Statistical methods for the analysis of complex omics data

Tissier, R.

1 Introduction

1.1 Introduction

In the last decades, technical developments in biomolecular research have made it possible to collect various omics measurements such as, gene expression, transcriptomics, proteomics, metabolomics, and glycomics. All these measurements, have improved our knowledge of the biological functions in the human body and the mechanisms which get activated in complex diseases. Prediction of disease phenotypes using such omics data in addition to classical environmental factors has also been made possible, opening thereby new research directions for personalized medicine.

Despite the broad availability of omics measurements, the statistical analysis does not always match the complexity of the data generating process. In particular, data collected in complex study designs such as family studies are often analysed without properly taking into account the sampling mechanism and the relatedness of family members. Moreover, incorporating biological or empirically derived information is not always exploited minimizing thereby the potential of state-of-the-art prediction approaches. Furthermore, prediction models are typically based on a unique omic source, thereby, neglecting the potential gain by combining multiple sources of omic predictors. This fact limits the accuracy of personalized prediction and work has to be done on the integration of multiple omic sources in prediction models as there is, actually, no state-of-the-art approach for this type of prediction problem. The development of advanced statistical methods to address the aforementioned complexities is the topic of this thesis. In particular, we present: (i) methods for modelling associations between phenotypes and omics data while cor-

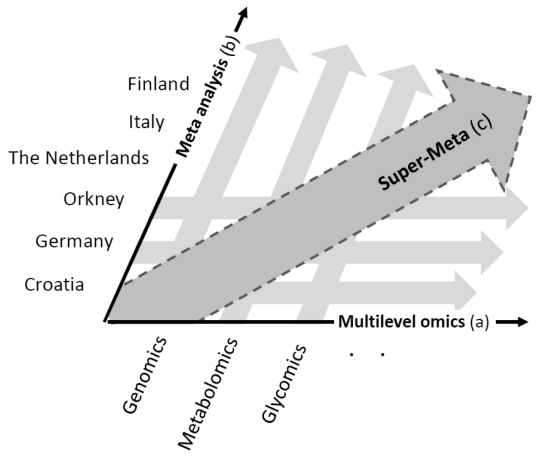

Figure 1.1: Diagram of the super meta-analysis combining several datasets as well as integrating multiple omic sources available in the MIMomics consortiums.

recting for the sampling mechanism in family studies, (ii) methods to build networks of omic features which are collected in family data, (iii) methods to improve prediction by adding information of the correlation structures in group penalization models using only one omic source, and (iv) the extension to several omic sources prediction models.

The research conducted in this thesis was part of the European collaborative project: Methods for Integrated analysis of Multiple Omics datasets (MIMOmics). The goals of the project were: (1) the development of a statistical framework of methods for all analysis steps needed for identifying and interpreting omics-based biomarkers, and (2) to integrate such data derived from multiple omics platforms within studies and across studies and populations. The second goal of the project, namely the development of a super meta-analysis framework, is visualized in Figure 1.1. To establish this super meta-analysis framework the development of robust statistical methodologies which are able to take into account the dependence between omic features, relatedness of individuals in the studies, high dimensionality of datasets and the sampling process were needed.

The methods developed in this thesis can be applied on the (a) and (b) axis of Figure 1.1. In particular, the methods that will be presented in the next chapters focus on the proper analysis of separate omics data under complex study designs and the integration of the various omics in predicting disease outcomes. The proper analysis of omics data under complex study designs allows the integration of the results via meta-analyses (axis (b)) and the analysis of multiple phenotypes simultaneously (axis(a)), while integrating various omics sources in a single prediction model grants the possibility to combine several measurements from one study (axis (a)). The development of such methods was necessary to achieve super meta-analyses.

The rest of this introductory chapter is organized as follows. We will first discuss

the key ingredient linking all thesis' chapters, namely the dependence between random variables and present the measures used in the coming chapters to quantify it. Next, we will present the modelling of this dependence in three settings: (i) between individuals i.e. when analysing data from family studies, (ii) between omic features, i.e. when building networks and prediction models and (iii) between outcomes measured on the same subjects. Finally, we will close with a short presentation of the chapters included in this thesis.

1.2 Measure of dependence: Pearson correlation coefficients

One of the most common measures of dependence is the correlation which captures a particular type of dependence, namely linear dependence. Let X and Y two random variables with finite variances σ_x^2 and σ_y^2, the correlation of X and Y which is denoted by ρ_{xy} is given by:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x^2 \sigma_y^2}, \tag{1.1}$$

where σ_{xy} is the covariance between X and Y.

In the case where multiple random variables are recorded, e.g. multiple omic features available in a dataset M, the correlation coefficient ρ_{xy} can be applied on all possible pairs of features leading thereby to the correlation matrix $\mathbf{R_M}$. Modelling such linear associations between omic features in the context of multivariate regression models or network methods is our main concern in Chapters 2-6.

In particular, in Chapters 2 and 3 the correlation matrix of the error terms of multivariate regression models is used to model the dependencies between multiple features measured on the same members of the same family. In Chapters 4-6 the correlation matrix is the input to construct weighted networks. Weighted networks, in general are defined as an adjacency matrix $\mathbf{A} = [a_{ij}]$, where each coefficient a_{ij} represent how close features i and j are. Each non zero coefficient in the matrix $\mathbf{A}$ represent an existing edge between two nodes in the network. One straightforward approach to compute a network of features is to compute their correlation matrix. More sophisticated methods have been developed to obtain more relevant or more interpretable networks. Specifically, in this thesis, we use the weighted gene coexpression network analysis (WGCNA , Zhang and Horvath (2005)) which uses a soft thresholding approach to make the adjacency sparser. The thresholding used in WGCNA is designed to produce network following a free scale topological criteria. This criteria, explained in Chapters 4-6, allows network to follow a hub model. Hubs models are believed to be representative of certain biological mechanisms, and are especially relevant in gene expression (Zhang and Horvath, 2005), helping investigators to identify groups of related features with a meaningful biological interpretation of these groups.

A limitation of the correlation coefficient (Eq 1.1) is that it cannot distinguish direct

from indirect linear dependencies. For instance, let X, Y and Z three random variables of interest. Figure 1.2 illustrates one example of possible relationships between them, where the presence of a link between the nodes implies the presence of a direct linear dependence. In this case, the fact that X and Y are both linearly related to Z will lead to ρ_{xy} different from 0 indicating the existence of correlation between them. This dependence is indirect. Making the distinction between direct and indirect dependencies is necessary when trying to identify groups of biologically related features. In this case, the use of partial correlation (Fisher, 1924) is preferred to avoid confounding effects.

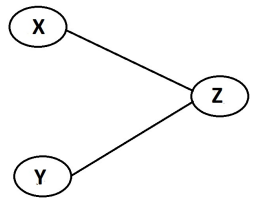

Figure 1.2

The partial correlation, is the correlation between X and Y given other variables, i.e. the conditional dependence between them. In particular, for the triplet of random variables (X, Y, Z) the partial correlation ρ^*_{xy} of X, Y given Z is written as:

$$\rho_{xy} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho^2_{xz}}\sqrt{1-\rho^2_{yz}}}$$

In the case of a high number of omics features in a dataset M, the partial correlation for all pairs of variables can be derived in terms of the correlation matrix $\mathbf{R_M}$ by $\mathbf{R^*_M} = \text{scale}(\mathbf{R^{-1}_M})$, where for a matrix $\mathbf{A}$, $\text{scale}(\mathbf{A}) = \text{diag}(\mathbf{A})^{-1/2}\mathbf{A}\text{diag}(\mathbf{A})^{-1/2}$, with $\text{diag}(\mathbf{A})$ the vector of the diagonal elements of $\mathbf{A}$. Note that when the number of variables exceeds the number of samples in the dataset, the correlation matrix is not invertible. To compute the partial correlation matrix in this case and to make it sparser regularized regression methods can be used, as described in Chapter 5. In particular, penalty functions are introduced to shrink correlation coefficients either separately or as a group in order to allow for inversion of the covariance matrix while retaining only relevant coefficients, forcing other partial correlation coefficients to be equal to 0. Furthermore, the use of group penalization grants the possibility to include a priori knowledge on existing relationships between variables. The partial correlation matrix is a key ingredient of the Gaussian graphical models (Lauritzen, 1996). These models are used in Chapter 5

and can be applied in Chapter 6. Graphical models are used to provide a representation of the existing direct interdependence between several variables allowing investigators to identify groups of features. Compared to WGCNA this approach does not force the network to follow a hub structure. The groups of identified features with graphical models are often smaller since they only contain features having direct interdependences.

Apart from the linear dependence between omic features, another potential source of correlation in genomic data is the relationships between members of the same family. Methods for modelling such dependence is the topic of the next section.

1.3 Family studies

Family studies are often used in genetic research to understand the role of genetics and shared environment in the etiology of disease. In the last years, in addition to genetic markers, several omic measurements are being collected for existing family studies in order to further improve our understanding of human diseases. In family studies, a commonly used design oversamples families enriched with the disease under study, i.e. we only recruit families with at least a certain number of cases. This is the so called multiple-cases family study design. Statistical inference under such a design is known to be robust to population stratification and efficient for detecting rare genetic variants as they tend to aggregate within families. Despite the strengths of family studies, we should acknowledge that recruiting disease-enriched families is harder than the sampling in case-control studies. Moreover, the statistical analysis of family data requires sophisticated approaches (de Andrade and Amos, 2000; Kraft and Thomas, 2000) which explicitly model the familial relationships and deal with the biased sampling design.

1.3.1 Modelling between subject correlation in family studies

Regarding the within families correlations, mixed-effects models are typically used. Let Y_i be a quantitative phenotype for family $i = 1, \ldots, n$ and X_i the $n_i \times (p+1)$ design matrix of p omic features. The linear mixed effects model to study the association between X_i and Y_i in a family i is written as:

$$Y_i = \beta X_i + \sigma_G g_i + \sigma_E e_i + \sigma_\epsilon \epsilon_i \tag{1.2}$$

with β the fixed effects parameter vector, g_i the $n_i \times 1$ random effects vector which models the familial genetic correlation, e_i the shared environment $n_i \times 1$ random effects vector and $\epsilon_i \sim N(0; \sigma^2_\epsilon I_i)$ the $n_i \times 1$ residual error terms vector. The genetic and environmental variances are σ^2_G and σ^2_E, respectively. The random variables, g_i, e_i and ϵ_i are assumed to be independent and to follow the multivariate normal distribution with $g_i \sim N_{n_i}(0; K_i)$ and $e_i \sim N_{n_i}(0; E_i)$, where E_i is the environmental correlation matrix often defined as a unit matrix, as all family members share the same environment. K_i is the relationship matrix representing the genetic relatedness between family members.

Let a and b be two members of family i, then the coefficient of the relationsgip matrix between a and b is $k_{iab} = 2^{-d(a,b)}$, where $d(a,b)$ is the genetic distance between family members a and b. The coefficient of relatedness k_{iab} represents the probability that a random allele is shared identical by descent (IBD, Thompson (2008)) by a and b, i.e. the probability that the allele is inherited from a common ancestor. From equation 1.2 it follows $\mathbf{Y_i} \sim (\beta \mathbf{X}_i, \sigma_G^2 \mathbf{K}_i + \sigma_E^2 \mathbf{E}_i + \sigma_\epsilon^2 I_i)$. For a sample of randomly selected families, the parameters β, σ_G, σ^2, and σ_ϵ can be estimated by maximizing the likelihood function L:

$$\mathbf{L} = \prod_i P(\mathbf{Y}_i | \mathbf{X}_i)$$

For binary phenotypes Y_i, generalized linear mixed model such as probit mixed models or logistic models are used. Such models are used in chapters 2-3 of this thesis.

1.3.2 Modelling ascertainment in family studies

The analysis of family studies is complicated by the oversampling of disease enriched families also known as ascertainment. To derive unbiased estimates of the omics effects on disease phenotypes and heritability related parameters in this case, we need to correct for the chosen sampling scheme. In the literature several approaches have been proposed to address this issue. Namely, the prospective, retrospective and joint likelihood approach (Kraft and Thomas, 2000). Let E be an exposure (catagorical variable), Y the case-control status (binary variable), and S the ascertainment process.

Under the prospective likelihood approach, we condition on the sampling process as shown in equation below:

$$P(\mathbf{Y} \mid \mathbf{E}, \mathbf{S}) = \frac{P(\mathbf{E}, \mathbf{Y}, \mathbf{S})}{P(\mathbf{E}, \mathbf{S})} = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E}) P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{S} \mid \mathbf{E})},$$

which can be further simplified by assuming complete ascertainment, i.e. for all individuals included in the sample $P(S \mid Y) = 1$. We obtain:

$$P(\mathbf{Y} \mid \mathbf{E}, \mathbf{S}) = \frac{P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{S} \mid \mathbf{E})}$$

For multiple-cases family studies the denominator $P(\mathbf{S} \mid \mathbf{E})$ can be easily modelled. However, for more complex sampling design, modelling the ascertainment process can be challenging. In such cases, the retrospective likelihood is preferred as this approach corrects implicity for the ascertainment if the ascertainment process depends only on the case-control status.

The retrospective likelihood is based on modelling the distribution of covariates conditional on the outcome and the ascertainment and can be expressed as follows:

$$P(\mathbf{E} \mid \mathbf{Y}, \mathbf{S}) = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E}) P(\mathbf{E} \mid \mathbf{Y})}{P(\mathbf{S} \mid \mathbf{Y})} = P(\mathbf{E} \mid \mathbf{Y}) \tag{1.3}$$

By application of Bayes'rule, in equation 1.3, The probability of $\mathbf{Y}$ given $\mathbf{E}$ becomes:

$$P(\mathbf{E} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{E})}{P(\mathbf{Y})} = \frac{P(\mathbf{Y} \mid \mathbf{E})}{\sum_E P(\mathbf{Y} \mid \mathbf{E}) P(\mathbf{E})}$$

As previously stated, the main advantage of the retrospective likelihood is the fact that the ascertainment does not need to be modelled. However, this approach does need to model the distribution of the exposure E within families. Therefore, specific assumptions have to be made which provide biased parameter estimates in case of model misspecification. Another drawback of this approach is the loss efficiency by possibly over-conditioning on the phenotype Y of interest and the ascertainment event (Kraft and Thomas, 2000).

The last approach, the joint likelihood, is based on modelling the joint distribution of the exposure and phenotype given the sampling process and is given as follows:

$$P(\mathbf{Y}, \mathbf{E} \mid \mathbf{S}) = \frac{P(\mathbf{E}, \mathbf{Y}, \mathbf{S})}{P(\mathbf{S})} = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{E}) P(\mathbf{Y} \mid \mathbf{E}) P(\mathbf{E})}{\sum_E P(\mathbf{S} \mid \mathbf{E}) P(\mathbf{E})}$$

This approach combines both disadvantages of the prospective and retrospective likelihood as both ascertainment process and the distribution of the exposure within the family have to be modelled, but is the most efficient as it needs the weakest conditioning (Kraft and Thomas, 2000). Indeed, this approach relies only on the conditioning on the ascertainment process. In the specific case of a family study following a multiple cases design and the exposure of interest is a single nucleotide polymorphism (SNP), both the ascertainment and distribution of the SNP within the family can be modelled.

1.3.3 Examples of family studies

In this thesis data from two family studies are analysed. The Leiden Longevity Study (LLS, Schoenmaker et al. (2006); Houwing-Duistermaat et al. (2009)) is a family-based study set up to identify mechanisms that contribute to healthy ageing and longevity. The inclusion criteria of the study are sibships with at least two alive nonagenarian siblings. Several secondary phenotypes and GWAS data were measured for the offspring of these siblings and their partners. Since the offspring have at least one nonagenarian parent, they are also likely to become long-lived. Therefore, the set of offspring and their partners corresponds to a multiple cases design with related subjects where the offspring are considered as cases and their partners as controls. 421 families with 1671 offspring (cases) and 744 partners (controls) have been included in the study. In Chapter 2, we study the relationships between SNPs and metabolites measured in LLS. Namely, triglyceride levels and glucose levels.

The Leiden Family Lab study on Social Anxiety Disorder (LFLSAD, Bas-Hoogendam et al. (2018)) is a two generation multiplex family study aiming to identify endophenotypes linked to the social anxiety disorder (SAD). Families were considered eligible for inclusion when they contained at least one adult with a primary diagnosis of SAD and whom had at least one child living at home with the proband, showing SAD symptoms.

In addition to these probands other family members were included in the study leading to 9 families with a total number of samples of 132. In Chapter 3, we aim to identify endophenotypes, i.e. heritable phenotypes associated with a primary phenotype of interest, using electroencephalography (EEG) measurements.

1.4 Secondary phenotypes

In genetic studies, apart from the genetic variants and primary disease phenotype e.g. case-control status, a number of omic and non-omic phenotypes are collected as well. These additional phenotypes are known as secondary phenotypes. For instance, in the LLS in addition to case-control status and GWAS, metabolites, classical environmental factors, etc., are measured. Similarly, in the Leiden Family Lab Study omics and fMRI data are available.

In these studies, one of the main research questions is to identify genetic variants associated with these additional secondary phenotypes. In the context of LLS this would help us investigate the presence of pleiotropy, namely the existence of genes associated with multiple phenotypes. The study of pleiotropic effects is important to understand the underlying biological mechanisms of complex diseases. Identifying pleiotropic effects can improve personalized medicine as well. Since specific genetic variants may show strong associations with multiple traits but in opposite directions (Solovieff et al., 2013), identifying pleiotropic effects will help to better prevent and identify possible side effects after gene therapy or genome editing treatments (Solovieff et al., 2013; Gratten and Visscher, 2016). In the LFLSAD, one of the primary objectives is to identify endophenotypes for social anxiety disorder and the genetic variants associated with them. A trait is declared as endophenotype of a specific disease if it is associated with the disease status, if it manifests whether illness is active or in remission (state-independent), and when the trait and the disease status co-segregate within a family. The search for endophenotypes is important as psychologic diseases are complex to diagnose and diagnosis can be subjective. Therefore, identification of genetic variants or biomarkers associated with the disease is difficult. Studying instead the association between genetic variants and highly heritable disease-related phenotypes is needed to understand the relationship between psychological disorders and the genome.

In both studies testing for genetic variants associated with secondary phenotypes is complicated by the sampling mechanism. In both designs, as explained in Section 1.3.2, there is over-representation of cases and we may obtain biased estimates of the association between secondary phenotypes and genetic variants or biomarkers if this is ignored. In the literature several ad-hoc solutions have been initially proposed: testing for association only on cases, testing for association only on controls, or simply adjusting for the case-control status in the regression model used. However, none of these methods properly corrects the sampling mechanism and the relationships between the primary phenotype, the secondary phenotype, and the genetic variants. Figure 1.3 illustrates 6 scenaria for the possible relationships between a SNP (G), a secondary phenotype (X), the case-control

status (Y) and the sampling process (S) (Monsees et al., 2009) .

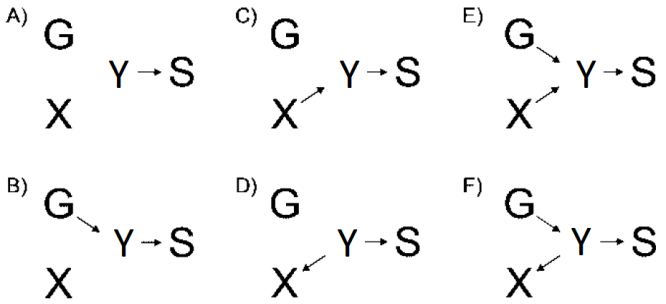

Figure 1.3: Directed acyclic graphs representing the different relationship between a SNP G, a secondary phenotype X, a primary phenotype Y and the sampling process S. Here we assume that the sampling process depends only on the primary phenotype. A: There is no association between G, X, and Y, B: G influences Y, C: X influences Y, D: Y influences X, E: G and X influences Y, F: G influences Y and Y influences X. Bias will occur when estimating the effect of G on X in scenarios B to F. Scenarios D and F induce reverse causality problems and are not considered in this thesis.

In general, the primary and secondary phenotypes are expected to be correlated as they are collected on the same individual. In this case and for the multiple cases studies we consider in this thesis, the sampling distribution of the secondary phenotypes in the study sample is not representative of its distribution in the general population. A naive analysis which ignores this feature will lead to biased estimates of the effect of genetic variants on the secondary phenotype.

In case-control studies, inverse-probability-weighting approaches (Richardson et al., 2007; Monsees et al., 2009) have been proposed to deal with the sampling mechanism on the primary phenotype. Inverse-probability-weighting is an alternative to regression-based adjustment of the outcomes. This approach focus on the idea that individuals have unequal probabilities to be sampled. To correct for bias induced by the sampling mechanisms and obtain proper estimates in the population of interest individuals are weighted by their inverse probability to be included in the study. Therefore, giving a larger weight to individuals having a small probability to be included in the study. This approach is very efficient and simple but can create imbalance if weights are not properly computed. Therefore, proper modelling of the probability of being included in the study is needed. For family studies, the use of inverse-probability-weighting methods is challenging (Rodríguez-Girondo et al., 2018) because we need to compute the probability that a family is recruited in the study which is not available for our studies. Alternatively, the retrospective likelihood approach can be used. The retrospective likelihood, as explained in Section corrects 1.3.2, implicitly for the ascertainment as follows:

$$P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y}, \mathbf{S}) = \frac{P(\mathbf{S} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}) P(\mathbf{G}, \mathbf{X} \mid \mathbf{Y})}{P(\mathbf{S} \mid \mathbf{Y})} = P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y})$$

Thus under the assumption that the ascertainment depends only on the case-control status, we require modelling: $P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y})$ which using the Bayes rule is further re-written as:

$$P(\mathbf{X}, \mathbf{G} \mid \mathbf{Y}) = \frac{P(\mathbf{X}, \mathbf{Y} \mid \mathbf{G})P(\mathbf{G})}{P(\mathbf{Y})}$$

To study the effect of a genetic variant on secondary phenotype we need to explicitly model the correlation between X and Y. In the case where X and Y are both continuous then we can assume multivariate normal.

In our case, we have a mix of outcomes and thus we build the joint distributions using latent variables. This model is presented in detail in Chapter 2.

1.5 Correlated predictors

An important complication in the discovery of biomarkers associated with a phenotype and in the prediction of disease phenotypes is the complex correlation structure of features and the fact that most phenotypes are associated with a combination of biomarkers from various omic sources. As a matter of fact, only few diseases are single gene disorders and most of the disease are due to a complex combination of biological and environmental factors.

In this case, separate analysis of omic features using univariate regression models is not advisable. The non-independence of the separate statistical tests and the strong multiple testing correction penalty, due to the high number of omic features, limit the ability of univariate models to discover new associations. A simple solution is then to include all variables of interest as covariates in a multiple regression model. Let Y be the quantitative phenotype of interest and $X = (X_1; \ldots; X_K)$ a set of biomarkers. The linear model is written as:

$$\mathbf{Y} = \alpha + \sum_{k=1}^{K} \beta_k \mathbf{X}_k + \epsilon$$

with α the intercept of the model, β_k the effect size of the k^{th} biomarker, and ϵ the vector of residuals. Even though this method is rather simple, due to the high dimensionality and presence of collinearity in omic features, this model, often, cannot include the whole set of predictors. This is only possible in combination with regularized regression such as the ridge regression (Hoerl and Kennard, 1970) or lasso regularization (Tibshirani, 1996) or variable selection techniques. The lasso regularization is widely used as this approach also forces numerous effect sizes, depending on the size of the penalty, to be equal to zero leading to simpler models. However this approach is also sensitive to the correlation structure between omic features. It will randomly select one variable from a set of strongly correlated features leading to models hard to reproduce and to understand the underlying biological mechanisms. Ridge is appropriate for correlated features, but does not shrink coefficients in the model towards zero leading to complex models which are

hard to interpret. Finally, the presence of strong correlations increases the possibility of confounder effects and therefore the quantity of false positive associations.

A solution to overcome these issues is to incorporate the correlation structure in the model. In particular, a two-step procedure may be followed. In the first step, the goal is to identify groups of closely related variables. In the second step, this grouping information is used in the statistical analysis. For the first step, there are two possibilities: either use a Biology- or a data-driven approach. In the biology-driven approach the idea is to incorporate the knowledge about pathways, i.e. groups of single omic features working on a specific cellular function. However, this approach has some limitations: first, the relationship between the variables from the same pathways are not always linear and thus these variables might not be correlated, and second, our knowledge about pathways is still incomplete and therefore we incorporate only a partial picture of the data structure in the model. For the data-driven approach the idea is to empirically derive the correlation structure of the data and to apply clustering algorithms in order to identify clusters of strongly correlated variables. Network construction methods are discussed in Chapters 5 and 6.

Once the groups of omic features have been identified, we can proceed with the statistical analysis of step 2. As far as testing is concerned, one approach is to use the grouping information for dimension reduction and then test for association between omics and phenotypes. In particular, this can be done either by selecting the most "important" variable of the cluster based on a specific criterion or by using a summary measures such as the mean or the first principal component (Pearson, 1901). Association between phenotypes and the summary measures can then be tested in order to detect the group of variables related to these phenotypes. Advantages of such an approach are that it is straightforward to summarize clusters of features in one variable and that we considerably reduce the number of tests to be performed. A downside of this approach is that the use of summary measures makes it hard to reproduce the original results. In prediction models, the grouping information can be incorporated in the statistical analysis, via group penalization methods (Yuan and Lin, 2006; Jacob et al., 2009; Simon et al., 2013; van de Wiel et al., 2014). Thereby, only the most important groups needed for the prediction will be selected leading to more stable and easier to interpret prediction models. Note that recently methods have been developed which allow features to be in different clusters allowing the data to mimic more closely the reality as many biomarkers are part of several pathways. These methods are presented and discussed further in detail in Chapters 5 and 6.

The approaches discussed above have been investigated and applied for a single omic source. Extensions to the simultaneous analysis of multiple omic sources are not straightforward. The integration of multiple omic sources is challenging due to the existing heterogeneity between the different omic sources, in terms of dimensionality, scale, and possible differences in noise structure. These complications make the identification of relationships between them difficult and no state-of-the-art method for integrating different omic sources is available. In Chapter 6 we discuss possible approaches to incorporate several omic sources in prediction models.

1.6 Outline of the thesis

The rest of this thesis contains 5 chapters. As explained in the previous sections, the analysis of omics data can be complicated by several sources of correlation. Table 1.1 shows the correlation structures handled in the different chapters. All chapters may be read in any preferred order, as they have been published or submitted independently. However, we feel that the order in which this thesis has been organized enhances the understanding of the links between the topics of the different chapters. In particular, even though Chapters 2 and 3 both focus on the ascertainment correction for secondary phenotypes in family study designs, we feel that Chapter 3 should be read after 2 as the methodology applied in Chapter 3 is developed in Chapter 2. Chapter 4 may be regarded as the link between the first part (Chapters 2 and 3) and the second part of the thesis (Chapters 5 and 6). In Chapter 4, we still consider family study designs but we use an alternative approach to test for associations between omics data and disease phenotypes, namely correlation networks. Correlation networks are the key ingredient in Chapters 5 and 6, where novel network-based approaches are presented to perform prediction of outcomes in population-based studies with one and multiple omic sources, respectively.

Dependencies	Individuals	Outcomes	Features
Chapter 2	✓	✓	
Chapter 3	✓	✓	
Chapter 4	✓		✓
Chapter 5			✓
Chapter 6			✓

Table 1.1: Overview of the between units dependencies modelled in the different chapters of this thesis.

In Chapter 2, we present a novel approach for the analysis of secondary phenotypes in multiple-cases family studies, i.e. families selected for having at least a certain number of cases. In particular, we work under the retrospective likelihood approach and explicitly model the dependence of the secondary phenotypes and the case-control status using a latent variable approach. A shared random effect is assumed to model the association between the primary and secondary outcome. For the analysis of the primary and secondary phenotypes properly chosen mixed-effects models are used to address the familial relationships. The performance of this approach is empirically evaluated in terms of bias, type I error and robustness to model misspecification. We use the LLS to illustrate the methods.

Chapter 3 explores the performance of the approach developed in Chapter 2 in another very common design for family studies, the proband design. In this design, families are selected based on the phenotype of specific family members. For the analysis of primary

phenotypes in this case, the conditioning on proband approach is typically considered. This approach has been recently applied for the analysis of secondary phenotypes (Greenwood et al., 2007; Turetsky et al., 2015) collected under the proband design. However, the dependency between the primary and secondary phenotypes is not modelled. Therefore, in the context of proband designs we compared our method presented in Chapter 2 with the conditioning on proband approach. Both methods are compared in terms of bias in the estimates of genetic effect on secondary phenotypes and heritability in an extensive simulation study. The relative performance of the two methods has been illustrated on electroencephalography (EEG) data from the LFLSAD.

Chapter 4 presents weighted gene coexpression analysis (WGCNA) with family data using a meta-analysis approach. To take into account between family variation, we proposed to perform the WGCNA on each family separately and to combine the obtained results using a meta-analysis approach. This approach was compared with two ad-hoc applications of WGCNA: (1) ignoring the family structure and (2) decorrelation of the gene expression via use of mixed models. To compare their performance, each method was applied on the simulated dataset provided by the Genetic Analysis Workshop 19 (GAW19).

Chapters 5 and 6 present network-based approaches for the prediction of health outcomes using omic sources. In particular, Chapter 5 investigates the combination of network analysis to identify clusters of correlated variables and the incorporation of this information in group penalization in order to improve stability and prediction ability of prediction model using a single omic source. We have considered several combinations of network analysis methods and group regularization approaches. Specifically, as network construction approaches we have used WGCNA and gaussian graphical modelling and as group regularization approaches we have considered: the group lasso, sparse group lasso, and adaptive group ridge. These combinations are compared with common regularization approaches such as lasso, ridge, and elastic net in terms of prediction ability and variable selection via double cross-validation. All methods have been applied to two different datasets: (1) the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study where gene expression and metabolomics at baseline were used to predict BMI after 7 years follow-up, and (2) the publicly available breast cancer cell line pharmacogenomics dataset in which we predict the response to treatment of cell lines using gene expression.

Chapter 6 extends Chapter 5 by allowing for the integration of several omic sources in the prediction model. To combine both datasets in the prediction model several approaches have been investigated. The first approach is to perform the network analysis separately on each omic source and combine them in the group regularization approach to predict the outcome. Even though this approach is robust to heterogeneity between omic sources, it is not able to capture interactions between omic sources. The second approach we considered is to apply the approach of Chapter 5 on the stacked omic sources. Even though this approach, potentially, can identify groups of related features coming from various omic sources, it is highly sensitive to difference in scale and heterogeneity between sources. The last approach, is relatively close to the first approach with one ad-

ditional step consisting of building a new network of summary measures of the clusters obtained in the first step. Clusters containing related summary measures from different omic sources are obtained, and clusters containing features from both omics sources can then be derived from them. Finally, the group penalization is performed using an overlapping group lasso approach allowing the variables to be in different groups. The performance of these approaches has been assessed using metabolomics and gene expression data from the DILGOM study and CNV and gene expression from the breast cancer cell line pharmacogenomics dataset to predict the same outcomes as in Chapter 5.

R codes of the methods developed in Chapters 2-3, Chapters 4-5, and supplementary materials of the different chapters can be found at the git repository:
https://github.com/RenTissier