-molecules, possibly integrating different data sources with different interaction profiles in a single consensus network [12]. Importantly, biological networks show a scale-free as discussed in [13]. This implies that only a small number of nodes in the network are highly connected and that the majority of nodes are connected to only a few neighboring nodes. Usually, connected nodes in such networks are said to be “interacting”. However, this does not necessarily mean that the nodes (or the compounds they represent) are physically interacting. Note that several so-called physical interaction networks may miss true interactions as well and often contain non-functional interactions [14]. Gene regulatory

networks (GRNs) represent directed functional linkages existing between genes and regulatory elements most frequently associated with transcription factors [15].

In this work, the envisaged biological networks are functional GRNs for which “interactions” depict either direct or indirect regulatory relationships [16]. The framework we develop relies on hybrid tree-based variable selection and GRN inference. One of the advantages of trees and ensembles of trees [17] is their ability to effectively and rapidly dissect complex data spaces such as those generated by gene expression microarrays. Trees and random forest methodology belong to the class of recursive partitioning methods, that aim to recursively partition the space spanned by all input variables into partitions of observations with similar responses. The final partitions may be characterized by highly complex interactive patterns between input variables, although care has to be taken when interpreting interactions in the context of random forests [18]. For a general overview on classification and regression trees, we refer to [19].

Single tree-based models can over-fit data at hand and, hence, to underestimate classification errors. Several measures can be taken to overcome these issues, including the building of unpruned trees on multiple bootstrap samples as implemented in Breiman Random Forests (*RF*) [20] and the separation of variable selection and node splitting steps [21]. Such a separation is implemented in Conditional Inference Trees (*CIT*) and Conditional Inference Forests (*CIFs*) [1,21]. At the heart lies an unbiased tree algorithm [21,22] that do not artificially favor splits in variables with many categories or continuous variables [23]. *CIFs* present several advantages over classical *RFs* including separation of node selection and splitting steps to overcome tree-based variable selection bias [5], resampling with replacement to handle ensemble variable selection bias introduced by bootstrap sampling [21], a conditional permutation scheme to deal with correlated input features [1], and the possibility of natural threshold selection for variable importance measures (*VIMs*), as we will show later. Hence, a *CIF*-based methodology theoretically encompasses categorical and continuous input variables that are possibly inter-related and measured on different scales, hereby paving the way for combined analysis of multiple data sources. For these reasons, and having integrative analyses of heterogeneous and interconnected omics data in mind, we chose *CITs* and *CIFs* as the basis of our novel network construction and inference methodology, despite the fact that random forests rather than *CITs* or *CIFs* are widely applied in bioinformatics contexts [24]. In

the belief that computational efficiency can be reached by optimizing the program code and *CIT/CIF* underlying algorithms, we focus on investigating the impact of parameter choices in *CIT* and *CIF* (e.g., related to multiple testing correction and the number of randomly selected variables at each tree node) on the performance of proposed gene regulatory network construction methods in synthetic and real-life data settings.

The *CIF<sub>mean</sub>* source code and run scripts developed in the context of this manuscript are freely available through via <https://bitbucket.org/kbessonov/cifmean> and [www.statgen.ulg.ac.be](http://www.statgen.ulg.ac.be).

## 5.3. Methods

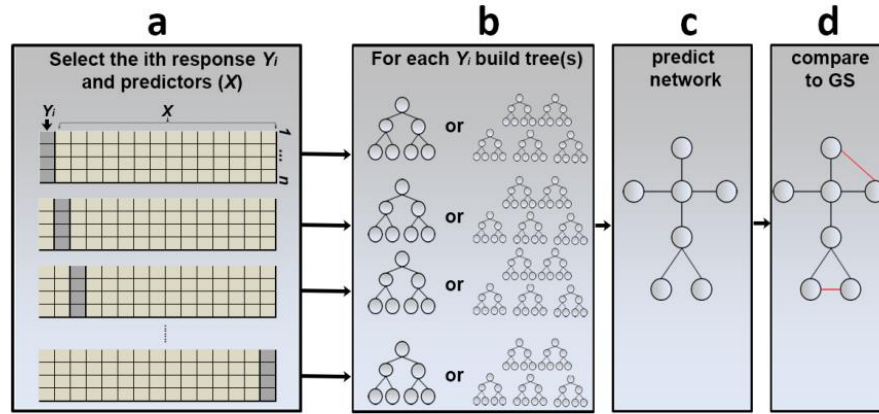
### 5.3.1. Data sources

We obtained publicly available gene expression data from the DREAM 2, 4 and 5 challenges [4,25,26] and the GEO public repository (GEO #: GSE43488).

In particular, we used gene expression data on 3456 *E.coli* genes from DREAM2, containing 320 transcription factors (TF), for 300 subjects [27]. The 320 TFs were considered as input variables to our proposed strategies (Figure 5.1). As the gold standard (GS) network we took evidence from *RegulonDB* [28] of experimentally verified regulator (TF) - target gene (TG) relationships.

The DREAM4 In Silico Network Challenge data only contained synthetic microarray expression data derived from 5 GS networks, each with 100 nodes [4,25,29]. Thus, each dataset contained 100 genes collected on 100 samples. Since no list of potential regulators existed, all 100 genes were considered as input variables.

The DREAM5 Network Inference Challenge data consisted of three GS networks 1-3, with respectively 1643, 4511 and 5950 genes. The GS network 1 data contained synthetic (simulated) gene expression data represented by 1643 genes and 195 regulators for 805 samples. The GS network 2 real-life *E.coli* expression data was characterized by the 4511 genes containing 334 TFs collected on 805 subjects.



**Figure 5.1:** Gene regulatory network framework based on *CIT/CIF*, adapted from [17,30]

- a) Given gene expression data for a number of subjects or individuals, consider iteratively each gene expression as output (response) and remaining gene expression as input (predictors).
- b) Construct a conditional inference tree (*CIT*) or conditional inference forest (*CIF*) per input/output.
- c) Per node, aggregate over all available tree(s) to obtain a variable importance measure ( $VIM_{global}$ ). Construct a non-symmetric adjacency matrix and, hence, a directed network.
- d) Compare the obtained predicted network to a gold standard, whenever such a standard is available or use performance metrics such as area under the ROC curve (AUROC) or area under the precision-recall curve (AUPR).

The DREAM5 GS network 3 also involved real-life data on an organism, this time *S.cerevisiae*. The corresponding gene expression data included 5950 genes containing 333 regulators and was collected on 536 samples. For each scenario, the entire set of regulators was used as starting set for variable selection.

As a case study, we took human microarray expression data from a type 1 diabetes (T1D) study in children [31], obtained via the public GEO database (GEO #: GSE43488). As the gold standard we considered the verified set of transcription factor–target gene sets from [27] that used a variety of sources, including the Transcriptional Regulatory Element Database (TRED) [32], Pazar [33], PubMed, and the Transcription Regulatory Regions Database (TRRD) [34], among others. The resulting unique list of gene-gene pairs was composed of 1617 genes (245 TFs and 1372 target genes). These 1617 genes, evaluated on 121 samples, served as input to the considered analytic tools. A summary of the available data is given in Table 5.1.

**Table 5.1:** Data characteristics

Data	GS available	Real-life	Nr of genes	Nr of TFs	Nr of Samples
DREAM2 ( <i>E.coli</i> )	Y	Y	3456	320	300
DREAM4 network 1	Y	N	100	100**	100
DREAM4 network 2	Y	N	100	100**	100
DREAM4 network 3	Y	N	100	100**	100
DREAM4 network 4	Y	N	100	100**	100
DREAM4 network 5	Y	N	100	100**	100
DREAM5 network 1	Y	N	1643	195	805
DREAM5 network 2 ( <i>E.coli</i> )	Y	Y	4511	334	805
DREAM5 network 3 ( <i>S.cerevisiae</i> )	Y	Y	5950	333	536
Case study: T1D ( <i>Human</i> )	N*	Y	1617	245	121

\*tentative gold standard was built based on the prior knowledge of the transcription factor–target gene interactions extracted from public databases (see Methods)

\*\*all 100 genes were used as potential TFs; no list of potential regulators was specified

### 5.3.2. CIT/CIF-based network inference methodologies

A schematic representation of our proposed GRN framework is given in Figure 5.1 (adopted from [17,30]). In particular, for a given omics data set, with molecular information that can be mapped to a gene, for instance transcriptome data, and assuming a one-to-one mapping of transcripts to genes, each transcript (gene) is subsequently taken as output (response) and the remaining transcripts (genes) are taken as input (predictor variables). For each response, a *CIF/CIT* is constructed and for each response-predictor gene pair a variable importance measure (*VIM*) is calculated (see below). These measures per gene are either based on a single *CIT* or are aggregated over several *CITs* in gene-based *CIFs*, depending on the view taken to construct a network from trees. In general, a (statistically) “significant” *VIM* for gene *X* in predicting gene *Y* will lead to a connection between *X* and *Y* in the network. Because of the direction of prediction, the connection is presented as a directed edge, naturally giving rise to a directed network (i.e. GRN). The so-called predicted network is compared to a gold standard (when available), using network prediction performance criteria as suggested by [26,29]:

- 1) the area under the receiver operating characteristic curve ( $AUROC$ ),
- 2) the area under the precision-recall curve ( $AUPR$ ), and
- 3) the *DREAM* challenge specific score.

The ROC curve plots the sensitivity (i.e. true positive rate) versus 1 minus specificity (i.e. 1 minus the true negative rate) and is well-known in statistics. Precision-Recall curves or PR curves are often used in Information Retrieval and offer an alternative to ROC curves for skewed class distributions. An algorithm may be a good performer based on ROC but not based on PR. Whereas recall is defined as the true positive rate, precision is defined as the fraction of examples classified as positive that are truly positive. When the number of unconnected nodes exceeds the number of connected nodes in the GS networks, as is the case with GRNs, more information about comparative performance of methods can be retrieved from precision-recall curves [35]. For more details about ROC-PR comparisons, we refer to [35]. The overall score summarizes performance over several network scenarios and is defined as in [26] as the mean of the (-  $\log_{10}$ -transformed) network specific  $p$ -values  $p_{PR}$  and  $p_{ROC}$ . The PR and ROC  $p$ -values are derived from the original  $AUPR$  and  $AUROC$  values by comparison of obtained areas with those obtained from a simulated null distribution based on 25,000 random networks [26].

In what follows, we briefly describe the network inference schemes considered in this work. Each of these schemes involved particular choices of *VIMs* and, hence, different gene-gene network (GRN) building strategies:

**CIT:** Here, the global null hypothesis of independence between any of the predictors and the response under consideration is tested by means of the conditional distribution of linear statistics in the permutation test framework of [36]. When this hypothesis cannot be rejected, the procedure stops. Otherwise, the predictor with the strongest association to the response is selected. We define the node's variable importance measure as its measure of association with the response (i.e. the  $p$ -value of the corresponding association test) and denote it as  $VIM_{node}$ . When a variable appears multiple times in the tree, the  $VIM_{node}$  value corresponding to the largest node for that variable (i.e. with the largest sample size) is taken. Next, the most optimal split for that node is sought (i.e. the

split that maximizes a split statistic). The split statistic is based on standardized linear statistics as before. For more details, we refer to [5].

**CIF and  $CIF_{cond}$ :** In this work, both GRN inference schemes build an ensemble of conditional inference trees via the R *party* package version 1.0-11 [1,5,6]. The *cforest\_control()* function therein defines parameters that control the tree building. Unless stated otherwise, we passed the following parameters described in [6] to *cforest\_control()*: *teststat="quad"*, *testtype="Univariate"*, *fraction=0.632*, *replace=F*, *mincriterion=0.95*, *minsplit=20*, *ntree=1000*, *mtry=k/3*. Note that *teststat="quad"*, *testtype="Univariate"*, and *replace=F* correspond to the recommendations given in [21], so as to construct unbiased random forest. The *mtry* parameter (i.e. the number of variables randomly selected at each node) was set to  $k/3$ , with  $k$  representing the total number of possible predictors in the data as recommended by [37]. While the *cforest()* function creates ensembles of trees from a training section of the input data, the function *varimp()* uses the out-of-bag (*OBB*) samples to calculate the importance of each predictor variable with respect to target response. In particular, for each gene predictor / gene response pair, the *varimp()* function outputs the mean decrease in accuracy (*%IncMSE*), indicating how much the mean square error (MSE) increases after permutation of the *OBB* samples averaged over all trees of the forest. Thus, large values of *%IncMSE* are suggestive of a gene pair's importance. Because for forests, a node's variable importance is aggregated over several trees, we denote it by  $VIM_{global}$ . In practice, its calculation was made by the *varimp()* function with parameters *nperm=100* and *OBB=T*. Hence, we used a total of 100 data permutations and OOB samples in the testing phase. In the case of  $CIF_{cond}$  the *conditional* parameter in the *varimp()* function was set to *true* (i.e. variable importance was assessed via the conditional importance measure of [23]), while in *CIF* it was set to *false*.

**$CIF_{mean}$ :** In contrast to *CIF* and  $CIF_{cond}$ , we passed the following tree growth parameters to the function *ctree\_control()*: *teststat="quad"*, *testtype="Univariate"*, *fraction=0.632*, *replace=F*, *mincriterion=0.95*, *minsplit=20*, *ntree=1000*, *mtry=k/3*. Because in  $CIF_{mean}$  variable importance is assessed within a statistical framework, based on formal testing and *p*-values, the obtained *p*-values were compared to a significance level of 0.05 (i.e. *mincriterion=0.95*). In addition, a minimum number of 20 individuals were required in a node before it was considered for node splitting (*minsplit=20*). A node's variable importance  $VIM_{global}$  was aggregated over several trees

according to the formula in the Eq. 5.1, with  $n(X_j)$  - the number of trees that contain the variable  $X_j$  as a node and  $p_{X_j}^t$  - the  $p$ -value related to the association test between predictor gene  $j$  and response gene  $i$  in tree  $t$  of the ensemble ( $VIM_{node}$ ). As before, when gene  $j$  ( $X_j$ ) occurs twice in the same tree, only the  $p$ -value corresponding to the largest sample node is considered. We use  $CIF_{mean}$   $p$ -value to refer to network inference strategies in which  $VIM_{global}$  is calculated using  $p$ -values aggregated via the Eq. 5.1. The other  $VIM_{node}$  aggregation schemes including the Fisher's combined, the 95<sup>th</sup> quantile-based, the weighted mean were not further tested as they provided a lower performance in small-scale DREAM4 data. In case  $testtype="Teststatistic"$  in the `ctree_control()` function above, not  $p$ -values but raw test-statistics are used to aggregate over trees. We refer to this strategy as  $CIF_{mean}$  *test-statistic*.

$$VIM_{global} = a_{ij} = \frac{\sum_t^T p_{X_j}^t}{n(X_j)} \quad \text{Eq. 5.1}$$

**Breiman RF:** We implemented classic random forest (building 1000 trees) with the *randomForest* library (version 4.6-7) in R [20,38] and the default options with  $mtry=k/3$ . Similar to the *CIF* and *CIF<sub>cond</sub>* methodologies described above,  $VIM_{global}$  importance measures were permutation based and reflect the mean decrease in accuracy ( $\%IncMSE$ ) before and after permutation of *OOB* samples. The *OOB* samples were derived based on sampling with replacement (bootstrapping) equivalent to the  $replace=T$  in the *CIF*. They were computed via the function `importance(...)[, "%IncMSE"]`.

**The conditional inference framework and multiple-testing:** Previously, we indicated that we based the stopping criterion during node selection in *CIT* or *CIF* on univariate (multiple testing uncorrected)  $p$ -values as invoked by  $testtype="Univariate"$  in `ctree_control()`. However, it is also possible to use a stopping rule based on test statistics rather than  $p$ -values. In comparison to the second, the first does not make assumptions about the nature of large-sample distributions. Currently, in the software, it is only possible to explicitly account for multiplicity in the node selection, when using a stopping rule based on  $p$ -values ( $testtype="Univariate"$ ), either by relying on Bonferroni ( $testtype="Bonferroni"$ ) or Monte Carlo ( $testtype="MonteCarlo"$ ) strategies. In practice, with Bonferroni correction, a node's variable importance measure  $VIM_{node}$  is calculated

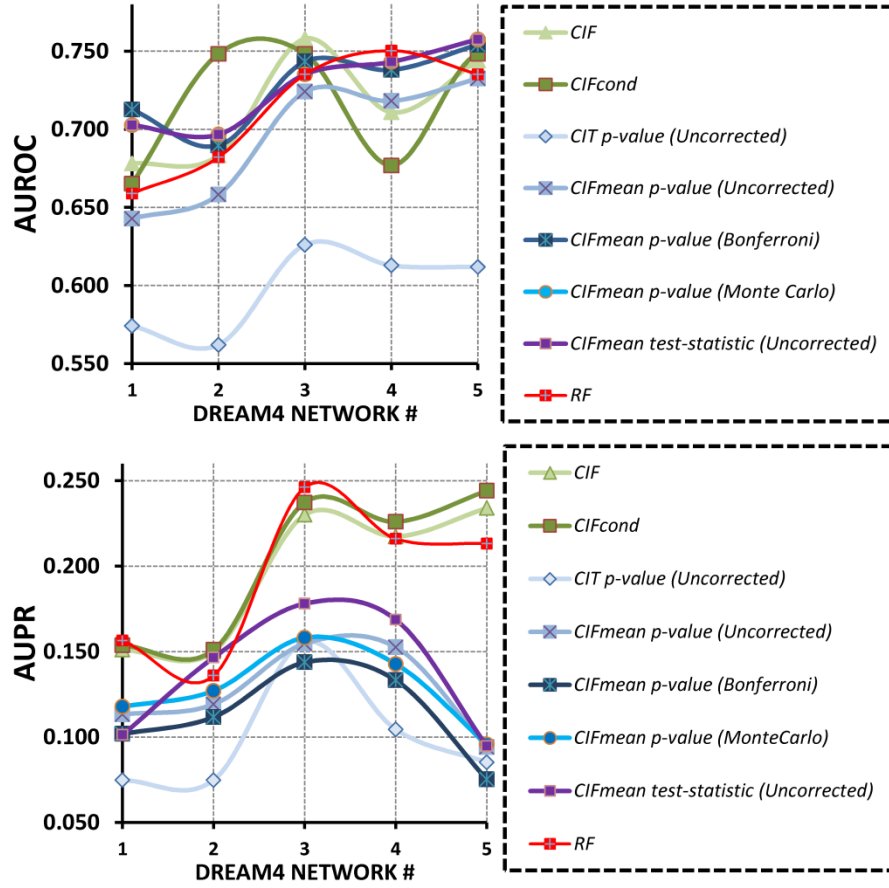
using the formula  $1-(1-p_{raw})^k$  where  $k$  is the total number of input predictor variables minus 1 [6]. The Monte Carlo multiple-testing corrected  $p$ -values are attached to a node ( $VIM_{node}$ ) calculated based on permutations and *min-p* procedure [6].

In the next section, we report results of extensive simulation studies, using the aforementioned network inference schemes and assess their robustness to altered parameter choices. In addition, we explore their utility on real-life data applications and formulate recommendations of our proposed GRN framework in data integrative contexts.

## 5.4. Results

### 5.4.1. Evaluation of *CIT/CIF*-based GRN inference with DREAM4 data

Based on the DREAM overall score criterion (see Methods section), the best performers were *CIF<sub>cond</sub>* (34.24) and *CIF* (33.92), followed by *RF* (33.50) and *CIF<sub>mean</sub>* based on aggregation of test-statistic (27.39) rather than  $p$ -values (23.75, 23.61, 23.23) (Figure S5.1). Amongst the *CIF<sub>mean</sub>* methodologies based on  $p$ -value aggregation, the best performers were GRN methodologies that utilized multiple testing (MT) correction. The Monte Carlo based MT correction was the most effective (23.75), closely followed by Bonferroni (23.61). The prediction performance of GRNs derived from a single tree (*CIT*  $p$ -value (Uncorrected)), compared to *CIF<sub>mean</sub>*  $p$ -value (Monte Carlo) was 1.8x lower: 13.17 compared to 23.75. The *AUROC* and *AUPR* in the 5 networks separately showed quite diverse performance trends amongst the considered GRN inference methods, as can be observed from Figure 5.2. The strong performance of *CIF<sub>cond</sub>* is confirmed by both *AUROC* and *AUPR*. However, it is quite computational intensive strategy (Table 5.2). The computations of methods detailed in Table 5.2 were run on a single core of an Intel L5420 processor clocked at 2.50 Ghz. For 100 genes and 100 subjects, Breiman's *RF* implementation was the fastest method, closely followed by *CIF<sub>mean</sub>*  $p$ -value (Bonferroni). For this reason, and because it is an easy-to-implement strategy giving rise to a statistically grounded threshold for variable importance, we will focus on *CIF<sub>mean</sub>* and will investigate how we can further optimize its performance.



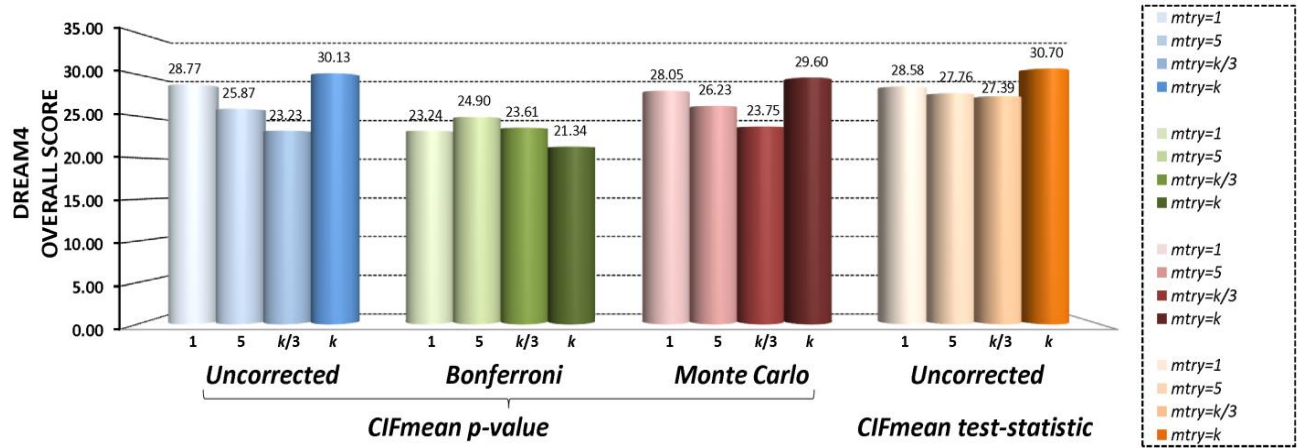
**Figure 5.2:** DREAM4 performance results –  $mtry=k/3$ . *AUROC* and *AUPR* expressed performance of considered GRN inference methodologies for each of the 5 DREAM4 networks included in the study and described in the methods Section 5.3.2. Table S5.1 complements this figure with specific *AUROC* and *AUPR* values.

The parameter  $mtry$  can have a large impact on GRN inference performance, as can be seen from Figure 5.3 for  $CIF_{mean}$ . The highest DREAM4 overall scores were obtained for  $mtry=k$ , hence using all possible input predictors, with the exception of  $CIF_{mean}$   $p$ -value (Bonferroni). For the latter approach,  $mtry=k/3$  seemed to be a reasonable choice.

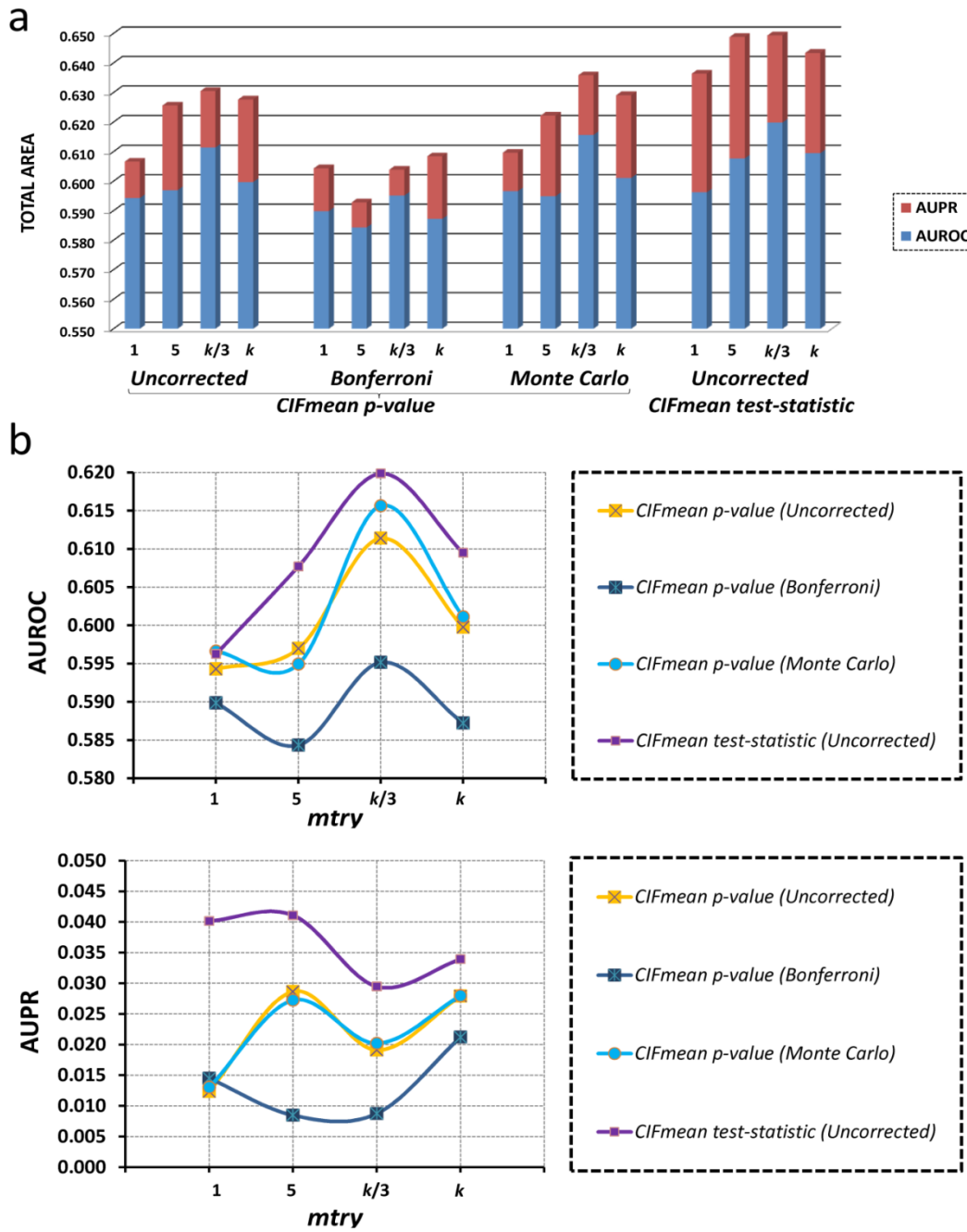
**Table 5.2:** Runtime estimates of the family of CIF methods on a single CPU

Method	Min*	h*
<i>CIT p-value (Uncorrected)</i>	0.30	0.005
<i>CIF</i>	41,600 (416 x 100)	693
<i>CIF<sub>cond</sub></i>	120,000 (1,200 x 100)	2,000
<i>CIF<sub>mean</sub> p-value (Uncorrected)</i>	12.35	0.205
<i>CIF<sub>mean</sub> p-value (Bonferroni)</i>	3.8	0.063
<i>CIF<sub>mean</sub> p-value (Monte Carlo)</i>	1,288	21.5
<i>CIF<sub>mean</sub> test-statistic (Uncorrected)</i>	14.6	0.24
<i>RF</i>	0.79	0.013

\*The input consisted of 100 genes and 100 samples (i.e. DREAM4 data). The estimated times assume serial runs without any parallelization (single thread). For settings description associated to each method, please refer to the Methods section.



**Figure 5.3:** DREAM4 performance results – variable *mtry*. The performance of the *CIF<sub>mean</sub>* methods at various *mtry* values assessed via the DREAM4 overall score. Overall scores are averages over 5 networks.



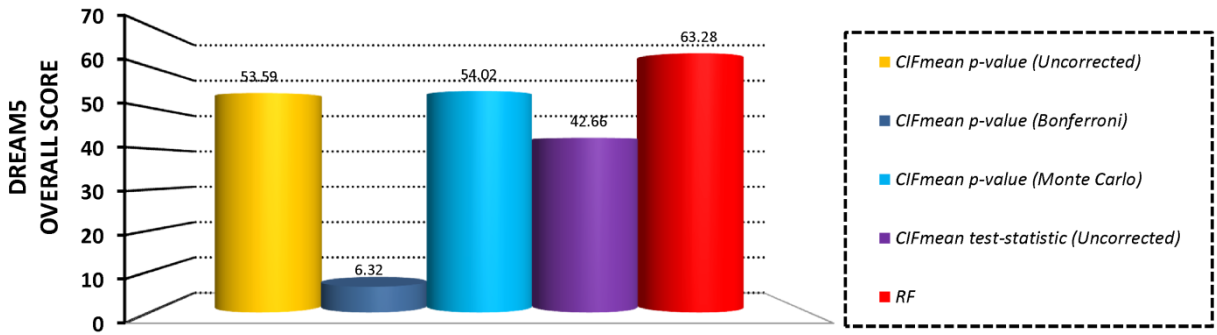
### 5.4.2. Evaluation of $CIF_{mean}$ -based GRN inference with DREAM2 data

The top performer based on  $AUROC$  and  $AUPR$  was  $RF$  followed by  $CIF_{mean}$  *test-statistic* (*Uncorrected*) and  $CIF_{mean}$  *p-value* (*Monte Carlo*) (supplementary Figure S5.2). The Monte Carlo multiple-testing correction provided the best performance amongst the  $CIF_{mean}$  *p-value* methods (supplementary Figure S5.2). Note that since only a single data scenario was available for DREAM2, it was not possible to compute a DREAM global score for each method but instead, we considered the sum of  $AUROC$  and  $AUPR$ .

Contrary to DREAM4 results, the optimal  $mtry$  parameter based on  $AUPR$  and  $AUROC$  across all  $CIF_{mean}$  methods, with the exception of Bonferroni, was at default value of  $k/3$ . In case of  $CIF_{mean}$  *p-value* (*Bonferroni*), the highest performance was reached at  $mtry$  maximal value of  $k$  (Figure 5.4).

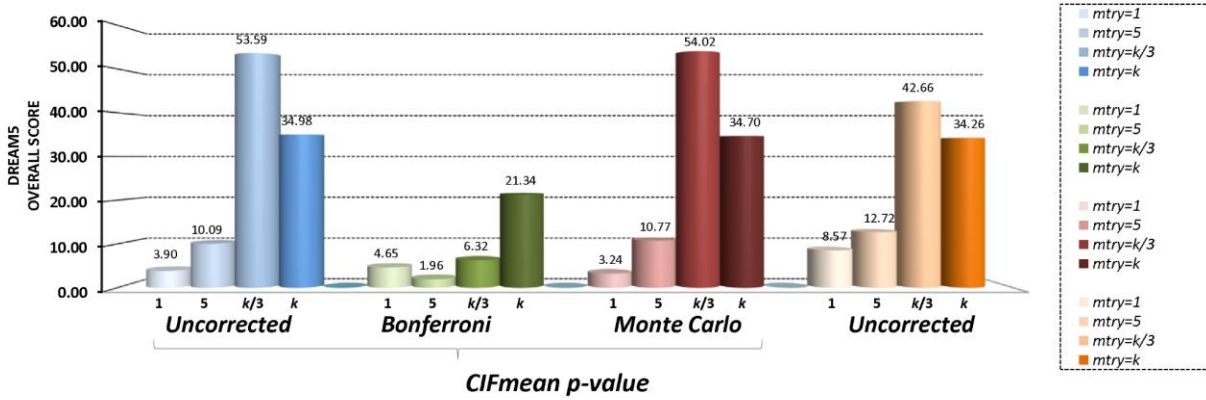
### 5.4.3. Application of $CIFs$ to DREAM5 data

Using the DREAM5 overall score developed in [26], the best performers were  $RF$  (63.28), followed by  $CIF_{mean}$  *p-value* (*Monte Carlo*) (54.02) and  $CIF_{mean}$  *p-value* (*Uncorrected*) (53.59). Very low performance was demonstrated by  $CIF_{mean}$  *p-value* (*Bonferroni*) showing  $\sim 8.5x$  performance drop compared to  $CIF_{mean}$  *p-value* (*Monte Carlo*): 6.32 compared to 54.02 (Figure S5.3 and Figure 5.5).



**Figure 5.5:** DREAM5 performance results -  $mtry=k/3$ . The GRN inference performance levels across  $CIF_{mean}$  methodologies. Performance is quantified via the DREAM5 overall score as defined in for instance [26]. Table S5.3 complements this figure with specific  $AUROC$  and  $AUPR$  values.

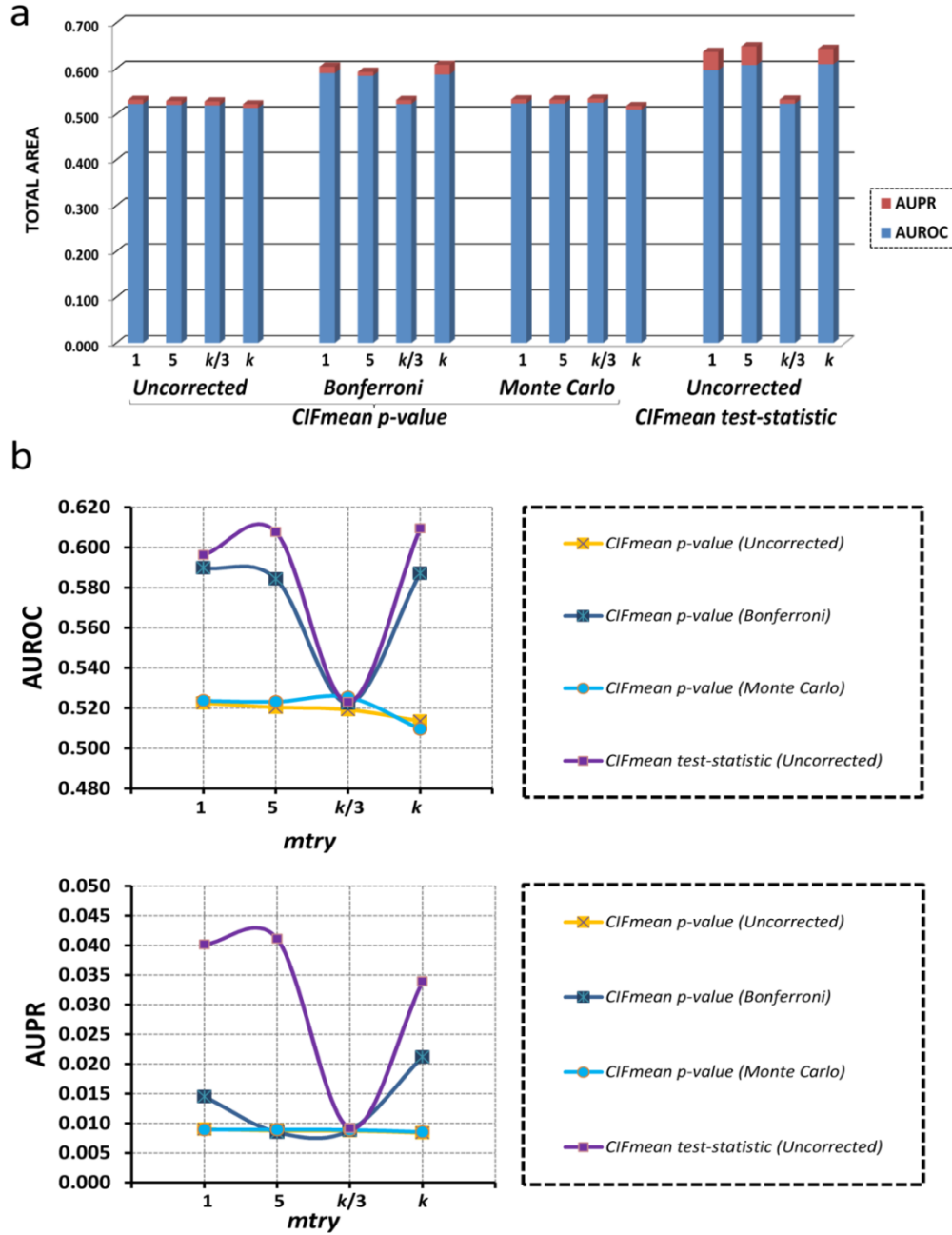
The default  $mtry$  value of  $k/3$  provided the optimal performance for  $CIF_{mean}$  uncorrected for multiple testing and  $CIF_{mean}$  with Monte Carlo based multiple testing corrected  $p$ -values. For  $CIF_{mean}$  with Bonferroni corrections and for  $CIF_{mean}$  test-statistic (*Uncorrected*), the maximal value of  $mtry$  provided the highest performance (Figure 5.6).



**Figure 5.6:** DREAM5 performance results – variable  $mtry$ . The performance of the  $CIF_{mean}$  methods at various  $mtry$  values were assessed based on DREAM5 overall score averaged over 3 DREAM5 networks.

#### 5.4.4. Case Study: T1D data

Since for these data 1617 nodes and 212 samples were available, we restricted attention to  $CIF_{mean}$ -based GRNs. The performance in human real-life data from T1D patients significantly dropped across all  $CIF_{mean}$  methods. The performance differences between  $CIF_{mean}$  methodologies were minimal though. The best performer at default  $mtry=k/3$  was  $CIF_{mean}$   $p$ -value (*Monte Carlo*) followed by  $CIF_{mean}$   $p$ -value (*Bonferroni*) (supplementary Figure S5.4). The impact of the  $mtry$  parameter across  $CIF_{mean}$  methods greatly varied (Figure 5.7a). In case of the  $CIF_{mean}$   $p$ -value method with Monte Carlo multiple testing correction, the highest performance was achieved at the default  $mtry$  setting of  $k/3$ , although the performance is rather stable across the considered values for  $mtry$ . With Bonferroni correction, the maximum value of  $mtry$  at  $k$  gave the most performance benefits. The  $CIF_{mean}$   $p$ -value (*Uncorrected*) method showed the lowest performance changes with varying  $mtry$  parameter values (Figure 5.7a). The default  $mtry$  value of  $k/3$  was clearly suboptimal for  $CIF_{mean}$  test-statistic (*Uncorrected*) and  $CIF_{mean}$   $p$ -value (*Bonferroni*). The significant drop in AUROC and AUPR is even more striking in (Figure 5.7b).



**Figure 5.7:** The T1D Case study performance results – variable  $mtry$ . **a)** Performance of the  $CIFmean$  methods based on the AUROC and AUPR. **b)** A more detailed view of the AUROC and AUPR dynamics as a function of the  $mtry$  parameter. As the gold standard the verified set of TF/TG sets from [27].

From a practical point of view, considering a threshold  $p$ -value of 0.01, the GRN inferred with the best performer  $CIF_{mean}$   $p$ -value (Monte Carlo) in this data scenario, highlighted a total of 89

interactions. Amongst them is a highly significant pair involving forkhead box P1 (*FOXP1*) and the IL-2 receptor- $\alpha$  (*IL2RA*), with respective  $p$ -value based global variable importance measure of 0.0057 (Table 5.3). Both *IL2RA* and *FOXP1* are well known T1D markers involved in IL-2 signaling pathway [39]. Table 5.3 lists other significant pairs linked to *IL2RA*.

**Table 5.3:**  $CIF_{mean}$   $p$ -value (Monte Carlo) significant pairs from the T1D dataset

Gene A	Full Name A	Gene B	Full Name B	$p$ -value
<i>SMAD7</i>	SMAD family member 7	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.002308
<i>FOXP1</i>	forkhead box P1	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.0057
<i>FOXO3</i>	forkhead box O3	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.0073
<i>BCL3</i>	B-cell CLL/lymphoma 3	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.010556
<i>FOXA2</i>	forkhead box A2	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.01156
<i>STAT1</i>	signal transducer and activator of transcription 1, 91kDa	<i>IL2RA</i>	interleukin 2 receptor, alpha	0.013757

## 5.5. Discussion

Networks come in different flavors, depending on their aim and the biological entities that serve as input during their construction. Examples of networks include gene regulatory networks (GRN) [40], co-expression networks [41-43], differential networks [44], metabolic networks [45]. Our inferred networks were directed. Genes were taken as nodes and “variable importance measures” were taken as weights to edges. The measures were derived from conditional inference trees (*CITs*) or conditional inference forests (*CIFs*). The reason for relying on conditional inference trees rather than classic regression trees was that we were ultimately interested in developing a network inference method that enables the integration of different data types (for instance, methylome, genome and transcriptome data). These data types generate measurements on differential scales, requiring re-scaling in order to avoid biased selection of features. Specifically, Breiman’s Random Forests [20] are known to be biased towards features with larger number of possible splits [21].

In addition, correlations between features are frequent in biological data (e.g., co-expression networks rely on “correlations” between gene expressions). Rather than reducing the data to obtain independent features (e.g., via components theory which would complicate node definition and

interpretation), a method that can directly deal with correlated features is highly desirable. Our results showed that the conditional inference forests (*CIF*) framework can outperform classic Random Forest, especially when features are correlated or are of different measurement types as was demonstrated in DREAM4 data (Figure 5.2 and Figure S5.1).

In particular,  $CIF_{cond}$  applied to relatively small data from DREAM4 (100 nodes and sample size of 100), outperformed all other considered methods based on *AUPROC* and *AUPR* performance measures, including *RF*. The added value of  $CIF_{cond}$  to *RF* seems to be rather small at first sight, but given its theoretical optimality in the presence of correlated data (as is the case here: multiple genes are co-expressed), we would generally favor  $CIF_{cond}$  over *RF*. Note that only weak correlation patterns existed between gene expressions in DREAM4 data, averaged over all networks, only  $2.20 \pm 0.91\%$  of gene pairs showed a correlation  $> 0.3$  (supplementary Figure S5.5). Interestingly, for network 4 the *AUROC* of  $CIF_{cond}$  was largely suboptimal to *RF*, whereas for *AUPR*, the  $CIF_{cond}$  slightly outperformed *RF* (Figure 5.2). Clearly, single tree-based techniques are not to be recommended for GRN inference purposes (Figure 5.2).

Interestingly, having a closer look at DREAM4 scenarios and Figure 5.2 and *AUPR*,  $CIF_{mean}$  with a stopping-rule based on test statistics rather than *p*-values outperformed all other  $CIF_{mean}$  methodologies. This may be due to the fact that  $CIF_{mean} \text{ test-statistic (Uncorrected)}$  does not make any assumptions about the shape or nature of the test statistic's distribution. Hence, it would be interesting to investigate in more detail the relation between GS network properties, the nature of the input variables and the performance of  $CIF_{mean} \text{ test-statistic (Uncorrected)}$ , possibly combined with a *maxT* [46] approach to derive multiple testing corrected *p*-values. The same observation was made for DREAM2 data (Figure 5.4).

Among the *p*-value based  $CIF_{mean}$  methodologies,  $CIF_{mean} \text{ p-value (Monte Carlo)}$  was the best performer for DREAM4 (Figure S5.1) and DREAM2 (Figure S5.2). For DREAM5 data scenarios,  $CIF_{mean} \text{ p-value (Monte Carlo)}$  did not only outperform all *p*-value based  $CIF_{mean}$  methodologies, but also  $CIF_{mean}$  based on test statistic (Figure 5.5). In DREAM5,  $CIF_{mean} \text{ p-value (Monte Carlo)}$  was closely followed by the not correcting for multiple testing  $CIF_{mean}$  strategy -  $CIF_{mean} \text{ p-value (Uncorrected)}$  (Figure 5.5 and Figure S5.3). All these results seem to indicate the added value of

adjusting for multiplicity during node selection, despite it being more computationally intensive (Table 5.2).

The most optimal  $mtry$  value highly depended on the data scenario, respectively DREAM2, DREAM4 and DREAM5 (Figure 5.3, Figure 5.4 and Figure 5.6). For  $CIF_{mean} p\text{-value (Uncorrected)}$  methodology, the most optimal values were respectively  $k$  and  $k/3$ . For  $CIF_{mean} p\text{-value (Bonferroni)}$  they were 5 and  $k$  with  $k/3$  being a reasonable alternative. For  $CIF_{mean} p\text{-value (Monte Carlo)}$  they were  $k$  and  $k/3$  and for  $CIF_{mean} test\text{-statistic (Uncorrected)}$  they were  $k/3$  and  $k$ . It is only for DREAM5 data that the number of samples largely exceeds the number of the input variables (i.e. TFs) considered for analysis. Hence DREAM5 more closely resembles a classic regression context, compared to the other data scenarios, for which it has been shown that there is little improvement by using unpruned bagging strategies (i.e.  $mtry=k$ ). However, all data scenarios would typically not be handled in a classic regression framework. The higher the discrepancy between the number of samples compared to the number of input variables ( $p \gg n$ ; as is the case for most real-life data examples with human samples), the more we expect  $mtry=k$  to do well, as was observed for T1D data (Figure 5.7). For practical reasons, our  $CIF$ -based GRN inference framework takes  $mtry=k/3$  as a default.

From a theoretical point of view and on small data sets  $CIF_{cond}$  is to be preferred (Figure 5.2 and Figure S5.1). From a practical point of view, more work is needed to use  $CIF_{cond}$  principles for GRN inference purposes. Based on the DREAM4 data represented by networks composed of 100 nodes, the computation time of  $CIF_{mean} p\text{-value (Uncorrected)}$  was 12.35 minutes versus 0.79 minutes for RF (Table 5.2). Analyzing 4511 nodes of the DREAM5 network 2 took  $CIF_{mean} p\text{-value (Uncorrected)}$  3232 minutes versus 6054 minutes for RF. Hence, it seems that the larger the data, with the same  $mtry$  parameters, the larger the computation time advantage of  $CIF_{mean}$  over RF may be. Clearly, as  $CIF_{cond}$  already took approximately 2 hours analyzing 100 node network versus 12.35 minutes for  $CIF_{mean} p\text{-value (Uncorrected)}$ , it is infeasible to use it on large data sets at the moment. Modifying  $CIF_{cond}$  to reduce computation time is the subject of future projects.

The GRN inference in eukaryotic expression data is complex [47]. Therefore the drop in performance of  $CIF_{mean} p\text{-value (Monte Carlo)}$  (Figure 5.7), compared to for instance DREAM4

data (Figure 5.2), on type 1 diabetes (T1D) data is not surprising. Low correlation between expression levels between genes is possibly due to transcription factor regulation acting on the protein rather than on the transcript level via post-translational modifications, unknown latent variables, genes exhibiting functional overlaps, several levels of regulation not caused by transcription factor binding [48], epigenetic component and others. This may result in gold standard networks with heavy reliance on protein-protein binding but poor expression level changes [26]. Nevertheless, *CIF<sub>mean</sub>* identified highly relevant T1D genes connected to IL2RA, a well-known marker of T1D. This suggests the potentials of *CIF<sub>mean</sub>* – based GRN inference in unraveling biologically relevant mechanisms. Among the genes listed in Table 5.3 is *STAT1*, a member of the STAT protein family, which is critical in IL2 signaling and regulation of T cell activity. Perturbations in IL2 signaling pathway were found to be closely associated to onset of T1D, highlighting the importance of the immune system and cytokine signaling components [39]. The FOX family of proteins and BCL3 highlight the immune system involvement in T1D. The SMAD family of proteins is also key to T1D, as they are associated with TGF $\beta$  and BMP pathways. Mutations in *SMAD* genes are strongly associated with diabetes, as previously reported in [49]. Note that these results were obtained by taking the threshold of 0.01. The optimal threshold in the ROC curve (i.e. the point closest to the top-left part of the plot) was 0.0038 with specificity and 1-sensitivity confidence intervals of 0.55-0.56 and 0.45-0.52 respectively (based on 2000 bootstrap samples).

Finally, *CIF*'s separate node selection and splitting association steps coupled to general association measure based on framework developed by Strasser and Weber [36] offer opportunities to handle different input data types, for instance RNA-*seq* and microarray expression data. This framework is linked to derivation of permutation-based linear test-statistic measuring association between predictors and responses detailed in [50]. Performance of *CIF* based methodologies was not yet tested on RNA-*seq* data, which is often characterized by small sample sizes. Since RNA-*seq* transcriptome data are ideally modeled via a negative binomial regression model that considers overdispersion [51], we plan to expand the choice of association tests currently incorporated in *CIF* methodologies. In addition, since these tests will rely on a regression framework, they can potentially be adjusted for confounding factors. In brief, the *CIF* framework provides generality and flexibility for enhancements in many contexts, including integrative multi-omics data analysis.

In future work, our aim is to avoid a posteriori data integration (for instance fusing a methylome-transcriptome, genome-transcriptome, transcriptome-transcriptome networks via [52]), but to develop a feasible *CIF*-based gene regulatory network inference method that can handle methylome, genome and transcriptome data as joint input to predict gene expression.

## 5.6. Conclusions

In this work, we investigated the performance and practical use of conditional inference trees (*CITs*) and forest (*CIFs*) to infer gene regulatory networks from synthetic and real-life data. Synthetic data and data on model organisms were made available by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [4,25,29]. Real-life data on type 1 diabetes was obtained from the GEO public repository (GEO #: GSE43488). We have shown that conditional inference framework suggested by [5] offers interesting possibilities for data integration, provided computational efficiency can be enhanced. In real-life settings of high-dimensional biological data, we recommend to use the *CIF*-based GRN inference approach based on conditional inference forests and node-specific *p*-values, adjusted for multiplicity by Monte Carlo resampling. In addition, we recommend randomly selecting about  $1/3^{\text{rd}}$  of the input variables at each node in the forest. Although more computationally intensive, this approach is less dependent on the number of randomly selected variables at each node (*mtry*) than conditional inference trees with Bonferroni corrected *p*-values. Averaging node-specific *p*-values over trees in *CIF* ensembles and using these as variable importance scores to weight network edges, greatly facilitates construction of weighted networks such as GRNs. Indeed, for classic forests-derived variable importance scores it is not obvious to set a threshold defining which two nodes need to be connected in a network, unless *ROC* or *PR* curves are constructed based on gold standard and the optimal threshold is derived from those. The statistical framework that underpins *CIFs* naturally leads to setting an overall “significance” level, such as 0.01. Adopting this strategy on microarray gene-expression data for 121 type 1 diabetes patients and 1617 genes gave meaningful results, supported by the literature.

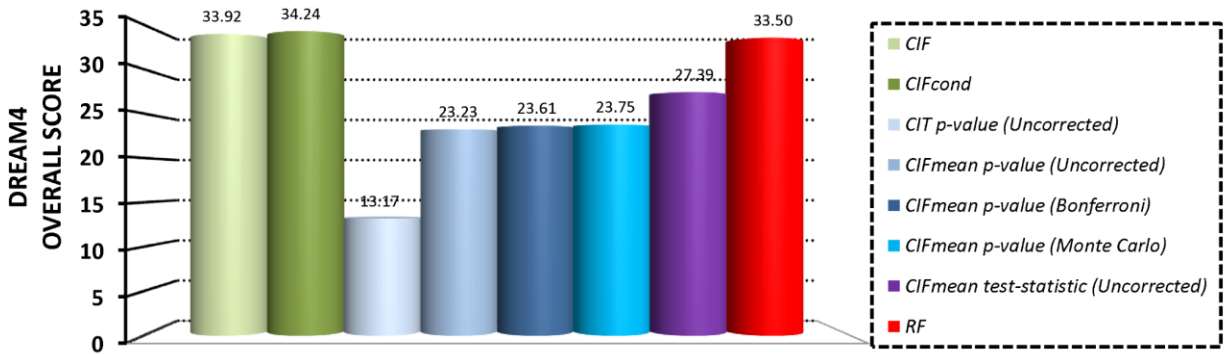
### 5.7. Chapter highlights

The *CIF*-based GRN inference methods developed in this chapter were compared amongst themselves and to classical *RF* using microarray expression data. Specifically, to our knowledge, it was the first time that *CIF*s have been applied for GRN inference (from microarray data). Methods built on *CIF<sub>mean</sub>* were shown to be scalable (avoiding computationally expensive permutation-based Variable Importance Measures), while providing an acceptable performance as compared to reference methods such as *RF*, *CIF* and *CIF<sub>cond</sub>* implemented in *randomForest* (version 4.6-7) [20,38] and *party* [6] R libraries. The *CIF<sub>mean</sub>* method can only work with single data source at a time. Thus, in the next chapter we will consider multiple omics sources at once to derive integrate gene-networks.

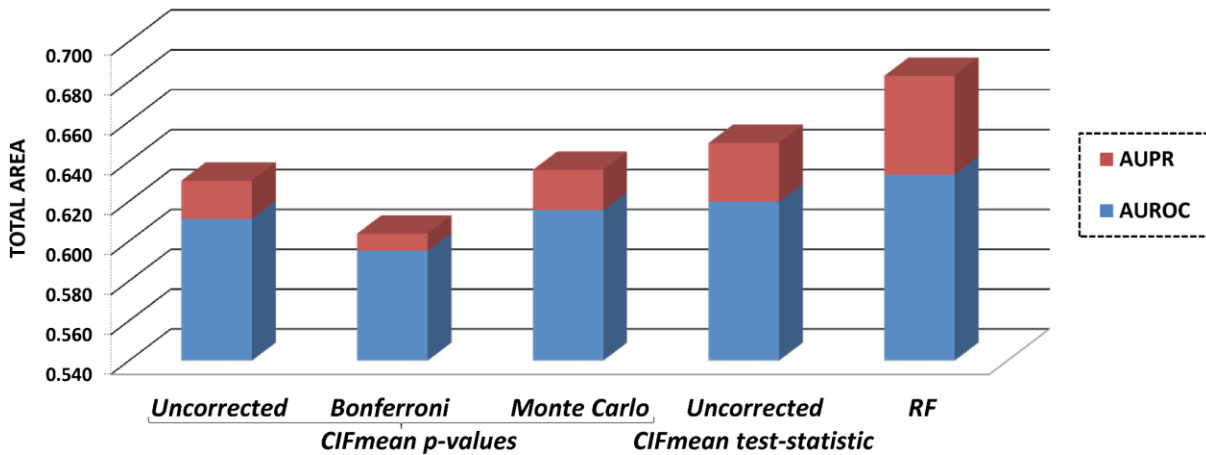
### 5.8. Acknowledgements and funding

This research was funded by the Fonds de la Recherche Scientifique (FNRS), in particular F.N.R.S. research grant n° 22333518 (KB) and research project convention n° 2.4609.11 (KVS, KB), and was carried out as part of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. We thank Pierre Geurts and Benjamin Dizier from the Systems Modeling Unit, Faculty of Applied Sciences, Université de Liège (Belgium) for constructive comments and suggestions throughout this project. Special thanks to Patrick E. Meyer from the Bioinformatics and Systems Biology Unit, Faculty of Sciences, Université de Liège (Belgium) for providing valuable feedback to earlier versions of this manuscript.

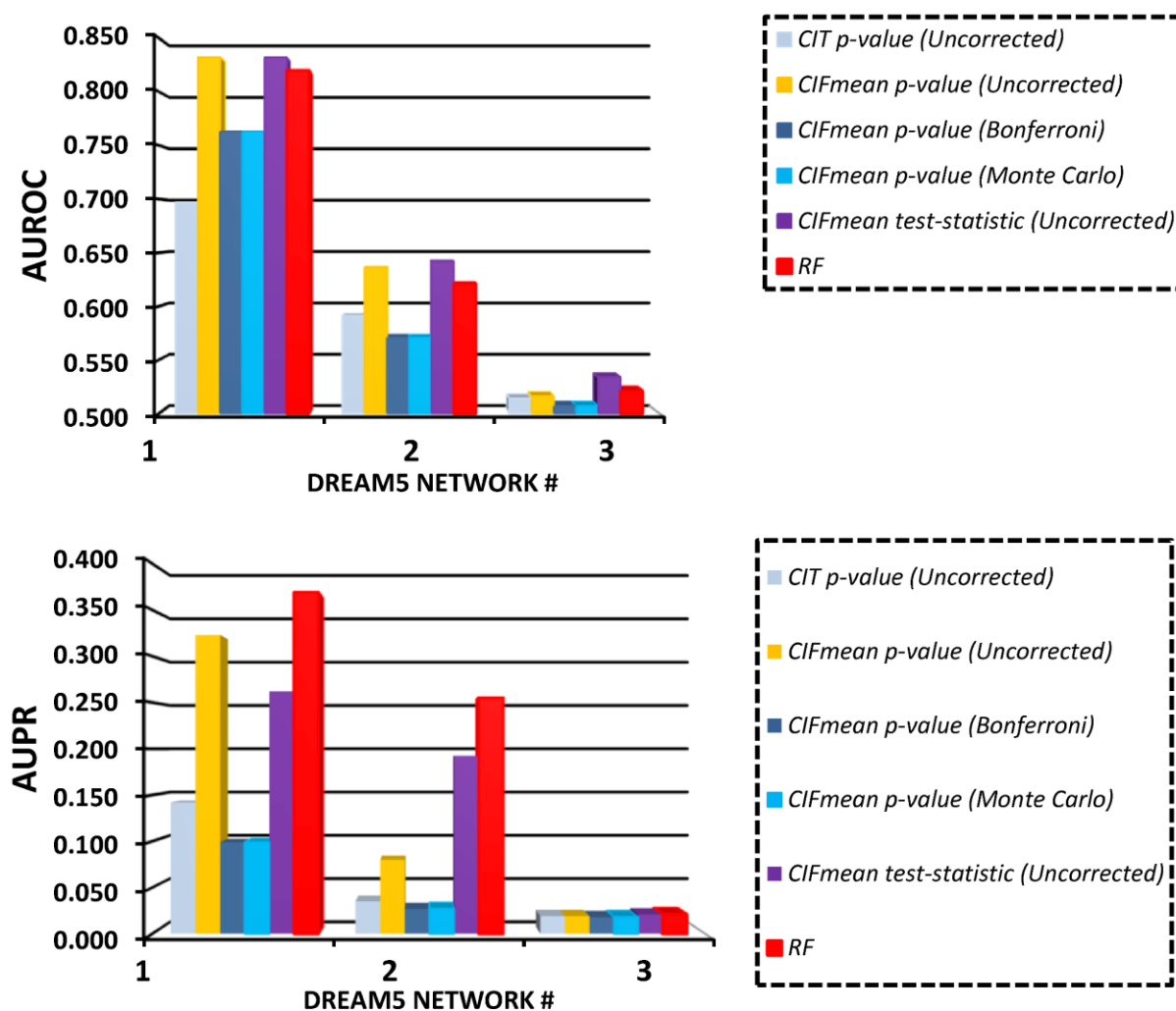
## 5.9. Appendix



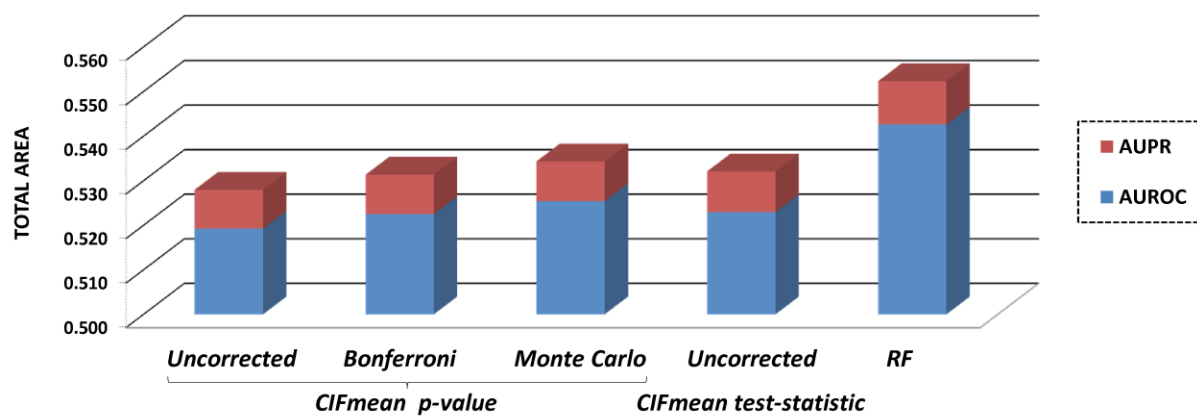
**Figure S5.1:** DREAM 4 performance results -  $mtry=k/3$ . The GRN inference performance levels across the 8 methodologies described in methods section. Performance is quantified via the DREAM4 overall score as defined in for instance [26].



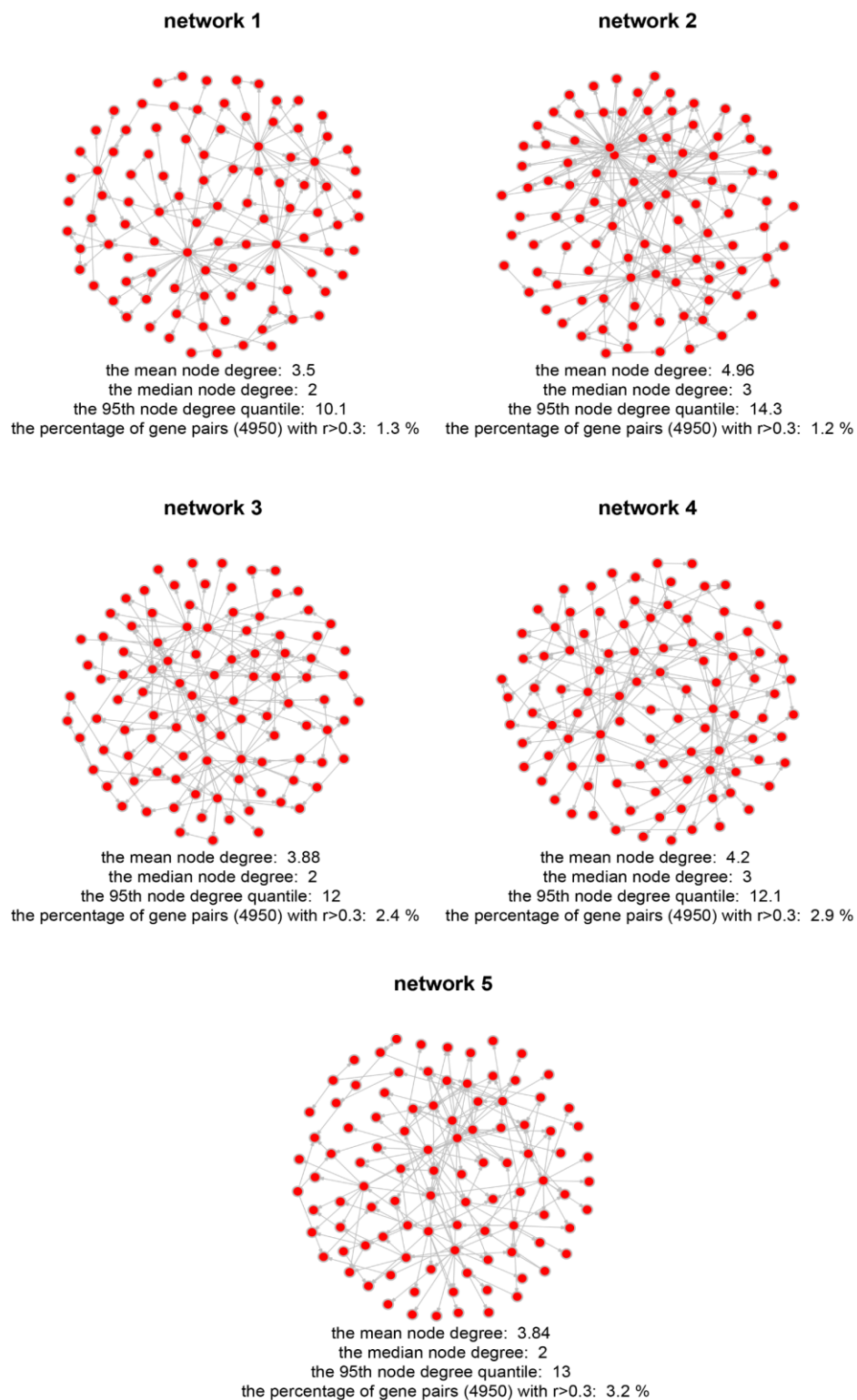
**Figure S5.2:** DREAM2 performance results -  $mtry=k/3$ . The performance of the  $CIF_{mean}$  and  $RF$  methods based on the total area of  $AUROC$  and  $AUPR$  using the default settings with the  $mtry=k/3$ .



**Figure S5.3:** DREAM5 performance results -  $mtry=k/3$  showing *AUROC* and *AUPR* per each network.



**Figure S5.4:** T1D Case study performance results -  $mtry=k/3$ . The performance of the  $CIF_{mean}$  methods based on the total area of AUROC and AUPR using the default settings with the  $mtry=k/3$ .



**Figure S5.5:** DREAM4 GS networks. The DREAM4 GS networks size 100 from 1 to 5 along with the basic network measures.

**Table S5.1:** DREAM4 methodology rankings - default settings

Method	Overall score*	Network 1		Network 2		Network 3		Network 4		Network 5	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
<i>CIF</i>	33.92	0.678	0.151	0.683	0.150	0.758	0.230	0.711	0.217	0.744	0.234
<i>CIFcond</i>	34.24	0.665	0.154	0.677	0.151	0.748	0.237	0.719	0.226	0.736	0.244
<i>CIT p-value (Uncorrected)</i>	13.17	0.574	0.075	0.562	0.075	0.626	0.156	0.613	0.105	0.612	0.085
<i>CIFmean p-value (Uncorrected)</i>	23.23	0.643	0.113	0.658	0.119	0.724	0.155	0.718	0.153	0.733	0.094
<i>CIFmean p-value (Bonferroni)</i>	23.61	0.713	0.102	0.690	0.112	0.743	0.144	0.738	0.133	0.754	0.075
<i>CIFmean p-value Bonferroni (ver. party)</i>	24.34	0.732	0.108	0.687	0.110	0.743	0.138	0.749	0.138	0.763	0.083
<i>CIFmean p-value (Monte Carlo)</i>	23.75	0.658	0.118	0.656	0.127	0.716	0.158	0.729	0.143	0.736	0.096
<i>CIFmean test-statistic (Uncorrected)</i>	27.39	0.703	0.101	0.697	0.147	0.735	0.178	0.743	0.169	0.758	0.095
<i>RF</i>	33.50	0.659	0.156	0.665	0.136	0.768	0.246	0.747	0.216	0.427	0.213

\*Based on the DREAM4 overall score criterion (see Methods section)

**Table S5.2:** DREAM2 methodology rankings - default settings

Method	AUROC	AUPR	precision (TP/TP+FP)					
			1*	2*	5*	20*	100*	200*
<i>CIT</i>	0.500	0.006	0.006	0.006	0.006	0.006	0.006	0.006
<i>CIFmean p-value (Univariate)</i>	0.611	0.019	0.048	0.080	0.089	0.196	0.130	0.020
<i>CIFmean p-value (Bonferroni)</i>	0.595	0.009	0.002	0.004	0.005	0.005	0.010	0.014
<i>CIFmean p-value (Monte Carlo)</i>	0.616	0.020	0.077	0.125	0.172	0.247	0.139	0.019
<i>CIFmean test-statistic (Uncorrected)</i>	0.620	0.029	0.333	0.182	0.278	0.417	0.242	0.024
<i>RF</i>	0.633	0.049	1.000	1.000	0.714	0.606	0.500	0.036

\*-precision values when number of true positives is 1,2,5,20,100 or 200

**Table S5.3:** DREAM5 methodology rankings - default settings

Method	Overall score*	Network 1		Network 2		Network 3	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
<i>CIT p-value (Uncorrected)</i>	0.69	0.697	0.142	0.591	0.035	0.514	0.019
<i>CIFmean p-value (Uncorrected)</i>	53.59	0.832	0.323	0.636	0.08	0.516	0.019
<i>CIFmean p-value (Bonferroni)</i>	6.32	0.763	0.099	0.571	0.027	0.507	0.018
<i>CIFmean p-value (Monte Carlo)</i>	54.02	0.832	0.322	0.641	0.081	0.516	0.019
<i>CIFmean test-statistic (Uncorrected)</i>	42.66	0.832	0.262	0.642	0.066	0.534	0.021
<i>RF</i>	63.28	0.82	0.369	0.621	0.088	0.521	0.021

\*Based on the DREAM5 overall score criterion (see Methods section)

## 5.10. References

1. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) **Conditional variable importance for random forests.** *BMC Bioinformatics* 9: 307.
2. Sherlock G (2000) **Analysis of large-scale gene expression data.** *Current opinion in immunology* 12: 201-205.
3. Goeman JJ, Buhlmann P (2007) **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 23: 980-987.
4. Marbach D, Schaffter T, Mattiussi C, Floreano D (2009) **Generating realistic in silico gene networks for performance assessment of reverse engineering methods.** *J Comput Biol* 16: 229-239.
5. Hothorn T, Hornik K, Zeileis A (2006) **Unbiased recursive partitioning: A conditional inference framework.** *Journal of Computational and Graphical statistics* 15: 651-674.
6. Hothorn T, Hornik K, Strobl C, Zeileis A, Hothorn MT (2014) **Package ‘party’.** *Package Reference Manual for Party Version 09-998* 16: 37.
7. Johnstone IM, Titterton DM (2009) **Statistical challenges of high-dimensional data.** *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367: 4237-4253.
8. Hardiman G (2004) **Microarray platforms-comparisons and contrasts.** *Pharmacogenomics* 5: 487-502.
9. Yao F, Coquery J, Le Cao KA (2012) **Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets.** *BMC Bioinformatics* 13: 24.
10. Lin YY, Liu TL, Fuh CS (2011) **Multiple kernel learning for dimensionality reduction.** *IEEE Trans Pattern Anal Mach Intell* 33: 1147-1160.
11. Zhu X, Gerstein M, Snyder M (2007) **Getting connected: analysis and principles of biological networks.** *Genes Dev* 21: 1010-1024.
12. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, et al. (2015) **Widespread macromolecular interaction perturbations in human genetic disorders.** *Cell* 161: 647-660.
13. Barabasi AL, Oltvai ZN (2004) **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 5: 101-113.
14. Levy ED, Landry CR, Michnick SW (2009) **How Perfect Can Protein Interactomes Be?** p11-p11 p.
15. Davidson E, Levin M (2005) **Gene regulatory networks.** *Proceedings of the National Academy of Sciences of the United States of America* 102: 4935-4935.
16. Cho DY, Kim YA, Przytycka TM (2012) **Chapter 5: Network biology approach to complex diseases.** *PLoS Comput Biol* 8: e1002820.
17. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) **Inferring regulatory networks from expression data using tree-based methods.** *PLoS One* 5.
18. Boulesteix AL, Janitza S, Hapfelmeier A, Van Steen K, Strobl C (2015) **Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests.** *Brief Bioinform* 16: 338-345.
19. Loh WY (2011) **Classification and regression trees.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1: 14-23.
20. Breiman L (2001) **Random forests.** *Machine learning* 45: 5-32.
21. Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) **Bias in random forest variable importance measures: illustrations, sources and a solution.** *BMC Bioinformatics* 8: 25.
22. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) **Conditional variable importance for random forests.** *BMC bioinformatics* 9: 307.
23. Strobl C, Hothorn T, Zeileis A (2009) **Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package.** University of Munich.

24. Boulesteix AL, Janitza S, Kruppa J, König IR (2012) **Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics**. University of Munich. 129-129.
25. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) **Revealing strengths and weaknesses of methods for gene network inference**. *Proc Natl Acad Sci U S A* 107: 6286-6291.
26. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. (2012) **Wisdom of crowds for robust gene network inference**. *Nat Methods* 9: 796-804.
27. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, et al. (2010) **Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data**. *Nucleic Acids Res* 38: e120.
28. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, et al. (2013) **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic Acids Res* 41: D203-213.
29. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, et al. (2010) **Towards a rigorous assessment of systems biology models: the DREAM3 challenges**. *PLoS One* 5: e9202.
30. Huynh-Thu VA, Wehenkel L, Geurts P (2013) **Gene regulatory network inference from systems genetics data using tree-based methods**. *Gene Network Inference-Verification of Methods for Systems Genetics Data*.
31. Kallionpää H, Elo LL, Laajala E, Mykkanen J, Ricano-Ponce I, et al. (2014) **Innate immune activity is detected prior to seroconversion in children with HLA-conferred type 1 diabetes susceptibility**. *Diabetes* 63: 2402-2414.
32. Jiang C, Xuan Z, Zhao F, Zhang MQ (2007) **TRED: a transcriptional regulatory element database, new entries and other development**. *Nucleic Acids Res* 35: D137-140.
33. Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, et al. (2009) **The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences**. *Nucleic Acids Res* 37: D54-60.
34. Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, et al. (2002) **Transcription Regulatory Regions Database (TRRD): its status in 2002**. *Nucleic Acids Res* 30: 312-317.
35. Davis J, Goadrich M. **The relationship between precision-recall and ROC curves**; 2006; Pittsburgh, PA.
36. Strasser H, Weber C (1999) **On the asymptotic theory of permutation statistics**.
37. Boulesteix AL, Janitza S, Kruppa J, König IR (2012) **Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2: 493-507.
38. Liaw A, Wiener M (2002) **Classification and Regression by randomForest**. *R news* 2: 18-22.
39. Hulme MA, Wasserfall CH, Atkinson MA, Brusko TM (2012) **Central role for interleukin-2 in type 1 diabetes**. *Diabetes* 61: 14-22.
40. Baitaluk M, Kozhenkov S, Ponomarenko J (2012) **An integrative approach to inferring gene regulatory module networks**. *PLoS One* 7: e52836.
41. Wang YX, Huang H (2014) **Review on statistical methods for gene network reconstruction using expression data**. *J Theor Biol*.
42. Horvath S, Dong J (2008) **Geometric interpretation of gene coexpression network analysis**. *PLoS Comput Biol* 4: e1000117.
43. Langfelder P, Horvath S (2008) **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 9: 559.
44. Ideker T, Krogan NJ (2012) **Differential network biology**. *Mol Syst Biol* 8: 565.
45. Maarleveld TR, Khandelwal RA, Olivier BG, Teusink B, Bruggeman FJ (2013) **Basic concepts and principles of stoichiometric modeling of metabolic networks**. *Biotechnol J* 8: 997-1008.

46. Westfall PH (1993) **Resampling-based multiple testing: Examples and methods for p-value adjustment**. John Wiley & Sons.
47. Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K (2009) **Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks**. *BMC systems biology* 3: 49.
48. Hu Z, Killion PJ, Iyer VR (2007) **Genetic reconstruction of a functional transcriptional regulatory network**. *Nature genetics* 39: 683-687.
49. McKnight A, Woodman A, Parkkonen M, Patterson C, Savage D, et al. (2009) **Investigation of DNA polymorphisms in SMAD genes for genetic predisposition to diabetic nephropathy in patients with type 1 diabetes mellitus**. *Diabetologia* 52: 844-849.
50. Zeileis A, Wiel MA, Hornik K, Hothorn T (2008) **Implementing a class of permutation tests: the coin package**. *Journal of Statistical Software* 28: 1-23.
51. Anders S, Huber W (2010) **Differential expression analysis for sequence count data**. *Genome biol* 11: R106.
52. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) **Similarity network fusion for aggregating data types on a genomic scale**. *Nat Methods* 11: 333-337.

## Chapter 6: Contributed work

### Synergetic interactions between theory and practice



#### Related publications:

Gadaleta F, Bessonov K\*, Van Steen K (2016) **Integration of Gene Expression and Methylation to unravel biological networks in glioblastoma patients.** (*Genetic Epidemiology* –accepted for publication)

Schleich F., Bessonov K, Van Steen K (2015). **Exhaled volatile organic compounds are able to discriminate between neutrophilic and eosinophilic asthma.** (*Patent #203-17 – submitted*)



## 6. Contributed work

### 6.1. Chapter summary

Methods development and practical application should go hand in hand. Therefore, we endeavoured on multiple projects in which interaction tools (such as conditional inference trees) and network-based integration (such as those encapsulated in a regression framework) can be applied, refined, and, thus, better understood.

This chapter briefly describes contributed work including projects involving conditional inference framework application to asthma sub-type classification (Section 6.2) followed by integration of methylome and transcriptome data via *Regression2Net* method (Section 6.3)

In Section 6.2, we briefly describe the work that was contributed to the University of Liege patent #203-17 - Method for the Diagnosis of Airway Disease Inflammatory Subtype. Using conditional inference forests (*CIFs*) inference framework introduced in previous chapters, we created binary classifier based on the most highly relevant volatile organic compounds (VOCs) that was able to discriminate between asthma sub-types. Kyrylo Bessonov's contribution consisted in the analysis of the exhaled breath profiles of patients diagnosed with various sub-types of asthma.

In Section 6.3 we describe *Regression2Net*. This involves network inference followed by network analysis. As disease case context we consider glioblastoma multiforme. We infer separate gene-gene networks only using transcriptome data (Expression-Expression interactions) or combining transcriptome and methylome data (including Expression-Methylation relations).

The *Regression2Net* integrated analysis project provided interesting results with regards to gene regulation by consideration of both methylation and expression components. In particular, strong methylation component, typical to cancers, and complex regulatory gene expression patterns were confirmed. The inferred networks highlighted common and distinct functional components shared between transcriptome and methylome data sources. For example, genes involved in energy metabolism pathways were shared between the two data components while genes involved in various cancer types, cell cycle control and immune system responses pathways were not. Overall,

the *Regression2Net* method is an example of the third integration strategy 3 introduced in Section 2.5 (analyze each data source separately). *Regression2Net* is based on penalized regression that was practically applied to concrete real-life dataset yielding biologically interesting results of methylation-expression gene regulation under the context of cancer data.

Kyrylo Bessonov's contributions to the *Regression2Net* project included joint code development of the pathway enrichment module, pathway enrichment analysis, results presentation, interpretation and validation in the context of the disease pathology, and final manuscript drafting.

## 6.2. Identification of asthma sub-types via breath profiles analysis (patent #203-17)

### 6.2.1. Section summary

This section of Chapter 6 briefly describes the work that was contributed to the University of Liege patent #203-17 - Method for the Diagnosis of Airway Disease Inflammatory Subtype. The contribution consisted in the analysis of the exhaled breath profiles of patients diagnosed with various sub-types of asthma. Using conditional inference forests (*CIFs*) inference framework introduced in previous chapters, we were able to create binary classifier based on the most highly relevant volatile organic compounds (VOCs) to discriminate between asthma sub-types.

**Problem:** Classification of asthma sub-types is important in clinical setting as subpopulation of asthmatics respond differently to anti-inflammatory treatment. Depending on four known types of asthma, the treatment therapies vary. Administration of the wrong therapy to asthmatic patients can cause side effects and low efficacy of the treatment [1]. The breath volatile organic compounds (VOC) can be used to classify asthmatics and are relatively easy to obtain. Using the sparse VOC data on 276 asthmatics can the *CIF*-based classifier provide acceptable performance based on AUROC and AUPR measures? Are identified VOCs individually provide a good binary classification performance? What are optimal quality control and *CIF* settings leading to highest classification performance?

**Results:** Three binary classification scenarios were considered eosinophilic / neutrophilic (*E/N*), eosinophilic / paucigranulocytic (*E/P*), neutrophilic / paucigranulocytic (*N/P*) asthma subtype classification scenarios. The *CIFs* under binary classification scenarios identified 3,7-dimethylnonane, nonanal, hexane, 1-propanol, 2-hexanone, 3-tetradecene and pentadecene as the most class discriminating VOCs. The classification AUROC and AUPR values for *E/N* were 0.8844 and 0.9193, for *E/P* were 0.9945 and 0.9757, for *N/P* 0.8459 and 0.8090

**Keywords:** classification, *CIFs*, VOCs, asthma

### 6.2.2. Introduction

Asthma is a complex disease with many inflammatory subtypes and complex underlying regulatory mechanisms as was mentioned in Chapter 4. Classification of asthma subtypes via volatile organic compounds (VOC) profiles was proven to be a successful and non-invasive approach in brain, prostate and lung cancers. Breath volatile organic compounds originate from endogenous products of metabolism or from ingested exogenous sources such as food and water. Application of VOC profiling in asthma patients holds a high potential for early disease detection and subtyping together with patient monitoring capabilities. There are four distinct inflammatory subtypes of asthma: eosinophilic, neutrophilic, mixed granulocytic and paucigranulocytic. Currently, the popular non-invasive detection includes the collection of mucus from lower airways (i.e. sputum) followed by cell count (eosinophils, neutrophils, etc.). Based on the sputum cell count, the four asthma subtypes are more precisely classified as follows: 1) eosinophilic subtype ( $>3\%$  eosinophils in the sputum); 2) neutrophilic subtype ( $\approx 76\%$  neutrophils); 3) paucigranulocytic subtype ( $<3\%$  eosinophils and  $<76\%$  neutrophils); 4) mixed granulocytic subtype ( $>3\%$  eosinophils and  $>76\%$  neutrophils) [2]. The appearance of new therapies to treat asthma calls for more accurate asthma sub-type detection. Some of the treatments include administration of inhaled corticosteroids (ICS) [3] and administration of monoclonal anti-IL-5 antibodies [4] and others. For example, in the case of eosinophilic asthma subtype, administration of ICS is highly beneficial since it quickly reduces the percentage of eosinophils contained in the sputum from asthmatics and represses the release of Th2 cytokines from lymphocytes and eotaxin from epithelial cells [3]. Each asthma subtype is characterized by different responses to the same therapy in terms of efficacy and adverse reactions. For example, inhaled corticosteroids treatment is the most popular and efficient in eosinophilic (allergic) asthma cases, but are inefficient or even detrimental in paucigranulocytic and neutrophilic asthma subtypes. Thus, it is clinically relevant to correctly diagnose the asthma subtype in order to increase treatment efficiency and reduce risks of adverse effects in patients.

The proposed analysis that addresses the above needs relies on feature selection and binary classification analysis based on conditional inference forest (*CIF*) framework [5] presented in Section 5.3.2. The identified VOCs provide the best discrimination between asthma types based on variable importance (*VIM*) and classification accuracy measures.

### 6.2.3. Methods

The exhaled breath of 276 patients was analyzed for the presence of volatile organic compounds (VOCs). The patients were recruited from the University of Liege (Belgium) asthma clinic between October 8, 2010 and January 2014. The 122 patients were classified having eosinophilic, 90 paucigranulocytic, 50 neutrophilic and 14 mixed granulocytic asthma. The mixed type of asthma was not included in the analysis due to low sample size and rarity. Gas chromatography and time-of-flight mass spectrometry (GC/MS) was used to identify VOCs present in exhaled breath from patients. Total of 3327 VOCs across samples were detected via GC/MS. Thus, the VOC numbers are in the 1-3327 range with VOC# representing the consecutive number of the input data matrix of 276 x 3328 dimensions including the class column (i.e. response). Conditional inference forests (CIFs) were used to build the ensemble of trees (i.e. forest) used to estimate discrimination importance (i.e. asthma subtype prediction) of each VOC. We tested binary eosinophilic / neutrophilic (*E/N*), eosinophilic / paucigranulocytic (*E/P*), neutrophilic / paucigranulocytic (*N/P*) asthma subtype classification scenarios. The main protocol consisted of three main steps:

**1) Data quality control (QC) criteria.** Since data was sparse, extra quality control step need to be taken. It consisted in the requirement of minimum of 10 subjects per each VOC in order to increase results reliability. The input data consists of area unit (AU) values of the gas chromatographic peaks of each detected VOC compound. The *E/N*, *E/P* and *N/P* scenarios contained 561, 714, and 429 QC'ed VOCs respectively.

**2) Tree building parameters.** The Conditional Inference Forest from *party* R package library [5,6] and *cforest\_control()* used the following parameters: *cforest\_control(teststat="quad", testtype="MonteCarlo", fraction=0.65, replace=F, mincriterion=0.99, minsplit=30, ntree=999, nresample=9999, mtry=0, maxdepth=0, savesplitstats = F)*. The default high number of resamplings fixed at 9999 allowed to more accurately estimate null distribution. The minimum number of samples per node variable fixed at 30 also provided the stringent selection criteria. These changes were taken as an extra precaution against data sparsity. A total of 999 conditional inference trees were built on the 75% of patient samples (i.e training data). The significance threshold of 0.01 ( $p$ -value < 0.01) specified by *mincriterion*=0.99 was applied. The stopping criteria included no restriction on the tree depth (*maxdepth*=0) and minimum of 30 samples per tree node

(*minsplit=30*). In addition, the set *mtry=0* parameter allowed to consider all predictor variables (i.e. 3327 VOCs) as node variables candidates.

**3) Selection of VOCs.** The variable selection consisted in the calculation of classical variable importance measure (*VIM*) defined by ‘mean decrease in MSE’ (*%IncMSE*) for all QC’ed predictor variables in each scenario (VOCs). Each analysis yielded a ranked list of predictor variables calculated with the help of *varimp()* function defined in the *party* library [5,6]. The *VIM* represented the strength of the VOCs asthma subclass prediction ability (i.e. discriminatory power).

**4) Classification performance.** The *CIF* from each scenario (*E/N*, *E/P* and *N/P*) was used to classify samples from the test data representing the 25% of the original samples. The standard classification measures including the area under the *ROC* and *PR* curves - *AUROC* and *AUPR* assessed *CIF* classification performance.

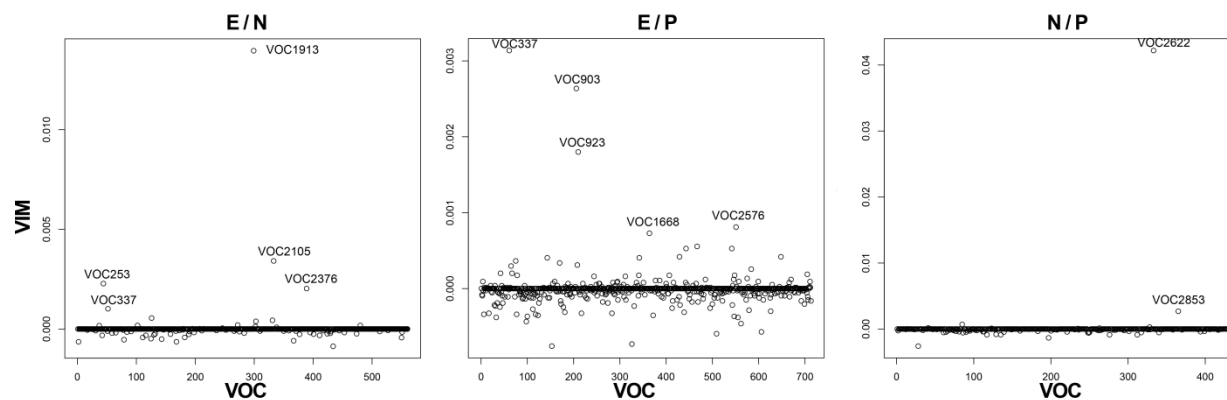
**5) Single VOC classification.** The binary classification via a single VOC was implemented in order to simplify and increase transparency of the classification process. The classification was done via the mean VOC areas of previously selected 9 VOCs via the *CIFs*. The mean areas were calculated for each VOC for each of the 3 considered asthma types (*E*, *N* and *P*). The default class was assigned to a class with the largest mean area. For example, the default class for VOC1913 is neutrophilic with area units (AU) of 14,343,580 and 66,581,501 corresponding to *E* and *N* classes. All patients with the selected VOC areas greater than the mean were assigned to the default class. Meanwhile, all patients with less or equal areas were assigned to the other class. For example, given two classes {eos} and {neutro} with the largest mean area of 66,581,500 units in the {eos} class, the patients would be classified according to these two simple classification rules: {AU >= 66,581,500.8} → {neutro}, {AU < 66,581,500.8} → {eos}. The performance of such simple classification rules was tested in all three scenarios and accuracy and precision are reported in Figure 6.2.4.

### 6.2.4. Results

The variable selection identified the most important variables with respect to asthma classification in the *E/N*, *E/P* and *N/P* scenarios. Only a few VOCs obtained higher *VIM* values across three scenarios as shown in Figure 6.2.1. Please refer to Table 6.2.1 for VOC chemical identity. VOC337 (hexane) was deemed important in both *E/N* and *E/P* with highest importance value in *E/P* scenario. Overlap in other VOCs was not observed. The highest *VIM* absolute values were seen in *E/N* classification scenario. Total of 5, 5 and 2 VOCs from the *E/N*, *E/P* and *N/P* scenarios were selected as the most important VOCs (Figure 6.2.1).

The areas of the selected VOCs, representing the compound amount, were compared in the three asthma sub-classes. The 9 boxplots in Figure 6.2.3 allow visually assess discriminatory power of each VOC based on another criterion – the area distribution and the mean of each asthma sub-class. The largest variance in areas between the 0.25 and 0.75 quantiles is quite significant especially in the neutrophilic sub-group (Figure 6.2.3) There was a positive trend observed between *VIM* values and the mean area differences between the sub-groups. For example, VOC1913 and VOC337 have the largest mean differences and also *VIM* values (Figure 6.2.1 and Figure 6.2.3).

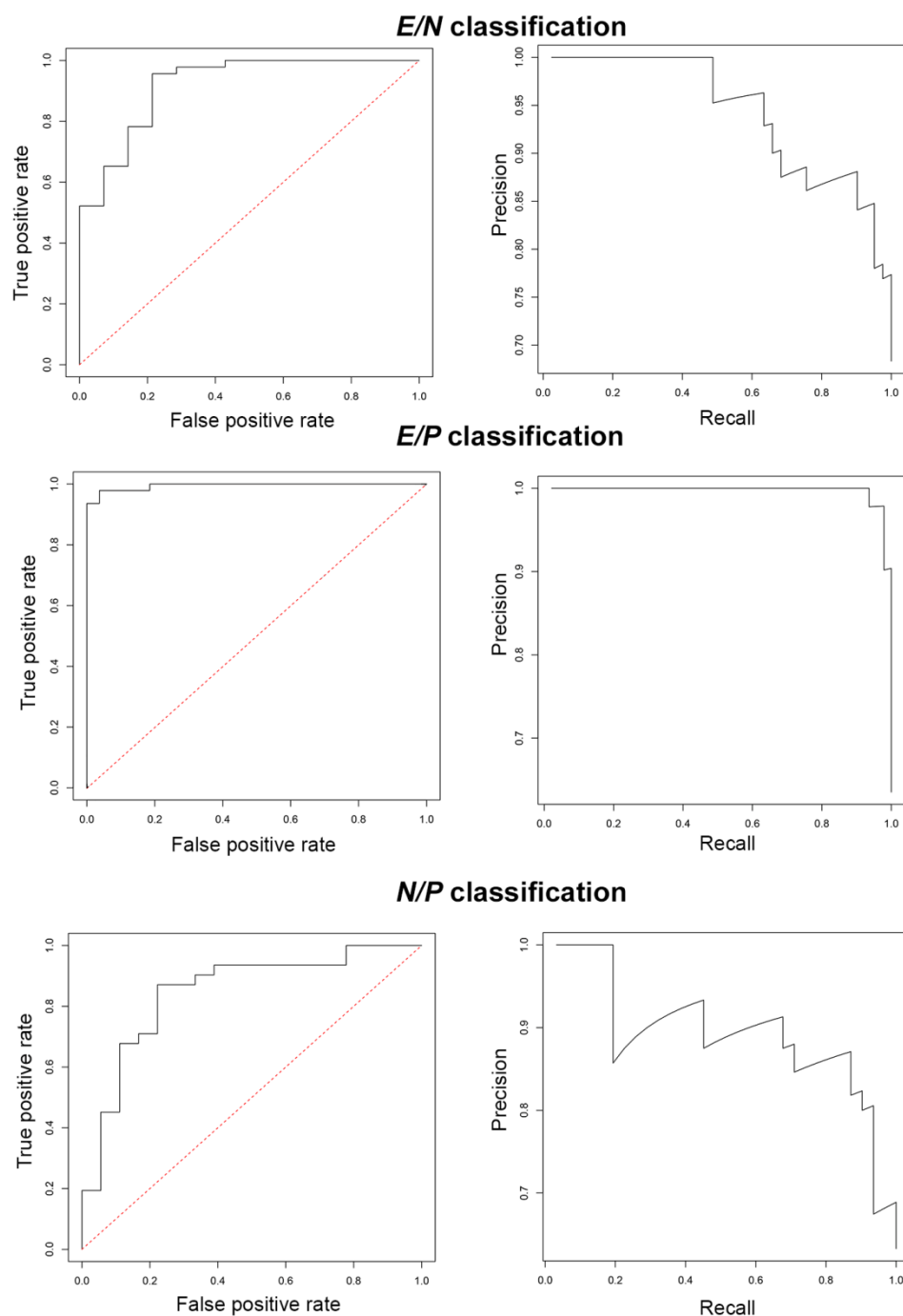
The box plots also showed presence of outliers in each top-performing VOCs shown as dots in Figure 6.2.3. The VOC areas are quite variable across patients even within each of three groups. After identification of most influential VOCs in predicting asthma sub-class, the previously built CIFs were tested under classification context. The adopted *CIF*-based method achieved extremely good classification performance close to theoretical maximums as indicated by *AUROC* and *AUPR* curves shown in Figure 6.2.3.



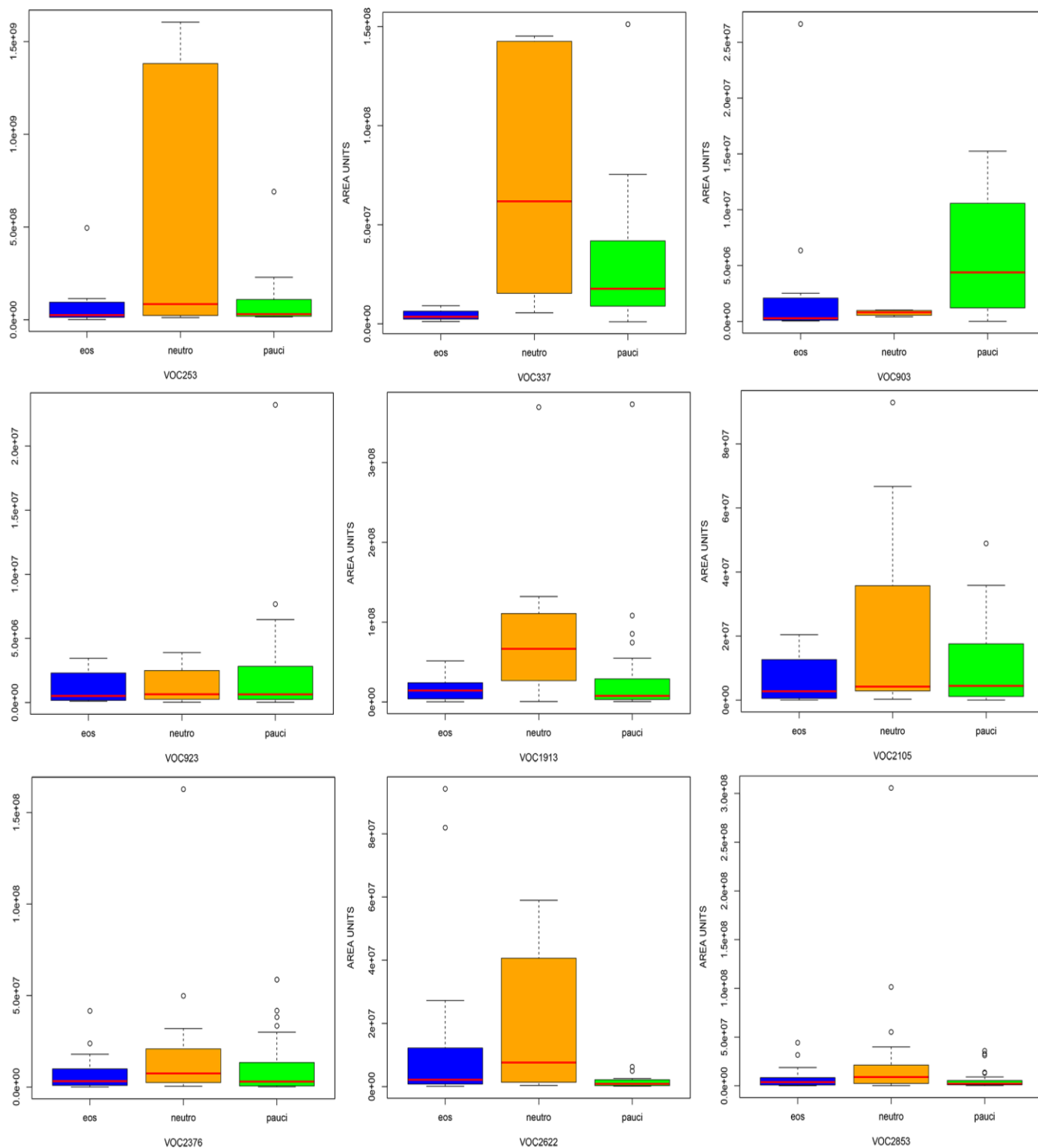
**Figure 6.2.1:** variable importance measure (*VIM*) in *E/N*, *E/P* and *N/P* binary sub-asthma classification scenarios calculated from conditional inference forests (*CIFs*). Legend: *E* – eosinophilic, *N* – neutrophilic, *P* – paucigranulocytic asthma. For chemical identities of highlighted VOCs please refer to Table 6.2.1.

**Table 6.2.1:** VOC mappings to compound names

VOC ID	Compound name	Formula
VOC1913	3,7-dimethylnonane	C <sub>11</sub> H <sub>24</sub>
VOC2105	nonanal	C <sub>9</sub> H <sub>18</sub> O
VOC2376	-	-
VOC337	hexane	C <sub>6</sub> H <sub>14</sub>
VOC253	1-propanol	C <sub>3</sub> H <sub>8</sub> O
VOC903	2-hexanone	C <sub>6</sub> H <sub>12</sub> O
VOC923	unknown	-
VOC2622	3-tetradecene	C <sub>14</sub> H <sub>28</sub>
VOC2853	pentadecene	C <sub>15</sub> H <sub>30</sub>



**Figure 6.2.2:** AUROC and AUPR curves from the *CIFs* binary classification of the *N/P*, *E/P* and *N/P* scenarios where E-eosinophilic, N- neutrophilic and P-paucigranulocytic athma sub-types. The red line diagonal line indicates random guess – the minimal classifier threshold. Table 6.2.2 lists areas under the curves.



**Figure 6.2.3:** VOCs box plots across three sub-types of asthma (Eosinophilic, Neutrophilic and Paucigranulocytic). The red line indicates mean area under the peak and ‘eos’, ‘neutro’ and ‘pauci’ are asthma types.

**Table 6.2.2:** *CIFs* classifier performance in asthma type binary classification

Classification task	AUROC	AUPR
<i>E/N</i>	0.8844	0.9193
<i>E/P</i>	0.9945	0.9757
<i>N/P</i>	0.8459	0.8090

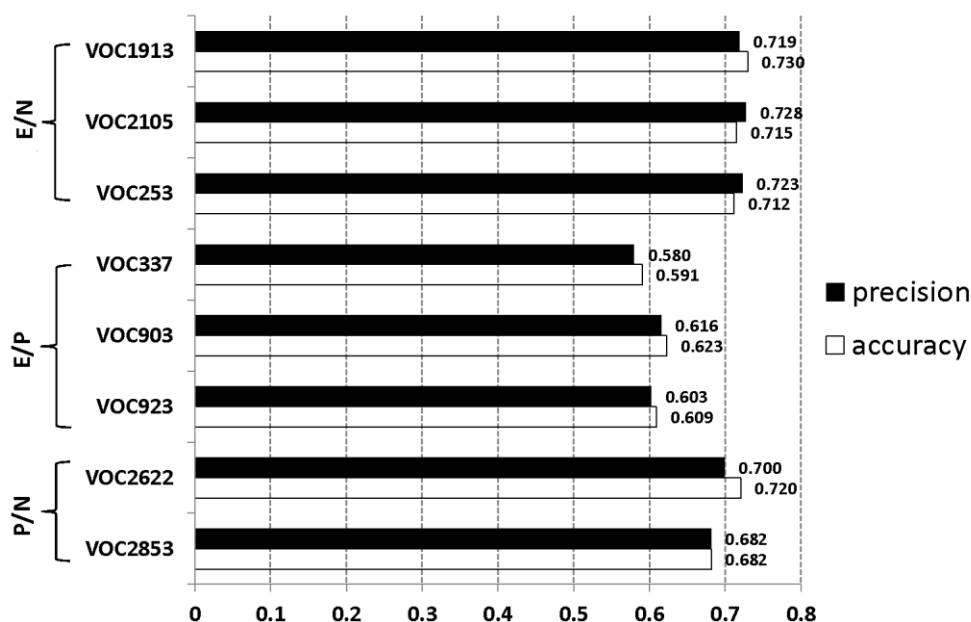
\* **legend:** *E/P* - Eosinophilic versus paucigranulocytic; *N/P* - neutrophilic versus paucigranulocytic; *E/N* - eosinophilic versus neutrophilic;

One of the important aims of this study was applicability, easy of use, and direct translation of the results to clinic settings. Since classification of new patients entails use of previously inferred *CIFs* composed of tree ensembles, the patient classification via *CIFs* requires computing facilities and bioinformatics expertise. A simpler classification relying on areas of individual VOCs was implemented as single VOC classification rules based on AU cutoff thresholds shown in Table 6.2.3. Performance of each VOC classifier rule was assessed via accuracy and precision. The selected 9 VOCs also show good performance (Figure 6.2.4) with accuracy and precision in 0.58-0.728 and 0.591– 0.73 ranges, respectively. The highest binary classification performance via single VOC approach is seen in *P/N*, followed by *E/N* and *E/P* scenarios.

**Table 6.2.3:** Single VOC classification rules based on area units (AU)

VOC	scenario	Classification rule
VOC1913	<i>E/N</i>	{AU >= 66581500.8} → {neutro}, {AU < 66581500.8} → {eos}
VOC2105	<i>E/N</i>	{AU >= 4221547.1} → {neutro}, {AU < 4221547.1} → {eos}
VOC253	<i>E/N</i>	{AU >= 84535600.7} → {neutro}, {AU < 84535600.7} → {eos}
VOC337	<i>E/P</i>	{AU >= 17700387.4} → {pauci}, {AU < 17700387.4} → {eos}
VOC903	<i>E/P</i>	{AU >= 4388495.2} → {pauci}, {AU < 4388495.2} → {eos}
VOC923	<i>E/P</i>	{AU >= 638517.7} → {pauci}, {AU < 638517.7} → {eos}
VOC2622	<i>P/N</i>	{AU >= 7618566.1} → {neutro}, {AU < 7618566.1} → {pauci}
VOC2853	<i>P/N</i>	{AU >= 8955716.6} → {neutro}, {AU < 8955716.6} → {pauci}

\* **legend:** *E/P* - Eosinophilic versus paucigranulocytic; *N/P* - neutrophilic versus paucigranulocytic; *E/N* - eosinophilic versus neutrophilic;



**Figure 6.2.4:** Accuracy and precision of each individual VOC under *N/P*, *E/P* and *N/P* scenarios where *E*- eosinophilic, *N*- neutrophilic and *P*-paucigranulocytic athma sub-types. Table 6.2.1 provides VOC number to chemical name conversion.

### 6.2.5. Discussion

The results proved that that patient classification via *CIFs* is possible despite scarcity and high variability of the VOC input dataset. The adopted *CIF*-based method achieved excellent performance indicators both in variable selection and classification domains (Figure 6.2.1 and Figure 6.2.2). Interestingly the number of VOCs with high *VIMs* was rather small indicating a complex architecture of the VOC input data as stated above. The most complex scenario was *N/P* where only 2 VOCs were selected and the lowest AUROC and AUPR values compared to other classification scenarios were observed (Table 6.2.2). Conversely, the *E/P* was the simplest classification scenario obtaining the largest AUROC and AUPR values. The specificity of *CIFs* was also highlighted by the minimal overlap between identified VOCs considering each binary classification scenario. This shows that *CIFs* are diverse across considered scenarios.

The boxplots provided insights into the meaning of the *VIM* values assigned by *CIFs* (Figure 6.2.3). Most of the 9 selected VOCs show low variability in their means but rather large variance

indicating the heterogeneity of data and fluctuation in areas (Figure 6.2.3). The VOCs with highest mean variation across the 3 asthma subtypes were VOC337, VOC903 and VOC1913. These three VOCs obtained highest *VIM* in E/P and E/N scenarios (Figure 6.2.1) indicating a positive link between *VIM* and mean differences across the groups. In addition, large variance in areas of the selected VOCs (Figure 6.2.3) suggested that use of areas of single VOCs for classification purposes is sub-optimal as compared to *CIF*-based classification.

Nevertheless, the classification performance of a single VOC was tested due to simplicity and ease of application in clinical settings. The classification rules applied to the VOC dataset are detailed in Table 6.2.3. In general, rules based on a single VOC achieved a lower classification performance as compared to *CIFs*. Still classification performance is acceptable with ~30% of estimated error margin. For *E/P*, *E/N* and *N/P* scenarios VOC203, VOC2622 and VOC1913 were the top performers, respectively. The classification rules in Table 6.2.3 can be used for simplified classification of patients in a clinical setting without access to computational facilities.

### **6.2.6. Conclusions**

In this work, we were able to apply the unbiased conditional inference forests framework under feature selection and classification contexts achieving very good performances in the variable selection and binary classification of asthma subtypes. The limitations of our approach lie in rather small sample size and heterogeneous VOC data requiring additional validation study. The application of *CIF*-based classifier is recommended due to its spectacular performance in the classification context. Nevertheless, in future the simplified classifier based on individual top ranked 9 VOCs will be tested first under the context of the asthma subtype diagnosis. Future clinical studies will validate results with possible development of a diagnosis device.

### 6.2.7. Section highlights

In this section we applied the *CIF* method to difficult classification task. The VOC data is sparse creating extra difficulties in data analysis such as stringent quality control and model inference steps on a training part of data. Here, the *CIF* binary classification results were excellent. The multiclass classifications might be problematic given data limitations. The section described on how practically apply classification scenarios to clinical setting. Further validation studies in new patients are needed to verify our findings.

### 6.2.8. Acknowledgments and funding

This research was in part funded by the Fonds de la Recherche Scientifique (F.N.R.S.), in particular the project “Integration and interpretation of "omics" biological data via networks and conditional inference forests” [KB], and was carried out as part of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

### 6.3. *Regression2Net* - Integration of Gene Expression and Methylation to unravel biological networks in glioblastoma patients

#### 6.3.1. Section summary

In this work we describe *Regression2Net*, a computational approach that is able to integrate gene expression and genomic or methylome data. The *Regression2Net* is a two-step fusion integration method (see Section 2.5). First, penalized regressions are used to build Expression-Expression (*EE*) and Expression-Genome (*EG*) or –Methylome (*EM*) networks. Second, network theory is used to highlight important communities of genes.

**Problem:** A new strategy needs to be devised to integrate multiple omic layers resulting in a single GRN. The expression (*E*) and methylome (*M*) data layers are used to generate *EMnet* and *EEnet* networks. The integration step is done during assembly of the *EMnet* and *EEnet* into *ANDnet*, *XORnet* and *INTnet* networks. The biological significance and relevance of the three network fusion strategies (*AND*, *XOR* and *INT*) needs to be determined and validated in the context of glioblastoma.

**Results:** When applying our approach *Regression2Net* to gene expression and methylation profiles from individuals with glioblastoma multiforme (GBM), we identified 284 and 447 unique sets of candidate genes potentially associated with the glioblastoma pathology. In-depth biological analysis of these networks revealed genes that are related to energy metabolism, cell cycle control (*AATF*), immune system response and various cancer types. Importantly, we observed significant over-representation of cancer-related pathways including glioma, especially in the methylation network. This confirms the strong link between methylation and glioblastomas. We, furthermore, identified potential glioma suppressor genes *ACCN3* and *ACCN4* linked to *NBPFI* neuroblastoma breakpoint family in the expression network. Numerous ABC transporter genes (*ABCA1*, *ABCB1*) present in the expression network suggest drug resistance of glioblastoma tumors.

**Keywords:** penalized regression, lasso, glmnet, glioblastoma, transcriptome, methylome

### 6.3.2. Introduction

Glioblastomas are aggressive brain tumors affecting glial cells of the central nervous system including astrocytes and oligodendrocytes. The exact causes are not fully understood but current experimental evidence suggests that its onset is linked to mutations in gene *p53*, the essential cell cycle control protein and the neurofibromin 1 - *NF1*, inhibitor of RAS signaling pathway [7]. In oligodendrocyte tumors, the key marker *OLIG2* that regulates oligodendrocyte differentiation is not expressed [8]. Additional studies confirm that the key markers of glial cancers are related to nerve cell development BMP-BMPR/RAS- APK and PI3K- activated signaling pathways. The genetic inheritance component of glioblastomas is thought to be weak based on heterogeneity of genetic alternations of known disease markers amongst subjects [9]. This heterogeneity enormously complicates the development of effective therapies. However, exploiting the availability of high-throughput omics data, together with the development of novel computational data integrative analysis techniques, may facilitate formulating targeted biological and clinical hypotheses. These hypotheses potentially speed up research and improve early detection and diagnosis of glioblastomas in clinical settings.

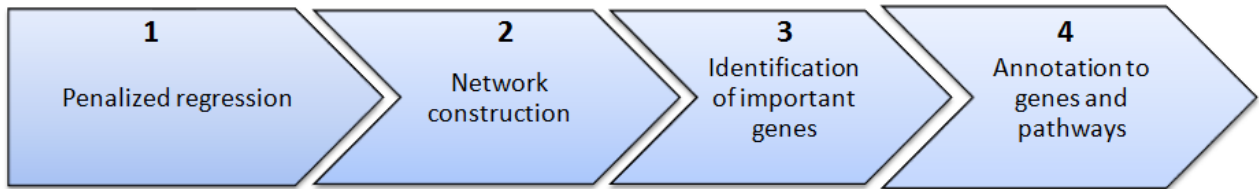
Several authors have indicated the added value of integrative omics analysis that involve the integration of at least two different omics data types, referring to different biological components in a cell. The methodologies to analyze such data are starting to emerge, with the biggest success stories reported for 2-omics analyses. Examples of 2-omics analyses include eQTL [10] and meQTL [11] analyses that respectively assess the influence of genetic and epigenetic markers on gene expression. Combining >2 omics data types is much more complex, given the hierarchical structure and interdependencies such data entails [12-14]. With a few exceptions, most methods integrate >2 different data sources by combining evidence obtained from pairwise analyses [14]. These evidence are often based on the derivation of standard measures of association, linking epigenetic markers to gene expression combined with gene expression analysis [15].

In this work, we develop a novel integration pipeline, *Regression2Net*, which combines information from for instance methylation and gene expression data via penalized regressions to construct gene-based networks. We, furthermore, applied our method to publicly available data for individuals affected by glioblastoma multiforme. The original glioblastoma pre-processed and

partially annotated data were obtained from [16]. By comparing the topology of the integrated networks derived from the inferred Expression-Expression and Expression-Methylation networks, and via subsequent functional analysis, we identified a list of genes with potential interest to the trait under consideration. In addition, our approach was capable of discovering biological mechanisms that may further enhance our understanding of the disease. Although experimental verification is needed to validate the novel formulated hypotheses in this work, our approach highlighted pathways that are significantly associated to glioblastoma. This suggests that our approach is one step forward towards translating *in silico* pieces of evidence to the patient bedside. In the next section, we motivate and provide details about the proposed integrative analysis pipeline *Regression2Net*.

### 6.3.3. Method

The *Regression2Net* uses a combination of penalized regression analysis and network theory to find structure in multi omics data. The main goal of *Regression2Net* consists in inferring gene network topologies and to derive meaningful key communities of disease-associated genes based on integrated data. To facilitate the explanation of our approach, we assume having transcriptome and epigenome data, collected on the same set of individuals. We split the entire analytic pipeline in four parts, which we describe below (Figure 6.3.1).



**Figure 6.3.1:** General workflow diagram of the *Regression2Net* methodology consisting of 4 stages.

**Part 1 – penalized regression.** Here, we consider each gene as response and its value is regressed (via penalized regression) against the remaining genes. Since both gene expression and methylation data may contribute to a gene measurement, we perform two types of regression: one in which only gene expression data are used, and one in which gene expression data (i.e. gene expression probes) are used at the response level and methylation data (i.e. methylation probes) as potential explanatory information. The regressions consecutively consider each gene’s expression as a dependent variable (response) and remaining gene expressions or methylation data as independent

variables, according to the type of regression. Mapping of probes to genes is based on genomic location of the probe to the nearest gene as in [16]. Note that, for instance, the number of methylation probes associated to a given gene may vary and hence multiple methylation probes may be considered as explanatory variables to gene expression. The strategy for variable selection and their “significance” assessment with penalized regression is based on principles outlined in [17] and was evaluated to synthetic data in [18]. In practice, *Regression2Net* currently leverages penalized linear regression with  $L1$ -norm penalty (Eq. 6.1). Given  $X_i$  the expression of gene  $i$  and the expression profiles of the remaining genes (referred to as  $X$ , for simplicity), the  $L1$ -norm penalized estimate consists of providing a solution for Eq. 6.1. The vector of regression coefficients  $\Theta$  determines the conditional independence structure among predictors.

$$\hat{\Theta}^{a,\lambda} = \underset{\text{s.t. } \Theta:\Theta_a=0}{\operatorname{argmin}} \left( \frac{1}{n} \|X_i - X\Theta\|_2^2 + \lambda \|\Theta\|_1 \right) \quad \text{Eq. 6.1}$$

One important feature of the  $L1$ -norm penalty consists in the tendency to shrink many coefficients to zero and to consequently remove them from the set of predictors  $X$ . This is an effective way to provide sparse solutions, which in turn lowers the variance of the selected regression coefficients. The variance that is lower than the one provided by non-penalized regression approaches is usually associated to the higher bias of the prediction, as explained in [19]. However, since our goal is to perform variable selection, we do not consider higher bias as a harmful limitation. It would be so if we were interested in predicting the expression value of the response genes. It is known that the crucial parameter that directly determines the rate of false positives and false negatives is the shrinkage factor  $\lambda$  in the Eq. 6.1. Regardless of a number of methods specifically designed to estimate  $\lambda$  reported in [20-22], we perform 10-fold cross validation on a subset of the dataset, which provides an optimal estimate of the shrinkage factor. At the end of the iterative procedure, two collections of genes (corresponding to the two types of regressions) with their “explanatory genes” are obtained.

**Part 2 – network construction.** All explanatory genes are subsequently connected to the genes they explain. In practice, all aforementioned connections are stored within an adjacency matrix  $A$ , the entries of which ( $A_{ij}$ ) being binary values 0/1 that show if gene  $i$  and gene  $j$  are connected or not,

thus obtaining two networks (one only using gene expression and one linking methylation to gene expression). The first network will be referred to as *EEnet* (Expression-Expression network) and the second to *EMnet* (Expression-Methylation network). In the presence of multiple probes per gene, two genes are connected in an *EMnet* (*EEnet*) network when at least one methylation (gene expression) probe is selected as significantly associated to the outcome ( $X_i$ ) by the regression algorithm.

**Part 3 - identification of important genes.** Here, the aim is to employ network-theory concepts to select the most important genes from the derived networks. These networks can be analyzed separately and results compared, or they can be integrated into a single combined network prior to analysis. A fundamental concept that needs to be clarified is the concept of “importance”. A simple procedure to select the most interesting (important) genes from a network is to consider its degree distribution. Note that the node degree or betweenness centrality of a specific node in a network is a local topological measure. These local topological network descriptions can be summarized into a global description of the network via the degree distribution  $p(k)$ . This distribution gives the proportion of nodes in the network having degree  $k$  (see Section 2.4.1). Therefore, a possible procedure to select “significant” genes would take into consideration highly connected genes in the *EEnet* and *EMnet* network, driven by the degree distributions of the respective networks. However, since degree correlations (i.e. dependency of two nodes being connected in the network on the node’s actual degrees) determine the actual network structure [23], *Regression2Net* adopts a different procedure.

Basically, aiming to increase the stability of important gene identification from different network resources (such as *EEnet* and *EMnet*), an integrated network is composed from its constituents *EEnet* and *EMnet*, prior to network interpretation. We do this in three possible ways, hereafter referred to as *ANDnet*, *XORnet* and *INTnet*. The edges in *ANDnet* are the edges that exist in both *EEnet* and *EMnet*. Here, *XORnet* is built by all the edges that are present in the *EMnet* but not in the *EEnet*. *INTnet* is a fused network of *EEnet* and *EMnet* using the approach of [16] adapted to gene-based adjacency matrices (with entries 0 and 1) underlying *EEnet* and *EMnet*, rather than similarity matrices between individuals (with numerical values which need to be normalized). In practice, un-normalized *EEnet* and *EMnet* networks are iteratively updated with information from

the other network, making them more similar after each step leading to a gradual shaping of the *INTnet* [16]. The finally fused network is converted to an adjacency matrix by replacing strictly positive matrix entries with 1. Once an integrated network is obtained via *ANDnet*, *XORnet* or *INTnet*, we select all connected genes (i.e. with node degree  $\geq 1$ ) from each network, giving rise to three lists of unique genes that enter the last part of the integrative analysis methodology.

**Part 4 - annotation protocol and pathway enrichment analysis.** In order to assess the significance of the selected genes in Part 3 in relation to the disease of interest, we perform an annotation and pathway enrichment analysis, supplemented by literature searches. In practice, we use the R package *biomaRt* (R version 2.20.0) to annotate genes from gene expression data and the R packages *GGHumanMethCancerPanelv1.db* to annotate genes from methylation panels. The selected annotation criteria include gene full name, chromosome name, ensemble gene and transcript IDs. KEGG pathway enrichment analysis is performed with the R package *KEGGprofile* [24] on non-overlapping genes from the unique gene lists derived from *ANDnet* and *XORnet* networks. The minimum threshold to accept a significant pathway is set to  $p\text{-value} < 0.05$  and is computed from a hypergeometric distribution for testing whether a pathway is over-represented in our gene list, compared to KEGG. Reported  $p$ -values are Bonferroni-corrected to deal with multiple testing.

The *Regression2Net* code is written in R is freely available at <https://bitbucket.org/kbessonov/regression2net/> and via <http://www.statgen.ulg.ac.be>

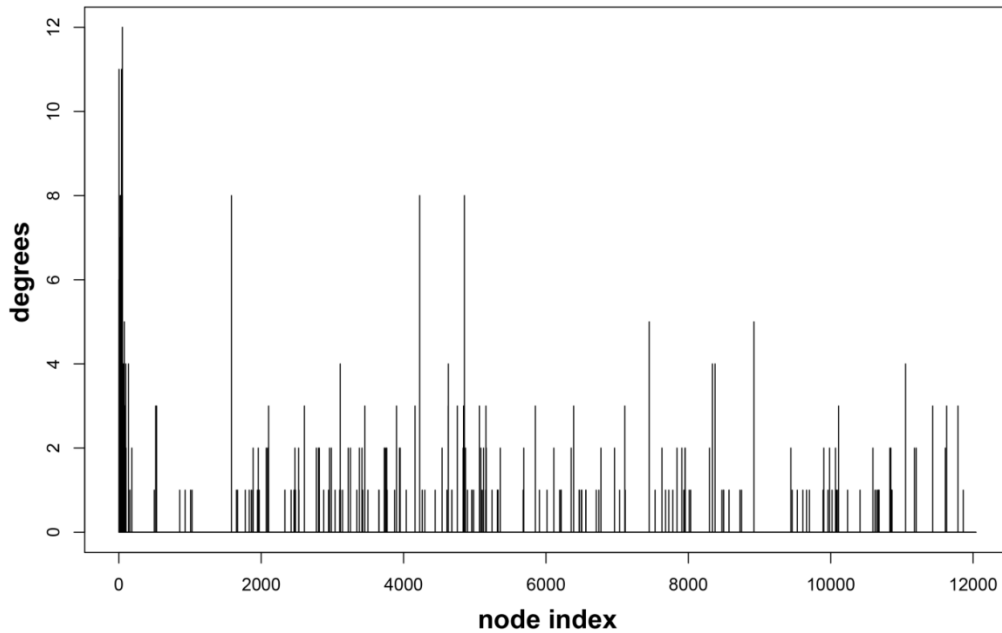
### 6.3.4. Data

The data at our disposal are heterogeneous datasets composed of gene expression and methylation profiles of 215 individuals affected by glioblastoma already considered in a study of patient similarity in [16]. DNA methylation probes (in total 1305) and mRNA (in total 12042) probes covered 680 and 12,042 genes, respectively. For more details about the platforms that generated the data, we refer to [16].

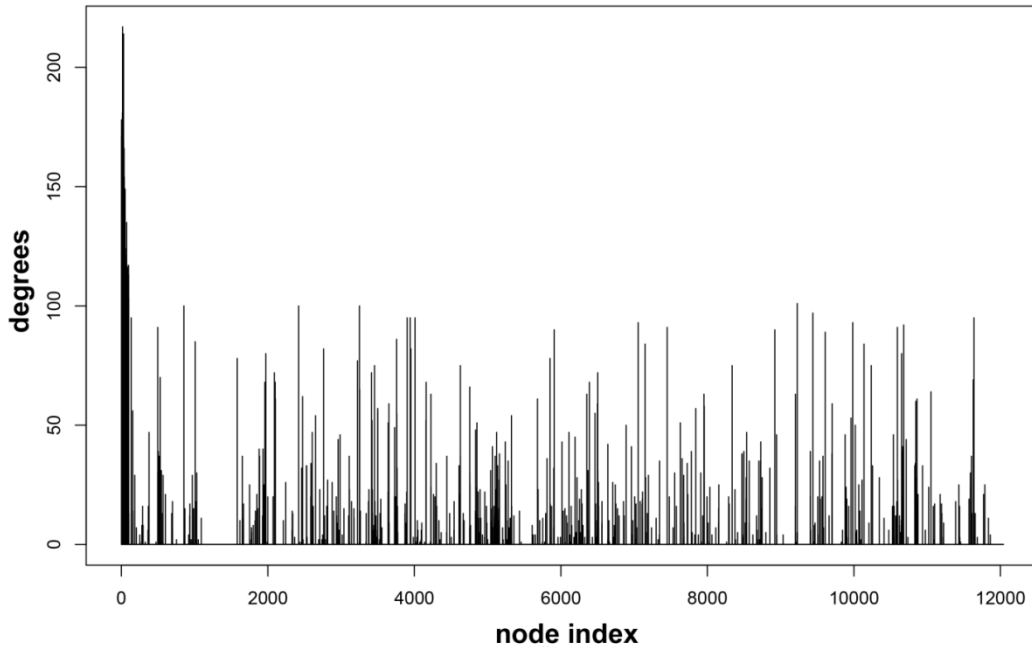
### 6.3.5. Results

#### 6.3.5.1. Data and network characteristics

The number of methylation probes associated with a given gene ranged between 0 and 4. These methylation probes are not uniformly distributed across the genome mapped to only 680 genes. We found a strong predominance of methylation probes located in the 5'UTR regions of the genes according to the Golden Gate Human Methylation Cancer Panel 1 [25] – Gene expressions were in 1-1 correspondence to gene expression probes. Basic topology analysis of the *ANDnet* and *XORnet* networks (see Method section) developed on real-life data for glioblastoma multiforme showed degree distributions in line with scale-free networks (Figure 6.3.2 and Figure 6.3.3). No distinct pattern between degrees and the relative number of methylation probes per gene has been found. From the *ANDnet* and *XORnet* networks, we identified 284 and 730 probes with node degree  $\geq 1$ , respectively. After gene mapping, this resulted in respective unique gene lists of length 284 and 447. These gene lists were submitted to in-depth biological analyses.



**Figure 6.3.2:** Total degree distribution of the *ANDnet* network of edges present in both *EEnet* and *EMnet*, Node index represents the node number



**Figure 6.3.3:** Total degree distribution of the *XORnet* network of edges present in *EMnet* but not in *EEnet*

### 6.3.5.2. Annotation of the *ANDnet* and *XORnet* unique gene lists

The aforementioned 284 and 447 unique gene lists were annotated to biological functions and pathways in order to provide biological context in relation to glioblastoma pathology. Amongst the 284 *ANDnet* genes (online Table S6.3.1) are the two Amiloride-Sensitive Brain Sodium Channels encoded by *ACCN3* and *ACCN4*. These genes were shown to be linked to the neuroblastoma breakpoint family *NBPFI* genes related to the development of glioblastoma [26]. The *NBPFI* genes are thought to be involved in brain development and the neuroblastoma onset [27]. When looking for the presence of transcription factors (TFs) amongst the genes of the *ANDnet* network, we noticed *AATF* and *ABT1*. They play an important role in the context of glioblastoma due to the fact that gene *AATF* controls crucial apoptotic cell death processes and gene *ABT1* is responsible for basal transcription control via interaction with class II promoter sequences and onset of schizophrenia [28,29]. Genes belonging to the ATP-binding cassette (ABC) are numerous in the *ANDnet* network. These transporter proteins are often involved in drug resistance [30]. Their strong presence amongst *ANDnet* genes suggests a complex gene regulatory mechanism that involves synergetic methylation and expression components. The complex regulation of the *ABC* genes has been confirmed in [26]. Overall, the *ANDnet* network is mainly composed of genes related to

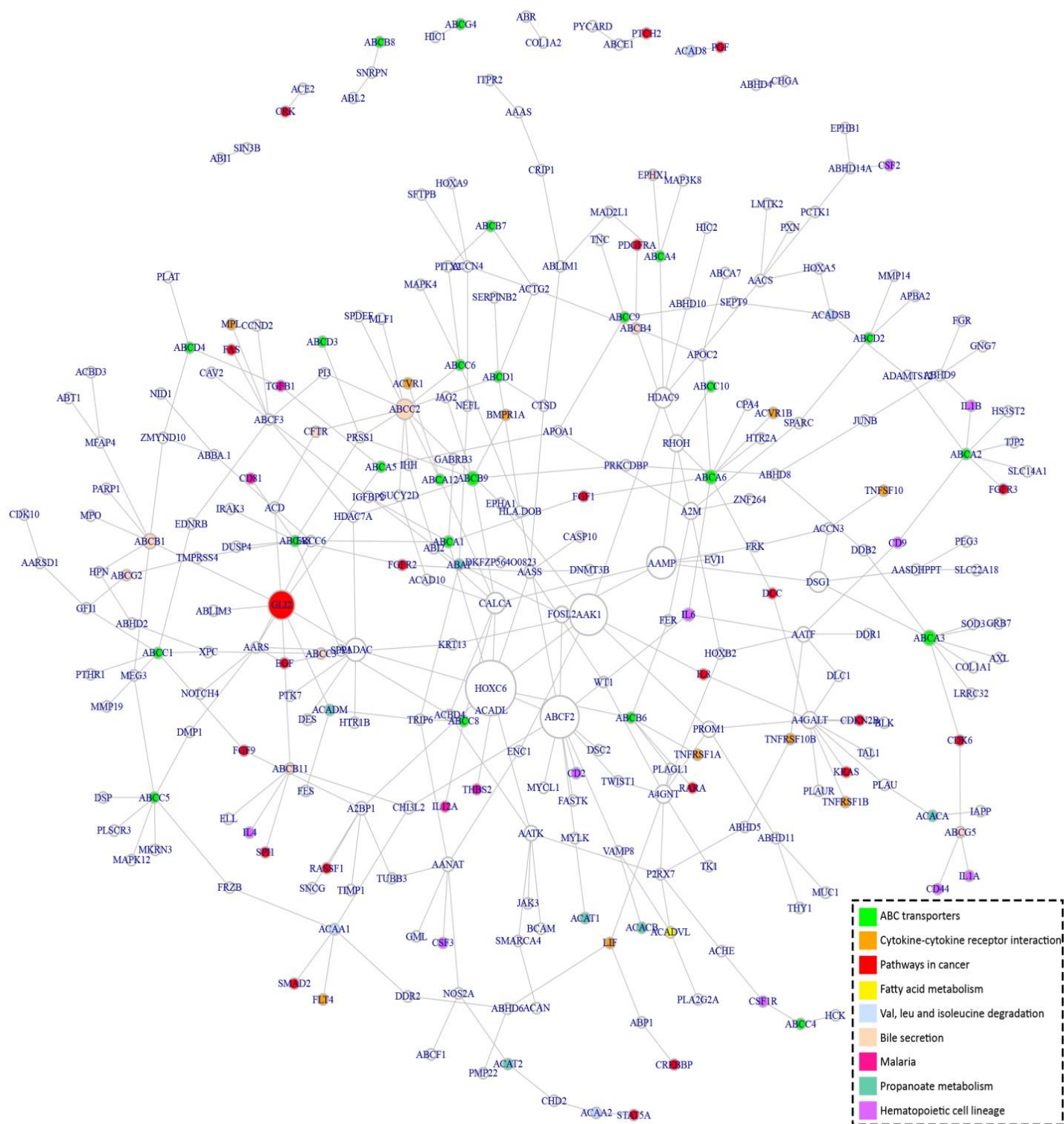
energy metabolism, while the *XORnet* network is formed by genes related to various cancer pathways, cell cycle control and immune system responses (online Table S6.3.2).

### 6.3.5.3. Enrichment analysis

The results of the KEGG pathway enrichment analyses are reported in Table 6.3.1, Table 6.3.2 and Table 6.3.3. Functional and pathway analyses of the 284 *ANDnet* genes revealed significant pathway enrichment in cancer-related genes, energy metabolism, ATP-binding membrane transporters, transcription regulation, cell cycle control proteins and other biological functions (Table 6.3.1 and Figure 6.3.4). Energy metabolism and ABC transporters genes are only significant in the *ANDnet* network. This shows that these biological processes can have both expression and methylation regulatory components [31]. Pathway analysis identified an important Glioma pathway (KEGG:hsa05214) enriched only in the *XORnet* network. The metabolic pathway in cancer (KEGG:hsa05200) is enriched in both the *XORnet* and *ANDnet* networks (Table 6.3.2). The following genes, exclusively present in the *XORnet*, are linked to KEGG:hsa05200: *AXIN1* - axin 1, *FGF7* - fibroblast growth factor 7, *FZD9* - frizzled class receptor 9, *NKX3-1* - NK3 homeobox 1, and *TGFB1* transforming growth factor, beta 1.






The relevance of some genes belonging to this pathway is supported by literature, specifically regarding gene *NKX3-1*, which is known to be implicated in prostate cancer development in adult mice [32], and gene *FGF7*, implicated in brain tumors [33].

In addition, the pathways identified in *INTnet* are most similar to those identified in *XORnet* (Table 6.3.2). Common pathways relate to cancer (KEGG:hsa05200) and glioma (KEGG:hsa05214).








**Figure 6.3.4:** *ANDnet* network overlap with the significant pathways. The highlighted genes belong to the significant pathways indicated in Table 6.3.1 while non-highlighted (white) genes have not been linked to any significant pathway. The size of each node is determined by betweenness, defined as the number of shortest paths going through the node

**Table 6.3.1:** *ANDnet* enriched pathways

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
<i>EEnet</i>						
<i>EMnet</i>						
<i>ANDnet</i>						
<b>KEGG pathway</b>			<b>gene ratio</b>	<b><i>p</i>-value</b>		
ABC transporters			32/44	8.45E-41		
Cytokine-cytokine receptor interaction			23/265	6.77E-05		
Pathways in cancer			25/327	2.17E-04		
Fatty acid metabolism			8/43	9.13E-04		
Valine, leucine and isoleucine degradation			8/44	1.12E-03		
Bile secretion			10/71	1.50E-03		
Malaria			8/51	4.01E-03		
Propanoate metabolism			6/32	9.43E-03		
Hematopoietic cell lineage			10/88	1.20E-02		















**Table 6.3.2:** *XORnet* enriched pathways

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
<i>EEnet</i>						
<i>EMnet</i>						
<i>XORnet</i>						
<b>KEGG pathway</b>			<b>gene ratio</b>	<b><i>p</i>-value</b>		
Pathways in cancer			79/327	2.52E-37		
Melanoma			22/71	1.05E-11		
Prostate cancer			22/89	2.06E-09		
Focal adhesion			33/200	3.79E-09		
Colorectal cancer			16/62	5.24E-07		
Bladder cancer			13/42	9.49E-07		
Toxoplasmosis			23/133	1.80E-06		

---

Cytokine-cytokine receptor interaction	34/265	1.81E-06
Small cell lung cancer	17/85	1.40E-05
Endometrial cancer	13/52	2.05E-05
p53 signaling pathway	15/69	2.19E-05
T cell receptor signaling pathway	19/108	2.48E-05
Regulation of actin cytoskeleton	28/214	2.80E-05
Basal cell carcinoma	13/55	4.42E-05
Chagas disease (American trypanosomiasis)	18/104	6.76E-05
ErbB signaling pathway	16/87	1.16E-04
Fc epsilon RI signaling pathway	15/79	1.59E-04
Pancreatic cancer	14/70	1.73E-04
MAPK signaling pathway	30/268	3.03E-04
Chronic myeloid leukemia	14/73	3.06E-04
Leishmaniasis	14/73	3.06E-04
Glioma	13/65	4.00E-04
Hedgehog signaling pathway	12/56	4.00E-04
Acute myeloid leukemia	12/58	6.13E-04
Renal cell carcinoma	13/70	1.02E-03
Adherens junction	13/73	1.71E-03
Cell cycle	18/128	1.72E-03
Axon guidance	18/130	2.16E-03
Hematopoietic cell lineage	14/88	3.47E-03
Toll-like receptor signaling pathway	15/102	5.11E-03
Malaria	10/51	5.93E-03
Neurotrophin signaling pathway	17/127	5.96E-03
Osteoclast differentiation	17/128	6.64E-03
Non-small cell lung cancer	10/54	1.05E-02
Natural killer cell mediated cytotoxicity	17/136	1.51E-02
Wnt signaling pathway	18/151	1.82E-02
Rheumatoid arthritis	13/92	2.56E-02
Leukocyte transendothelial migration	15/117	2.83E-02

**Table 6.3.3:** *INTnet* enriched pathways

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
<i>EE</i> net						
<i>EM</i> net						
<i>INT</i> net						
KEGG pathway	gene ratio		p-value			
Pathways in cancer	109/327		2.22E-22			
ABC transporters	34/44		3.15E-21			
Hematopoietic cell lineage	37/88		4.11E-10			
Cytokine-cytokine receptor interaction	69/265		9.94E-08			
Chagas disease (American trypanosomiasis)	36/104		6.23E-07			
Bladder cancer	20/42		2.24E-06			
p53 signaling pathway	27/69		2.37E-06			
Melanoma	27/71		5.02E-06			
Osteoclast differentiation	39/128		8.93E-06			
Focal adhesion	52/200		1.94E-05			
T cell receptor signaling pathway	34/108		2.63E-05			
Prostate cancer	29/89		9.18E-05			
Malaria	20/51		1.48E-04			
Pancreatic cancer	24/70		2.84E-04			
Toll-like receptor signaling pathway	29/102		2.22E-03			
Chronic myeloid leukemia	23/73		2.51E-03			
Leishmaniasis	23/73		2.51E-03			
MAPK signaling pathway	57/268		4.53E-03			
Non-small cell lung cancer	18/54		8.04E-03			
Fc epsilon RI signaling pathway	23/79		1.11E-02			
Regulation of actin cytoskeleton	47/214		1.19E-02			
Glioma	20/65		1.29E-02			
Toxoplasmosis	33/133		1.37E-02			
Cell cycle	31/128		3.72E-02			
Leukocyte transendothelial migration	29/117		3.83E-02			
B cell receptor signaling pathway	21/75		4.28E-02			

### 6.3.6. Discussion

We combined *EEnet* and *EMnet* into integrated networks: either *ANDnet*, *XORnet* or *INTnet*. The motivation to derive *XOR* type of networks is our belief that expression-based gene-gene interactions may be quite different from methylation-based gene-gene interactions, which may be of interest on its own. *INTnet* provides a more elaborate way of integrating *EEnet* and *EMnet* information based on the non-linear combination method of [16]. In our application, the *INTnet* approach gave rise to the same adjacency matrix as we would obtain by connecting two genes whenever a connection was present in the *EEnet* or in the *EMnet*. The highest number of significantly enriched pathways was identified with *XORnet* (Table 6.3.2), indicating the importance of methylation regulation in glioblastoma. Notably, whatever integrative network approach was followed, the quality of the integrated network would depend on the quality of the constituent networks.

Currently, in *Regression2Net*, mapping results for methylation probes to genes does not account for gene length, neither for the number of methylation probes in a gene. In the presence of multiple methylation probes per gene, two genes are connected in the *EMnet* network when at least one methylation probe has been selected by penalized regression. We consider this an acceptable strategy due to the fact that the mapped Expression-Methylation network (*EMnet*) of interest in our strategy is an unweighted one. Potentially, larger genes may have increased chances to get connected in a network, as those genes include a higher number of methylation probes.

In total, we identified a total of 10 pathways amongst the 38 *XORnet* KEGG pathways linked to cancer. The most interesting functional links amongst candidate genes of the *XORnet* network are those between genes *NCAM* – neural cell adhesion molecule and *FGF7* - fibroblast growth factor 7. The *FGF* competes with *NCAM* for FGF receptor binding (*FGFR*) that results in alteration of *FGFR* signalling [33]. Aberration in the expression levels of gene *NCAM* and excessive *FGFR* signalling have been shown to be correlated with tumor onset [34,35]. Thus FGF family of proteins also plays an important role in neurological disorders through alteration of *FGFR* signalling. In addition, the transcription regulation functions of the *XORnet* network are represented by transcription factors *FOSL2* and *SIN3B* involved in cell proliferation and other oncogenic activities (Table S6.3.2). We have shown that there is an added value in constructing and functionally

analyzing both *XORnet* and *ANDnet* networks, since they may give complementary information. A possible interpretation of the topology of the *ANDnet* network is that the 284 genes in the *ANDnet* network (Table S6.3.1) are controlled by both an expression and methylation component (two connected genes in *ANDnet* are by definition connected in both the *EEnet* and *EMnet* networks).

Importantly, a total of 25 out of 284 genes of the *ANDnet* network and 79 out of 447 genes of the *XORnet* network were linked to the KEGG metabolic pathway in cancer (KEGG:hsa05200). We consider this to be a significant result as it suggests that glioblastoma cancers seem to be strongly linked to the methylation component, which in turn perturbs the expression component [36]. This is directly reflected within the topology of the *EEnet* and *EMnet* networks. Enrichment analysis of the 447 genes of the *XORnet* network (genes with degree  $\geq 1$ ) showed consistent presence of pathways related to cancer and biological processes including various types of carcinomas, cell signalling and immune system responses. This supports evidence that cancers have a very strong methylation component, confirmed by many studies [31,36,37]. Also, KEGG pathway enrichment analysis performed on the genes of the *ANDnet* network identified 4 pathways that are common to both the *ANDnet* and *XORnet* network, including cytokine-cytokine receptor interaction, metabolic pathways in cancer, malaria and haematopoietic cell lineage pathways. This suggests that genes in these subsets may display highly complex regulations.

### 6.3.7. Conclusions

In this work we have described a computational method and integration methodology based on penalized regression and graph theory: *Regression2Net*. Using data on genome-wide gene expression and methylation we applied our method in the context of glioblastoma pathology. We biologically validated our findings by means of annotations, pathway enrichment analysis and literature searches. Our integrative analysis methodology, which includes the construction of *XORnet* and *ANDnet* networks, highlighted the added value of network integration prior to functional analysis. We were able to confirm the strong methylation component in glioblastoma pathologies. The evidence provided by our findings, supported by the literature, strongly suggests the potentials of our proposed strategy.

### 6.3.8. Section highlights

The *Regression2Net* method was successfully applied to real-life data. A total of 3 strategies were used including *AND*, *XOR* and *INT*. Functional analysis of the inferred networks showed strong links to cancer and cell cycle control pathways. The shortcoming of *Regression2Net* strategy is that it builds an unweighted GRN without assessment of statistical significance of network edges between gene nodes.

### 6.3.9. Acknowledgments and funding

This research was in part funded by the Fonds de la Recherche Scientifique (F.N.R.S.), in particular “Forestry in Integromics Inference” (Convention n° T.0180.13) [FG, KB, KVS] and the project “Integration and interpretation of "omics" biological data via networks and conditional inference forests” [KB], and was carried out as part of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

### 6.3.10. Appendix

**Table S6.3.1:** *ANDnet* 284 gene annotations. (See the online supplement).

**Table S6.3.2:** *XORnet* 447 gene annotations. (See the online supplement).

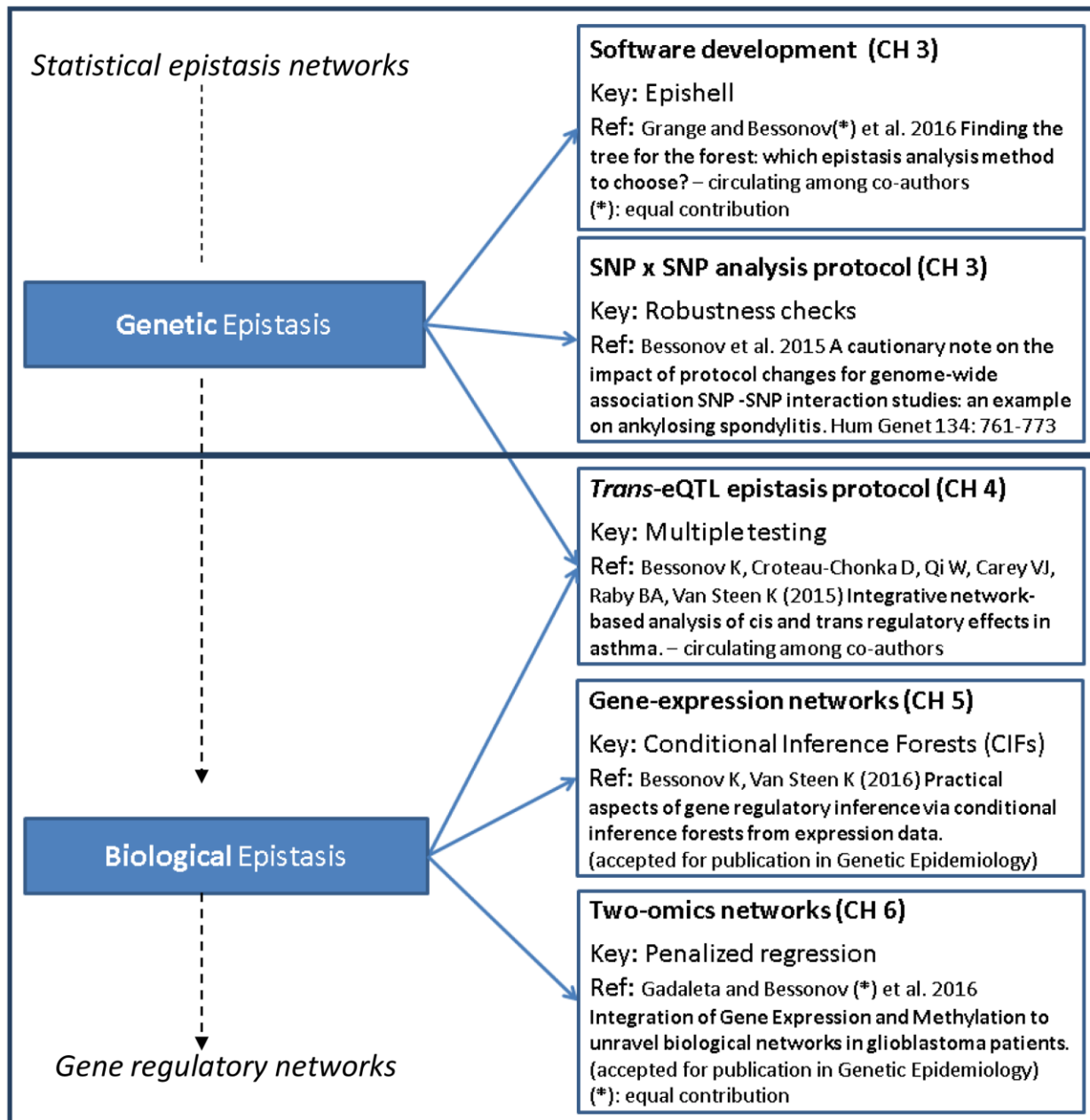
### 6.3.11. References

1. Hetherington KJ, Heaney LG (2015) **Drug therapies in severe asthma—the era of stratified medicine.** *Clinical Medicine* 15: 452-456.
2. Louis R, Godinas L, Schleich F (2011) **Induced Sputum-Towards Normal Values.** *Non Invasive Assessment of airways inflammation in asthma and COPD*: 113-123.
3. Pavord ID, Brightling CE, Woltmann G, Wardlaw AJ (1999) **Non-eosinophilic corticosteroid unresponsive asthma.** *Lancet* 353: 2213-2214.
4. Haldar P, Brightling CE, Hargadon B, Gupta S, Monteiro W, et al. (2009) **Mepolizumab and exacerbations of refractory eosinophilic asthma.** *N Engl J Med* 360: 973-984.
5. Hothorn T, Hornik K, Strobl C, Zeileis A (2010) **Party: A laboratory for recursive partytioning.**

6. Hothorn T, Hornik K, Strobl C, Zeileis A, Hothorn MT (2014) **Package ‘party’**. *Package Reference Manual for Party Version 09-998* 16: 37.
7. Zhu Y, Guignard F, Zhao D, Liu L, Burns DK, et al. (2005) **Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma**. *Cancer cell* 8: 119-130.
8. Marie Y, Sanson M, Mokhtari K, Leuraud P, Kujas M, et al. (2001) **OLIG2 as a specific marker of oligodendroglial tumour cells**. *The Lancet* 358: 298-300.
9. Kraus JA, Lamszus K, Glesmann N, Beck M, Wolter M, et al. (2001) **Molecular genetic alterations in glioblastomas with oligodendroglial component**. *Acta neuropathologica* 101: 311-320.
10. Franke L, Jansen RC (2009) **eQTL analysis in humans**. *Methods Mol Biol* 573: 311-328.
11. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, et al. (2014) **Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type**. *BMC Genomics* 15: 145.
12. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, et al. (2009) **Data integration in genetics and genomics: methods and challenges**. *Hum Genomics Proteomics* 2009.
13. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) **Methods of integrating data to uncover genotype-phenotype interactions**. *Nat Rev Genet* 16: 85-97.
14. Van Steen K, Malats N (2014) **Perspectives on Data Integration in Human Complex Disease Analysis**. In: Wang B, Li R, Perrizo W, editors. *Big Data Analytics in Bioinformatics and Healthcare*. 1 ed: IGI Global. pp. 284-322.
15. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, et al. (2014) **The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts**. *Genome Biol* 15: R37.
16. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. (2014) **Similarity network fusion for aggregating data types on a genomic scale**. *Nat Methods* 11: 333-337.
17. Meinshausen N, Bühlmann P (2006) **High-dimensional graphs and variable selection with the lasso**. *The Annals of Statistics*: 1436-1462.
18. Gadaleta F, Van Steen K (2014) **Discovering Main Genetic Interactions with LABNet Lasso-Based Network Inference**.
19. Tibshirani R (1996) **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society Series B (Methodological)*: 267-288.
20. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) **Least angle regression**. *The Annals of statistics* 32: 407-499.
21. Zou H (2006) **The adaptive lasso and its oracle properties**. *Journal of the American statistical association* 101: 1418-1429.
22. Hirose K, Tateishi S, Konishi S (2013) **Tuning parameter selection in sparse regression modeling**. *Computational Statistics & Data Analysis* 59: 28-40.
23. Hao D, Li C (2011) **The dichotomy in degree correlation of biological networks**. *PLoS One* 6: e28322.
24. Zhao S (2012) **KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway**. *R package version 1*.
25. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, et al. (2006) **High-throughput DNA methylation profiling using universal bead arrays**. *Genome research* 16: 383-393.
26. Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, et al. (2008) **A constitutional translocation t (1; 17)(p36. 2; q11. 2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes**. *PloS one* 3: e2207.
27. Janoueix-Lerosey I, Schleiermacher G, Delattre O (2010) **Molecular pathogenesis of peripheral neuroblastic tumors**. *Oncogene* 29: 1566-1579.
28. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, et al. (2009) **Common variants on chromosome 6p22. 1 are associated with schizophrenia**. *Nature* 460: 753-757.

- 
29. Gejman PV, Sanders AR, Duan J (2010) **The role of genetics in the etiology of schizophrenia.** *Psychiatric Clinics of North America* 33: 35-66.
  30. Lage H (2003) **ABC-transporters: implications on drug resistance from microorganisms to human cancers.** *International journal of antimicrobial agents* 22: 188-199.
  31. Phillips T (2008) **The role of methylation in gene expression.** *Nature Education* 1: 116.
  32. Abdulkadir SA, Magee JA, Peters TJ, Kaleem Z, Naughton CK, et al. (2002) **Conditional loss of Nkx3.1 in adult mice induces prostatic intraepithelial neoplasia.** *Molecular and cellular biology* 22: 1495-1503.
  33. Francavilla C, Loeffler S, Piccini D, Kren A, Christofori G, et al. (2007) **Neural cell adhesion molecule regulates the cellular response to fibroblast growth factor.** *Journal of cell science* 120: 4388-4394.
  34. Grose R, Dickson C (2005) **Fibroblast growth factor signaling in tumorigenesis.** *Cytokine & growth factor reviews* 16: 179-186.
  35. Vawter MP (2000) **Dysregulation of the neural cell adhesion molecule and neuropsychiatric disorders.** *European journal of pharmacology* 405: 385-395.
  36. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, et al. (2008) **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 455: 1061-1068.
  37. Esteller M (2005) **Aberrant DNA methylation as a cancer-inducing mechanism.** *Annu Rev Pharmacol Toxicol* 45: 629-656.

# Chapter 7: General Discussion





## 7. General Discussion and Future Perspectives

### 7.1. Introduction

The main theme of this thesis work is detection of biological interactions in multi omics-data via the development of integration methodologies including epistatic *trans/cis* eQTL MB-MDR, *CIF<sub>mean</sub>*, and *Regression2Net* covered in Chapters 3, 4, and 6. The above mentioned interaction mining methods aim to extract “useful knowledge” from multiple omics data sources – part of the Big Data to Knowledge (BD2K) process introduced in Section 1.1. Due to the complexity and broadness of covered topics, this thesis work in the network medicine and omics integration fields only addressed the ‘tip of iceberg’. Big interest in the data integration and network medicine fields [1] promises to fill the remaining knowledge voids in future and provide solutions such as personalized medicinal care solutions [2]. The later sections will link our main achievements throughout this thesis work in the context of the set aims introduced in Section 1.6. We start our discussion from GWAIS detecting statistical epistatic interactions between genotypes and phenotypes (Figure 1.4) transitioning to other types of interactions in the context of omics data integration.

### 7.2. Optimal GWAIS protocol

The results of Chapter 3 showed the importance of proper GWAIS protocol tuning and assessment of each protocol parameter. Due to significant variation in results we recommend a thorough description of parameters used in every GWAIS. It remains to be seen if the presented results generalize well to other datasets in the context of complex diseases as the heterogeneity (e.g., different sub-populations, batch effects, etc.) might have a significant impact on a final results variation. The dendrogram (Figure 3.3) turned out to be a very useful tool to show relative distances between GWAIS protocols graphically. Compared to original publication [3], this thesis provided additional results on MB-MDR application to genome-wide setting via *gammaMAXT* algorithm [4]. Unfortunately, due to high computational requirements of the classical exhaustive *MAXT* algorithm, we were not able to compare obtained *gammaMAXT* results with the *MAXT* ones on the genome-wide scale.

The main objective of Chapter 3 was a further evaluation of GWAIS protocol [5] via a case study on AS. The relative ranking of parameters based on their impacts on final results is very specific and can be easily used as the checklist in any GWAIS. The dataset size with and without prior variable pre-selection was identified as the most impactful factor (compare protocols #1 and #3, #5 and #9 of Figure 3.3). This finding is not surprising as datasets with a lower number of variables have a lower search space and, thus, narrower rank ranges resulting in lower overall distances between SNP pairs. We saw smaller overall distances amongst the GWAIS protocols relying on pre-filtered data as compared to the exhaustive ones (Figure 3.3). The reduction of the original dataset and variable pre-selection has a strong potential pitfall – removal of relevant variables limiting the discovery of novel epistatic interactions. Thus, variable selection should be ideally based on a well-founded hypothesis embedded in sophisticated algorithms. To this end, trees intrinsically implement recursive data partitioning dividing original data into subsets. The data structure of these subsets can be locally explored and relevant variables selected. The tree-based algorithms such as *RF* and *CIF* are well suited to explore interactions in omics datasets with large number of interactions as stated in [6]. Application of tree-based methods for variable selection to an existing GWAI protocol presented in Chapter 3 is promising, but would require a careful tuning on simulated or gold standard data. At present due to incompleteness of human interactome maps and lack of wet-lab empirical data hinders development of reliable gold standards [7].

The LD pruning effects on the final results in GWAS and GWAIS is an area of active research. The correction patterns of genotype frequencies represented by LD can give mild to significant impacts as was partially shown in Chapter 3. In case BOOST-like protocols relying on logistic regression framework, the impacts can be quite significant. Thus, incorporation of LD structures into epistasis detection algorithms is a desirable feature and helps identification of additional causal indirectly linked SNPs [8,9]. We found very few studies exploring the impact of LD on result reliability of the GWAS and GWAIS meriting further exploration. Further developments should consider incorporation of markers LD structure and development of a reliable gold standard allowing fine-tuning of each epistasis detection methodology. The LD structure is also of concern in eQTL studies [10,11] that link genomics and transcriptomic data. Elevated LD between a pair of markers might lead to false positives as discussed in the subsequent section.

### 7.3. *Trans*-eQTL epistasis protocol for eQTL detection

GWAIS considered in the previous section only assess the impact of the genotypic component (SNP) on a phenotypic trait (e.g., disease status). The eQTLs consider the impact of genotype (DNA) on gene expression levels (RNA). In the context of GWAS and GWAIS, eQTLs represent an intermediate layer that can potentially impact phenotype providing visual clue as to DNA-RNA interaction. Integrative eQTL studies covering both gene-gene interactions and phenotypic/functional components are becoming more widespread [12-14]. This allows to better account for the complexity of biological systems as the relationships between omics information layers are more complex than currently appreciated [15].

Another major issue of most eQTL studies is that they treat eQTL loci as totally independent entities failing to incorporate possible correlation structure (i.e. LD) existing between genetic markers [16]. Under this independence scenario, only additive and cumulative loci effects are taken into account that can miss a large number of important gene-gene interactions and loci with small marginal (i.e. main) effects [16]. Most of the current eQTL studies use linear regression to detect associations between genotype and expression. As explained in Section 1.3, the assumption of linear associations under biological context is rather primitive and can potentially ignore complex interaction patterns.

Gene expression intensity levels may be correlated due to other confounding factors, as apparent from gene co-expression networks [17]. Co-expression between expression levels of genes tends to correlate with common biological function and is important to account for in eQTL study in addition to LD patterns. Most of epistatic eQTL studies use models that consider expression traits separately assuming no correlation. In Chapter 4 we presented epistatic *trans/cis* eQTL MB-MDR method that also did not take into account correlations between the *cis* eQTL genes as it was very low (see Figure 4.7). Future eQTL studies should take into account both correlation structures existing between expression traits and between loci pairs as it is another source of false positives. For example, new approaches include eQTLs in the construction of weighted co-expression networks such as in [18]. To increase the accuracy of eQTL mappings one can resort to fine-mapping studies that offer a more detailed analysis and can verify previous eQTL mappings while

providing additional insights. For example, a fine-mapping study of importin 8 gene (IPO8) in human liver tissue samples identified several factors affecting eQTLs detection including mutations occurring within the 3'UTR region [19]. One simple naïve way is to assign different weights to interaction effects between eQTL loci based on corresponding LD value. Yet another way is to incorporate LD into an eQTL model as a covariance matrix. This was addressed by Guseva *et al.* [20] which designed an imputation-based transcriptome-wide association study (TWAS) that measured expression-trait associations based on *cis* eQTLs.

An open-ended question, especially in epistatic eQTL studies, is sensible significance assessment and proper multiple testing correction. In case of our MB-MDR based epistatic eQTL method presented in Chapter 4, the default significance level of 0.05 (e.g.,  $p$ -value  $< 0.05$ ) was adequate even at an extremely large interactome search space. The FWER rates were less or equal to 0.05 (see Section 4.4.2). The low correlation between *cis* eQTL genes might have kept FWER rates at acceptable levels with the need of additional multiple testing corrections. Still owed to the presence of other unknown confounder factors, the selected significance *trans/cis* epistatic eQTL threshold at  $p$ -value  $< 0.05$  might have provided overly optimistic results. Future releases of the *trans/cis* eQTL MB-MDR protocol should account for a greater number of confounding factors such as LD, gene expression correlation and missing/swapped data. In other words, selected  $p$ -value threshold might not be adequate at scenarios with more pronounced correlation structure. Another important issue with eQTL studies is a proper multiple testing correction especially in the case of epistatic eQTLs as discussed in Section 4.5 due to the presence of a huge number of genetic markers and expression traits. To this end, various strategies are introduced in Section 4.2 ranging from re-sampling [21] to step-wise selection of relevant markers [22]. One of the easiest and most obvious ways to correct for multiple testing is a Bonferroni's method but, in agreement with other studies, this method is too conservative [23]. Indeed, after application of Bonferroni correction to our *trans/cis* eQTL results only 3 SNP pairs were identified listed in Table S4.6. Another way to cope with multiple testing is to decrease the total number of hypotheses via calculation of the effective number of tests ( $M_{\text{eff}}$ ) as per Nyholt's method [24]. However, this approach has been shown to be susceptible to correlation structures between SNPs (i.e. LD) [23]. As discussed in Section 4.5, the adopted  $M_{\text{eff}}$  method by [25] based on utilization of eigenvalues of a correlation matrix takes advantage of correlations between expression traits. Nevertheless, the low number of highly

correlated expression traits did not provide any improvement over classical Bonferroni correction. Since MB-MDR adequately corrects for multiple testing as demonstrated by  $\text{FWER}_{\text{within}}$  (Figure 4.5) and low correlation patterns are seen between expression traits, no further corrections were necessary. Still, development of novel approaches that both take into account correlation structures present in genotype and expression data is still needed. The most promising method is mutual information based by Szymczak *et. al* [26]. This method uses mutual information measure which intrinsically incorporates correlation notion existing between variables assuming that one random variable contains information about the other. Also, in the Szymczak *et. al* method the significance of final association scores is based on  $10^5$  permutations. It still remains to be seen if Szymczak *et. al* [26] method can maintain adequate error rates under epistatic eQTL context involving interaction of SNP pairs possibly dependent on LD patterns.

Translation of computational predictions to clinical setting is very compelling. It is highly desirable to validate eQTL results experimentally via targeted point mutations of a selected loci and quantification of expression variations. The need of reliable experimental wet-lab eQTL validation protocols is especially relevant given that most detected eQTLs lack molecular mechanism(s) of association (i.e. biological epistasis). Integrative eQTL studies are more frequent nowadays pooling evidence from various sources. For example, several studies exist that integrate eQTL and PPI data [27,28] , eQTL and phenotypic data [20].

In Chapter 4 we presented a method that takes into account joint interaction effects of *cis* and *trans* loci. The proposed model better reflects a biological reality characterized by complex multi-level regulatory programs with underlying cross-talk between cell functional programs (e.g., cell cycle) [29]. The statistical framework of the method does not assume linear associations between genotype and expression and is semi-parametric with the inherent support of joint effects between several loci. The proposed MB-MDR method coupled to gene network analysis proved to be theoretically and practically sound as it is based on biologically relevant hypotheses combining benefits of statistical and network analyses. In addition, to our knowledge, this is the first study that showcased MB-MDR under the *trans/cis* eQTL epistatic context on real-life complex disease data. The major advantage of the proposed method lies in its novel integrated view on *trans/cis* eQTL regulation instead of classical one-way isolated individual mining of *trans* and *cis* eQTL

loci. In addition, a modular nature of the method allowed us to measure the overlap between 2-way *trans/cis* and 1-way *cis* only regulatory program components reaching 18.7% pathway overlap (Figure 4.4). Thanks to semi-parametric MB-MDR nature that lacks linearity assumption between predictor and response variables, the method allowed exploring interactions with the minimum specification of parameters. In addition, results from Chapter 3 showed that MB-MDR can successfully handle a moderate strength LD ( $r^2 = 0.75$ ) based on the results on ankylosing spondylitis WTCCC2 data where LD received the lowest impact ranking compared to other parameters (see Figure 3.3). Another study confirmed our previous conclusion showing that even in the presence of a very strong LD ( $r^2 = 0.9$ ) MB-MDR performance declined very mildly based on FWER increase of only 0.01 [30]. Although the impact of LD on MB-MDR performance is an opened ended question, the proposed *trans/cis* eQTL MB-MDR-based method can handle LD which is not the case with classical regression-based methods where LD causes multicollinearity issues that can't be easily remediated. [31]. The MB-MDR tolerance to LD was shown in Chapter 3, specifically Figure 3.3, where there was minimal distance between MB-MDR CODOMINANT protocols 5 and 6 applied on pre-selected data. Another benefit of the proposed method lies in its ability to concentrate on epistatic *trans* and *cis* eQTL loci interaction effects while correcting for main effects (one-way associations). Finally, the network component of the method allowed to offer valuable complementary global interaction view using as input an individually listed *trans/cis* eQTL results. Specifically, the topology of the inferred directed *p*-value weighted network (Figure 4.8) allowed identification of the key hub *trans/cis* eQTL genes unveiling molecular mechanisms of the complex and heterogeneous disease - asthma. Despite the advances in eQTL detection methods there is a large room for further improvements including accountability for the LD patterns, more accurate loci gene mapping, the inclusion of higher order interactions (e.g., 3-way interactions), integration of several biological evidence from different sources (e.g., expression, phenotypic impacts, etc.).

Finally, expansion of eQTL methods to deal with higher-order interactions is highly desirable. The incorporation of phenotype data into eQTL detection methods is also very promising pioneered by transcriptome-wide association study of Gusareva *et al.* [20]. By taking into account disease status, condition specific epistatic eQTLs can be detected making them highly relevant to clinical settings. An example of such effort that not only considers gene-gene interactions under co-expression but

also under co-regulation contexts via the shared set of eQTL loci is represented by the eQTL-based gene–gene co-regulation network (GGCRN) method[32].

From a biological point of view, eQTLs are frequently the hotspots of the co-expression and PPI networks involving genes sharing common biological function or process [16,29,33]. For example, a study of Zhu *et al.* indicated that co-expression and PPI networks in yeast show a large topological overlap in which *cis* eQTLs are more likely to be linked to the hub-genes [33]. Thus, an integrative analysis of eQTL results coupled to network topology and functional annotation assessments is important and can reveal relevant gene regulatory mechanisms linked to the dataset context.

#### **7.4. GRN inference via trees from microarray expression data**

The previous chapters of this thesis explored different methods to detect gene-gene interactions. Later functional analysis of interactions via pathway enrichment and literature search linked them to complex disease etiology highlighting plausible pathological mechanisms. Due to their versatility and performance, the tree-based methods were applied in the context of gene regulatory/transcriptional network inference describing gene-gene interactions. Both pure machine learning and hybrid statistical/machine learning tree-based methods were explored via Random Forest (*RF*) and Conditional Inference Forest (*CIF*) algorithms, respectively. In Chapter 5 we took a closer look at the *CIF* method which offers attractive features. These features are conditional permutation scheme, a solid statistical framework for estimation of node significance, stringent stopping criteria controlling tree growth. Especially the conditional permutation feature of *CIFs* was of great interest since biological variables display various degrees of correlation (e.g., LD structure between markers, co-expressed genes, transcription factor – target gene expression links, etc.). Due to computational limitations, the conditional permutation scheme was only tested on a simulated data [34] requiring further tests on a real-life omics data. Unfortunately, the Hothorn’s conditional permutation scheme implemented in the *party* R library [35] was not scalable to genome-wide and transcriptome-wide contexts due to its high computational demands especially accentuated during conditional permutation scheme (see Section 5.4.1).

In light of these shortcomings, a significantly speedier non permutation-based version of *CIFs* was introduced. This novel approach, *CIF<sub>mean</sub>*, averaged the node-specific association values, expressed in a form of a test-statistic or a *p*-value, without the need of additional computationally demanding permutation steps in the variable importance measure (*VIM*) calculation. In practice, this simplification of *VIM* calculation brought up significant speed gains at an expense of performance. The performance drop in most cases was not significant allowing *CIF<sub>mean</sub>* to compete effectively with the reference methods such as *RFs*. The time advantage of the *CIF<sub>mean</sub>* over *RF* was almost 2 fold in the DREAM5 network composed of 4,511 nodes (3,232 minutes versus 6,054 minutes, respectively).

The speed of tree-based methods is especially critical in directed network inference since all possible pair-wise interactions need to be considered. Thus, the complexity of exhaustive implementations is often exponential  $O(exp)$ . Hence, network inference is highly time-demanding and performance sacrifices are well justified. Fortunately, there exist alternatives including various variable aggregation strategies, dynamic programming and others. Some of variable aggregation strategies explore feature similarity and prior knowledge (e.g., biological relevance). Fortunately, tree-based methods for networks can be easily parallelized thanks to independent tree inference of the forest and *VIM* calculation for each gene-gene interaction.

One negative caveat associated with all tree-based methods is the need to tune them up tree inference parameters requiring possession of a gold standard and a training set. The common parameters to tune are *mtry* and *ntrees* referring to the number of randomly picked variables and number of trees in a forest, respectively. Our results also confirmed a strong *CIF* performance dependence on *mtry* parameter. A positive performance trend with increased *mtry* value was especially seen in *CIF*-based methods across synthetic and real-life datasets (see Chapter 5) with optimal value ranging between 5 and 1/3 of all input variables ( $mtry=k/3$ ). In this thesis, optimal parameters for *CIF<sub>mean</sub>* methods were highlighted across the diverse datasets (see Section 5.4).

The information on biological systems is increasing in diversity as more and more omic datasets become available thanks to advancements in high-throughput screening technologies and IT

infrastructure [15]. The holistic views of systems biology require clever and biologically sound omic data integration discussed in the subsequent section.

## 7.5. Integration

Increased accessibility to diverse high-throughput technologies steadily increases the number of omics data types representing various omics layers (e.g., genotypes, expression, methylation, copy number variation, protein-protein, microbiome, etc.). Many studies related to complex diseases contain one or more diverse sources of data requiring novel ways of data integration. As mentioned in Section 2.5, data integration and fusion are two different terms with integration taking a broader context [10,36]. Data processing is especially challenging in integrative context due to different data structures and formats [37]. When integrating data, one is inevitably faced with probe aggregation issue involving multiple probes mapping to the same entity (e.g., gene). One can adopt probe aggregation schemes based on physical and functional mapping, but there are alternative approaches where one can combine, for example, genomic and methylomic data in a “clever way” via kernel PCA summarizing a given region of interest (ROI) [38]. Compared to functional and physical probe mapping strategies, the ROI-based method seems to be the most promising as information from multiple probes is aggregated into a single construct. The ROI profiles can represent, for example, genes or pathways and will be built by calculating similarities measures based on diffusion kernel and PCA components. In addition, incorporation of graph structures makes it possible to incorporate valuable information from other omic layers (e.g., PPI networks). Currently in our lab the ROI based integration pipeline is being developed with future MB-MDR extension to accommodate ROI profiles.

The functional probe aggregation based on association tests between probes and target gene expression can be adopted for SNP probe aggregation. This approach mimics a typical eQTL study measuring marginal *cis* eQTL effects. We proposed selection of probe with the strongest association (e.g., minimal *p*-value). As association measure for the SNP probe aggregation, we suggest median test (MED) proposed by Ziegler *et al.* [26] since there non-normal distribution of the expression data is accounted for via mutual information measure coupled with permutation-

based significance assessment. MED is more powerful for long-tailed distributions [39].

Also, one needs to take into account the LD structure patterns existing between loci as commented in Chapter 3. Although MB-MDR in Chapter 3 was minimally affected by LD structure in pre-selected data, LD is still an issue to be addressed by novel association measures relevant for probe aggregation and eQTL contexts. For example, Gusareva *et al.* developed a novel method that takes LD structure into account via covariance matrix during the association quantification between *cis* SNP, expression and phenotypic trait in transcriptome-wide association study (TWAS) [20].

In Section 2.5 integration methods were broadly classified into three categories one of which were explored in this thesis work (strategy 3). Specifically, a separate analysis of each data source (see Section 1.3) was explored in the case of *Regression2Net* method where expression-expression and expression-methylation networks were separately inferred and combined into a final *ANDnet*, *XORnet* or *INTnet* networks. This methodology allows for different ‘flavors’ of network fusion integration highlighting either consensus or uncooperative interactions present in both data sources summarized by *ANDnet* and *XORnet*, respectively (Section 6.3.3). Nevertheless, this leaves a user with integration choices that are difficult to choose.

## 7.6. Perspectives

The rapid increase of genomic, transcriptomic, proteomic, metabolomic and other types of data from high-throughput sources, has increased the need for an integrative analysis requiring practical solutions to *Big Data to Knowledge* (BD2K) problem [40] (see Section 1.1). In the near future we should expect a steady increase of studies that utilise different types of integration [40]. Before taking a major leap towards the new generation of multi-omics interaction methods, several major theoretical and practical obstacles remain to be solved including standardization of omics data deposition, proper quality control, speedier and parallelized versions of currently available methods, improved visualization and accessibility. The future generation of integrative methods needs to address problems related to decreased run-time requirements, correlation patterns, and other hidden confounder factors. The new generation of integrative algorithms should incorporate quality measures to assess reliability of the input data in order to achieve significant improvements

in several performance criteria including increased accuracy and decreased false positive and false negative rates [40]. In addition, future studies should solve an open question of meaningful validation and gold standard procurement to allow for more accurate and unbiased assessment of new integrative methods under similar conditions [41].

*Big Data* expansion prompted the development of self-evolving genetic algorithms (GA) such as biological natural language processing (NLP) [42], incorporating automatic learning and a certain degree of mutation/alteration after integration of previously unseen new training samples. The self-evolving algorithmic feature seems to be very attractive in the context of *Big Data* era and NP-hard problems characterized by extremely large search spaces. Feature selection methods incorporating GA in the context of marker selection are already appearing [43,44].

An integrated understanding of interactions in the genome, the transcriptome, the proteome, the environment mediated by the underlying cellular network, gives a firm ground for future advances. Identification of the relevant system components via integrative approaches will better characterize complex diseases thru identification of the key functions to be possibly altered via drugs. It will require strong collaboration between machine learning, statisticians, computer scientists, and biologists to implement new data integrative algorithms that may lead to an increased understanding of complex diseases. Ultimately, we need to better understand cell functioning and the act under ‘think globally, act locally’ paradigm [7].

## 7.7. References

1. Silverman EK, Loscalzo J (2012) **Network medicine approaches to the genetics of complex diseases.** *Discov Med* 14: 143-152.
2. Roca J, Cano I, Gomez-Cabrero D, Tegnér J (2016) **From Systems Understanding to Personalized Medicine: Lessons and Recommendations Based on a Multidisciplinary and Translational Analysis of COPD.** *Systems Medicine*: 283-303.
3. Bessonov K, Gusareva ES, Van Steen K (2015) **A cautionary note on the impact of protocol changes for genome-wide association SNP x SNP interaction studies: an example on ankylosing spondylitis.** *Hum Genet* 134: 761-773.
4. Lishout FV, Gadaleta F, Moore JH, Wehenkel L, Steen KV (2015) **gammaMAXT: a fast multiple-testing correction algorithm.** *BioData Min* 8: 36.
5. Gusareva ES, Van Steen K (2014) **Practical aspects of genome-wide association interaction analysis.** *Hum Genet* 133: 1343-1358.

6. Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, et al. (2012) **SNP interaction detection with Random Forests in high-dimensional genetic data.** *BMC Bioinformatics* 13: 164.
7. Barabási A-L, Gulbahce N, Loscalzo J (2011) **Network medicine: a network-based approach to human disease.** *Nature Reviews Genetics* 12: 56-68.
8. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, et al. (2003) **Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans.** *Nat Genet* 33: 518-521.
9. Bush WS, Moore JH (2012) **Chapter 11: Genome-wide association studies.** *PLoS Comput Biol* 8: e1002822.
10. Van Steen K, Malats N (2014) **Perspectives on Data Integration in Human Complex Disease Analysis.** In: Wang B, Li R, Perrizo W, editors. *Big Data Analytics in Bioinformatics and Healthcare*. 1 ed: IGI Global. pp. 284-322.
11. Fish A, Capra JA, Bush WS (2015) **Are Genetic Interactions Influencing Gene Expression Evidence for Biological Epistasis or Statistical Artifacts?** *bioRxiv*: 020479.
12. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) **Systematic identification of trans eQTLs as putative drivers of known disease associations.** *Nat Genet* 45: 1238-1243.
13. Franke L, Jansen RC (2009) **eQTL analysis in humans.** *Methods Mol Biol* 573: 311-328.
14. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, et al. (2014) **Detection and replication of epistasis influencing transcription in humans.** *Nature* 508: 249-253.
15. Ge H, Walhout AJ, Vidal M (2003) **Integrating 'omic' information: a bridge between genomics and systems biology.** *Trends Genet* 19: 551-560.
16. Zhang W, Liu JS (2010) **From QTL Mapping to eQTL Analysis.** *Frontiers in Computational and Systems Biology*: Springer. pp. 301-329.
17. Horvath S, Dong J (2008) **Geometric interpretation of gene coexpression network analysis.** *PLoS Comput Biol* 4: e1000117.
18. Ponsuksili S, Siengdee P, Du Y, Trakooljul N, Murani E, et al. (2015) **Identification of common regulators of genes in co-expression networks affecting muscle and meat properties.** *PLoS One* 10: e0123678.
19. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, et al. (2011) **Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue.** *PLoS Genet* 7: e1002078.
20. Gusev A, Ko A, Shi H, Bhatia G, Chong W, et al. (2015) **Integrative approaches for large-scale transcriptome-wide association studies.** *bioRxiv*: 024083.
21. Zhang X, Huang S, Sun W, Wang W (2012) **Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study.** *Genetics* 190: 1511-1520.
22. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O (2015) **Fast and efficient QTL mapper for thousands of molecular phenotypes.** *Bioinformatics*: btv722.
23. Salyakina D, Seaman SR, Browning BL, Dudbridge F, Muller-Myhsok B (2005) **Evaluation of Nyholt's procedure for multiple testing correction.** *Hum Hered* 60: 19-25; discussion 61-12.
24. Nyholt DR (2004) **A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other.** *The American Journal of Human Genetics* 74: 765-769.
25. Li J, Ji L (2005) **Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix.** *Heredity (Edinb)* 95: 221-227.
26. Szymczak S, Igl BW, Ziegler A (2009) **Detecting SNP-expression associations: a comparison of mutual information and median test with standard statistical approaches.** *Stat Med* 28: 3581-3596.
27. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) **eQED: an efficient method for interpreting eQTL associations using protein networks.** *Mol Syst Biol* 4: 162.

28. Gligorićević V, Przulj N (2015) **Methods for biological data integration: perspectives and challenges.** *J R Soc Interface* 12.
29. Pilpel Y, Sudarsanam P, Church GM (2001) **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 29: 153-159.
30. Mahachie John JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, et al. (2011) **Comparison of genetic association strategies in the presence of rare alleles.** *BMC Proc* 5 Suppl 9: S32.
31. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, et al. (2013) **Collinearity: a review of methods to deal with it and a simulation study evaluating their performance.** *Ecography* 36: 27-46.
32. Li J, Wang L, Guo M, Zhang R, Dai Q, et al. (2015) **Mining disease genes using integrated protein-protein interaction and gene-gene co-regulation information.** *FEBS Open Bio* 5: 251-256.
33. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, et al. (2008) **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.** *Nat Genet* 40: 854-861.
34. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) **Conditional variable importance for random forests.** *BMC Bioinformatics* 9: 307.
35. Hothorn T, Hornik K, Strobl C, Zeileis A, Hothorn MT (2014) **Package ‘party’.** *Package Reference Manual for Party Version 09-998* 16: 37.
36. Oxley ME, Thorsen SN (2004) **Fusion or Integration: What's the Difference?** : DTIC Document.
37. Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, et al. (2014) **A practical data processing workflow for multi-OMICS projects.** *Biochim Biophys Acta* 1844: 52-62.
38. Fouladi R, Bessonov K, Van Lishout F, Van Steen K (2015) **Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis.** *Hum Hered* 79: 157-167.
39. Szymczak S, Scheinhardt MO, Zeller T, Wild PS, Blankenberg S, et al. (2013) **Adaptive linear rank tests for eQTL studies.** *Statistics in medicine* 32: 524-537.
40. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, et al. (2009) **Data integration in genetics and genomics: methods and challenges.** *Hum Genomics Proteomics* 2009.
41. Elena S, Gusareva NM, Kristel Van Steen (2016) **Omics integration to uncover genome-phenome associations.** *Human Genetics*.
42. Yandell MD, Majoros WH (2002) **Genomics and natural language processing.** *Nat Rev Genet* 3: 601-610.
43. Popovic D, Moschopoulos C, Sakai R, Sifrim A, Aerts J, et al. **A self-tuning genetic algorithm with applications in biomarker discovery;** 2014. IEEE. pp. 233-238.
44. Ang KS, Kyriakopoulos S, Li W, Lee D-Y (2016) **Multi-omics data driven analysis establishes reference codon biases for synthetic gene design in microbial and mammalian cells.** *Methods*.



## Chapter 8: CV and Publications





## 8. CV and Publications

This section provides the complete list of publications produced or written during my PhD training and my Curriculum Vitae.

### 8.1. Publication list (2012-2016)

- Bessonov K, Van Steen K (2016) **Practical aspects of gene regulatory inference via conditional inference forests from expression data**. (accepted for publication in Genetic Epidemiology)
- Francesco G<sup>¶</sup>, Bessonov K<sup>¶</sup>, Van Steen K (2016) **Integration of Gene Expression and Methylation to unravel biological networks in glioblastoma patients**. (accepted for publication in Genetic Epidemiology)
- Bessonov K, Gusareva ES, Van Steen K (2015) **A cautionary note on the impact of protocol changes for genome-wide association SNP x SNP interaction studies: an example on ankylosing spondylitis**. *Hum.Genet* 134:761-773
- Pineda S, Gomez-Rubio P, Picornell A, Bessonov K, Márquez M, Kogevinas M, Real FX, Van Steen K, Malats N. (2015) **Framework for the integration of genomics, epigenomics, and transcriptomics in complex diseases**. *Hum Hered.* 79(3-4):124-36
- Bollen L, Vande Casteele N, Peeters M, Bessonov K, Van Steen K, Rutgeerts P, Ferrante M, Hoylaerts MF, Vermeire S, Gils A. (2015) **Short-term Effect of Infliximab Is Reflected in the Clot Lysis Profile of Patients with Inflammatory Bowel Disease: A Prospective Study**. *Inflamm Bowel Dis.* 21(3):570-8
- Gusareva ES, Carrasquillo MM, Bellenguez C, Cuyvers E, Colon S, Graff-Radford NR, Petersen RC, Dickson DW, Mahachie John JM, Bessonov K, Van Broeckhoven C; GERAD1 Consortium, Harold D, Williams J, Amouyel P, Sleegers K, Ertekin-Taner N, Lambert JC, Van Steen K (2014) **Genome-Wide Association Interaction Analysis for Alzheimer's Disease**. *Neurobiol Aging.* 35(11):2436-43
- Fouladi R, Bessonov K, Van Lishout F, Van Steen K. (2015) **Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis**. *Hum Hered.* 79(3-4):157-67

Legend: <sup>¶</sup> - joint first author

## 8.2. Curriculum Vitae

5 years of academic training and research experience in Bioinformatics and Statistical Genetics

5 years academic teaching experience during PhD and masters

4 years of academic and 1 year of industry Biochemistry / Biotechnology lab experience

3 years of R, Python, C/C++ and Perl, MATLAB practical programming experience

Quick learner, hard working, enthusiastic, and dependable

Strong communication and interpersonal skills

6 years of university course teaching experience

Languages: English, French, Spanish, Russian, and Ukrainian

### PROFESSIONAL SKILLS

<i>Bioinformatics/ Computer Skills</i>	Intermediate in C/C++, Perl, R, DBs, Linux Linux admin skills, web site development and maintenance (WordPress) Software developer for data analysis pipelines Gene expression graph-based data clustering and functional analysis Protein modeling and molecular dynamics simulations with Gromacs®
<i>Genomic/Statistical</i>	Proficient in data mining and general omics data analysis Experience in the analysis methods involved in large-scale genetic association, eQTL, network-based studies for complex diseases
<i>Analytical Chemistry</i>	Working knowledge of HPLC/GC coupled with Mass Spectroscopy Experienced working with Waters ESI-MS, Bunker MALDI
<i>Immunology/Histology</i>	Immunofluorescence and Immunohistochemistry microscopy Tissue fixation, embedding and staining for slides
<i>Microbiology</i>	Good Aseptic Technique practices Gene knockouts using PCR cassettes and mutant selection
<i>Biotechnology</i>	Purification and separation protein techniques DNA/ RNA / Protein isolation and quantification: PCR, RT-PCR, primer design, PAGE SDS-PAGE, Western/Northern Blots, ELISA
<i>Microscopy</i>	Phase-contrast epifluorescence imaging

**EDUCATION**

**Ph.D in Engineering (Bioinformatics and Statistical Genetics)** **2012 – 2016**

University of Liege, Liege, Belgium

Thesis: From Statistical to Biological Interactions via Omics Integration

**M.Sc in Bioinformatics** **2010 – 2011**

University of Guelph, Guelph, ON, Canada

Thesis: Functional Characterization of the NSF1 (YPL230W) Gene using Correlation Clustering and Genetic Analysis in *Saccharomyces Cerevisiae*

**Honours B.Sc in Biochemistry(major)** **2006 – 2009**

University of Guelph, Guelph, ON, Canada

Summer Research Project 2007: Neurogenesis of stem cells from umbilical cord blood (UCB)

**Biotechnology Lab Technologist (Research)** **2003 – 2006**

Seneca College, Toronto, Ontario

**PROFESSIONAL EXPERIENCE**

**PhD student** **Sept 2012 - present**

University of Liege, Belgium

“Omics” data processing and useful knowledge inference. Detection of interactions in complex diseases. Teaching of courses.

**Bioinformatics Research Assistant – Dr. George Harauz** **May 2009 – Sept 2009**

University of Guelph, Ontario

Molecular Dynamics simulations setup and analysis of the MBP protein using GROMACS. Development of algorithms for quick hydrophobic moment calculation programmed in C++. ‘Wet-lab’ protein isolation and purification. Software development for quick data analysis and workflow processing pipelines.

**Teacher Assistant** **May 2010 – Dec 2011**

University of Guelph, Ontario

Thoroughly prepared and conducted Applied Biochemistry labs involving protein isolation and purification. Assisted students by giving a professional advice often receiving a positive feedback.

**Research Assistant – Dr. Mansel Griffiths, Food Science**

**Jan2009-Apr2009**

Canadian Research Institute for Food Safety, University of Guelph, Ontario

Microbiological characterization of phage virus. Phage coated membranes development for meat preservation.

**ESI-MS / MALDI Technologist - Antibodies**

**May 2008-Dec 2008**

Roche Diagnostics, Penzberg, Germany

Conducted MS Analysis of antibodies and other proteins using ESI-MS and MALDI. Automated some lab procedures. Optimized protocols. Collaborated with other researchers. Built mouse antibody database.

**Research Assistant – Rob Merrill's Protein Lab**

**Sept 2007-May 2008**

University of Guelph, Ontario

Assisted lab manager with ongoing projects and miscellaneous tasks including protein purification including French press of cells and centrifugation of lysates, running of purified protein samples on FPLC; Prepared and run SDS gels; made various buffers/solutions as needed

**Summer Research Assistant – Dean Betts Stem Cell Lab**

**May 2007-Sept 2007**

University of Guelph, Ontario

Tissue culture of mesenchymal stem cells (MSC) and embryo cultures. Differentiation experiments of MSCs

Completed independent summer project on neurogenesis potential of equine UBC stem cells

**Clinical Studies Technician**

**May 2006-Dec 2006**

Allied Research International, Mississauga, Ontario

Collected pollen samples to help with clinical studies of allergens in a chamber. Interacted and helped patients

**Science Tutor**

**Jan 2005-May 2006**

Seneca College, Toronto, Ontario

Tutored various science related subjects. Worked on one on one basis including grade 2 to 12 students. Lead and prepare Math workshops fostering student interactions under friendly atmosphere.

## **VOLUNTEERING**

Hospital Volunteer in Cancer Unit

**2005**

Sunnybrook & Women Hospital, Toronto

## **ACCOMPLISHMENTS - SCHOLARSHIPS**

Belgian federal FNRS ASP PhD scholarship 2012-2016

Best paper award in Bioinformatics track at SAC2011 conference, Taiwan, 2011

OGSST and OGS scholarship recipient 2011 for a total sum of \$15000

Graduate Student Teaching Conference, Guelph, 2011

2008 DAAD - Deutschen Akademischen Austauschdienstes Dienst RisePro Scholarship recipient

Summer Leadership and Research Program 2007 at Uof G, Ontario Veterinary College

SMILE Mentoring Certificate of Achievement, Seneca College

French immersion Explore program scholarship holder 2006