New approaches in *omics* data modelling

Author: Lara Nonell Mazelon

TESI DOCTORAL UPF / 2019

THESIS SUPERVISOR
Dr. Juan R. González
Barcelona Institute for Global Health

THESIS TUTOR
Dr. Robert Castelo Valdueza

a la meva família, LLIT

Agraïments

Voldria agrair a tots aquells sense els que aquesta tesi no hauria estat possible, tant per les aportacions científiques com les personals, no menys importants.

En primer lloc, gracias Juan Ramón, gracias por empujarme, hacer esto posible y estar ahí en los momentos clave, conduciendo esta tesis y ampliando mis miras científicas.

A la meva família, Ignasi: per ser-hi, sempre i per sempre. Per aquest recolzament incondicional en tot el que faig, amb el teu suport present. Ton i Lluís: heu hagut de patir una mare ausent però ara ja tornaré a estar present i podrem organitzar, visitar i fins i tot us arrebataré el "mando" de la tele amb propostes no futboleres, tremoleu! Passi el que passi us estimaré sempre. Continuant la línia familiar...Mama: esta determinación por aprender y hacer siempre cosas, aunque alguien pueda pensar que no corresponde; son tu herencia, de la que estoy muy orgullosa. Iván: siempre a mi lado en nuestro pequeño núcleo familiar, aunque no te lo creas tú también lo hubieses podido hacer. Papà: allà on siguis, hi ha una part de mí també teva. Meme: la luna es tuya. Ignasi i Pepita: des de la vostra parcel·la, el suport també incondicional ha estat molt important.

Alex Sànchez: gràcies per per introduir-me en aquest fascinant món de les *òmiques* i apostar sempre per mi. En aquest sentit, també estic agraïda a l'Eduard Barrabés, que ens va presentar en algun moment del passat. Vull també agrair a la Malu Calle, haver-me donat l'oportunitat d'ensenyar *òmiques* a nivell universitari, gaudeixo molt amb les classes, alhora que aprenc.

I ja centrats a l'IMIM, voldria agrair a molta gent que m'ha donat suport durant tots aquests anys. Montse Torà: per donar-me algunes llicències per a poder tirar endavant aquest projecte. Extenc l'agraïment a la institució i en especial a la Balbina, el Joaquim i el Ferran. Em quedo amd tots aquests anys passats al SAM, on hem après tantes coses juntes... Eulàlia: per compartir tants dies i encoratjar-me sempre a tirar endavant, la força de les mames treballadores i cansades és imparable, ja et trobo a faltar! Marta: successora sempre lluitadora i resolutiva al lab, junt amb el

Miquel, fareu un tàndem imparable; gràcies pel suport i per escollir-nos. Magda: gràcies per recolzar-me en el dia dia, i fer-me sentir com una bona ”jefa”, fins i tot en els últims temps caòtics, et trobaré a faltar... Gràcies a tots, també a la resta de companys i companyes que van passar pel SAM, han estat uns anys fantàstics i sempre ens quedarà el Nespresso i el Valor. La nova creació del grup de biologia computacional human compta amb la incorporació de la Júlia i el Joan (per fi un grup bioinformàtic!) gràcies per fer-ho fàcil i per tots els comentaris i suggerències. No em puc descuidar dels informàtics, Joan Marc, Josep i Alfons, gràcies pel suport i ajut en la programació en contenedors i gestió de cues, sense la qual no hauria pogut fer el primer article i altres anàlisis.

A les companyes i excompanyes Cris, Silvia H., Lara, Eva J., Kiko, Blanca, Gonzalo, Mar, Ana Ferrer, Anna Puiggros, Andrea, Sílvia R., Laura...; molt més que ciència compartida en alguns moments. També vull donar gràcies a tots aquells ”usuaris” del SAM, per transmetre saviesa i permetre'm aprendre contínuament...això ho puc fer extensible a tantes persones de l'IMIM, inclòs el comitè, PRBB i l'hospital que no sé si hi cabria tothom...

Vull també agraïr al Robert Castelo i Xavier Basagaña els consells científics i pràctics rebuts tots aquests anys.

Miriam: Sempre vas dir que ho faria i *here I am*, más vale tarde que nunca...Gràcies per ser-hi sempre també, sempre endavant!

Per últim, vull agrair a tots tots aquells companys, alumnes, familiars i amics a qui he escatimat hores i a tots aquells a qui no m'he acostat gaire en aquests temps per falta d'això mateix, TEMPS. Calamars i especialment brujis meves (tornaré!), A, C, B, S, M, P, V, F, O, N, L,...

Segur que oblido algú que mereix ser agraït i que amb la pressió del moment ha saltat temporalment de la meva memòria, gràcies i disculpes!

Abstract

The breakthrough in the technological field has allowed the extraction of large amounts of the so-called *omics* data. The analysis and integration of this type of data by means of advanced statistical and bioinformatics methods will allow the improvement in the management of diseases. The diversity and complexity of *omics* data has encouraged the development of hundreds of new statistical methods to meet this objective. Therefore, having the appropriate methods to accommodate different data distributions and modelling complex data structures becomes essential. This thesis presents advances in three directions in this regard. First, the study of several methods to assess non-linear associations which is relevant when assessing the effect of environmental exposures (i.e exposome) on complex diseases. The study is accompanied by the development of the R package *nlOmicAssoc*. Second, the simplex distribution is proposed to analyse methylome data since this distribution properly fits beta values that are generated in this type of studies. The extension to generalized linear models with simplex response is also proposed. Lastly, an R package, *HOmics*, has been developed to incorporate *a priori* biological knowledge into association studies by using Bayesian hierarchical models. It also implements methods to model the dependence between *omics* data, enabling data integration.

Resum

L'avenç en el camp tecnològic ens ha permès obtenir grans quantitats de les anomenades dades *òmiques*. L'anàlisi i integració d'aquesta mena de dades mitjançant mètodes estadístics i bioinformàtics avançats ha de permetre la millora en el maneig de les malalties. La diversitat i complexitat de les dades *òmiques* ha incentivat el desenvolupament de centenars de nous mètodes estadístics per a complir amb aquest objectiu. Per tant, és primordial disposar de mètodes que acomodin les distribucions adequades i modelin estructures de dades complexes. Davant d'això, aquesta tesi presenta avenços en tres direccions. En primer lloc, l'estudi de diferents mètodes per a analitzar associacions no lineals, molt rellevant en estudis d'associació entre exposicions mediambientals (i.e. exposoma) i malalties complexes. Aquesta anàlisi va acompanyada del desenvolupament del paquet de R *nlOmicAssoc*. En segon lloc, es proposa utilitzar la distribució simplex per analitzar dades metilòmiques, donat que aquesta distribució ajusta els valors beta generats en aquesta mena d'estudis. També es formula l'extensió a models lineals generalitzats amb resposta simplex. I per últim, el paquet de R *HOmics*, que incorpora coneixement biològic als estudis d'associació mitjançant models Bayesians jeràrquics. També implementa mètodes per modelar la dependència entre dades *òmiques*, permetent la integració de dades.

Resumen

El avance en el campo tecnológico nos ha permitido obtener grandes cantidades de los llamados datos *ómicos*. El análisis e integración de este tipo de datos mediante métodos estadísticos y bioinformáticos avanzados posibilitan la mejora del manejo de las enfermedades. La diversidad y complejidad de los datos *ómicos* ha incentivado el desarrollo de centenares de nuevos métodos estadísticos para cumplir con este objetivo. Por tanto, es primordial disponer de métodos que acomoden las distribuciones adecuadas y modelen estructuras de datos complejas. Esta tesis presenta progresos en tres direcciones al respecto. En primer lugar, el estudio de diferentes métodos que analizan asociaciones no lineales, muy relevante en estudios de asociación entre exposiciones medioambientales (i.e. exposoma) y enfermedades complejas. Este análisis va acompañado del desarrollo del paquete de R *nlOmicAssoc*. En segundo lugar, se propone utilizar la distribución simplex para analizar datos metilómicos, dado que esta distribución ajusta los valores beta generados en este tipo de estudios. También se formula la extensión a modelos lineales generalizados con respuesta simplex. Y por último, el paquete de R *HOmics*, que incorpora conocimiento biológico a los estudios de associación mediante modelos Bayesianos jerárquicos. También implementa métodos para modelar la dependencia entre datos *ómicos*, facultando la integración de datos.

Preface

I have been for more than ten years now connected to the *omics* world. It all started after a long travel around South America. I had been working in computing consultancy for several years but had the feeling I had betrayed my mathematical background. That was why, after giving birth to my first son Ton and thanks to a good friend's connection, I started my PhD at the Universitat de Barcelona in statistics, data analysis and biostatistics. There, Àlex Sànchez introduced me to the fascinating and innovative -at that time- microarray world. This technology was combining many interesting things: biological concepts, a rather complicated technology and biostatistical (this was still pre bioinformatics era) algorithms to analyse the data with the help of R programming. Everything was new and challenging for me.

A couple of years later, I gave birth to Lluís and right after that obtained my DEA on classification methods for microarrays. After that, I decided not to go on with my phD for several reasons but continued in the *omics* world as the analysis responsible of the Servei d'Anàlisi de Microarrays (SAM, IMIM). Since then, I have analysed data of several thousands of microarrays in my office, located at the first floor of the PRBB. For those chance occurrences that happen in life, at the end of the corridor, people from ISGlobal are having their offices. Juan Ramon, while passing through my office, used to knock my door to try to convince me to end up my PhD by finally writing a thesis. He insisted even after I had given up. But, things are sometimes unpredictable and here I am, thanks to his persistence, presenting you my third baby, this thesis.

The *pregnancy* has been a hard time, combining a full time job, some teaching and this research; but satisfaction requires always an effort, it is not the same to climb a mountain and reach the top after suffering the heat and fatigue for hours than going for a cool and relaxing walk. I have to admit that I have learned many things, technical skills but also my proficiency in the art of *resiliencing*...

Thesis overview

This thesis, entitled *New approaches in omics data modelling*, is organised in three sections containing several chapters to help the reader follow the work that I have carried out during these four years of thesis development.

Introduction Comprises a general introduction, the hypotheses and associated objectives of this thesis. Besides, it contains a list of the data sets used.

General introduction Includes the current state-of-the-art of the *omics* world and those technical and statistical concepts related to their data, useful to follow the posterior analyses and results.

Hypotheses Outlines the hypotheses set out through the thesis span.

Objectives Describes the objectives addressing the previous outlined hypotheses.

Data sets Lists the data sets that have been used in the three articles included in this thesis.

Results States the development of the three objectives, with the details of the papers written and packages developed in R.

Non-linear models to link exposome with *omic* data

Are methylation beta-values simplex distributed?

HOmics: Bayesian hierarchical models to analyze *omics* data with prior biological knowledge

Discussion Comments on the three papers and how I addressed the hypotheses and objectives through this study. It also discusses general issues found while developing this thesis. Enumerates the conclusions obtained in this thesis at the end of this part.

List of Figures

List of Tables

Contents

Part I

Introduction

Chapter 1

GENERAL INTRODUCTION

1.1 A world of *omes* and *omics*

In the continuous deciphering of biology and disease, there has been a great breakthrough of *omics* technologies in recent years, enabling the study of biology to an unprecedented detail. The term *omics* is derived from the Latin suffix *ome*, meaning mass or many. Thus, *omics* involve many measurements per endpoint. Genomics, the study of the genome, was the first studied *ome* and was coined by the geneticist Tom Roderick over a beer at a meeting held on the mapping of the human genome in 1986 [Yadav, 2007]. Since then, many other *omics* have been formulated. Transcriptomics, epigenomics, proteomics and metabolomics enable the study of the transcriptome, epigenome, proteome and metabolome, respectively. The exposome or the measurement of environmental exposures over the human lifespan [Wild, 2005], is one of the latest *omes* added to the list of such disciplines. Other *omes* such as the spliceosome, diseasome, phenome, microbiome or integrome have recently been coined.

Many of the *omics* can be considered as biological layers, being closely related to the central dogma of molecular biology, stated by Francis Crick in 1957 [Cobb, 2017]. The modern interpretation of the central dogma specifies that genetic information contained in the DNA flows into its final protein product in two steps: transcription and translation. First, the

transcription transforms the codifying DNA in messenger RNA (mRNA) and then the translation transforms the mRNA in the final protein, with the information to perform a specific function. At each step of this transformation, a different *omics* is highlighted (Figure 1.1).

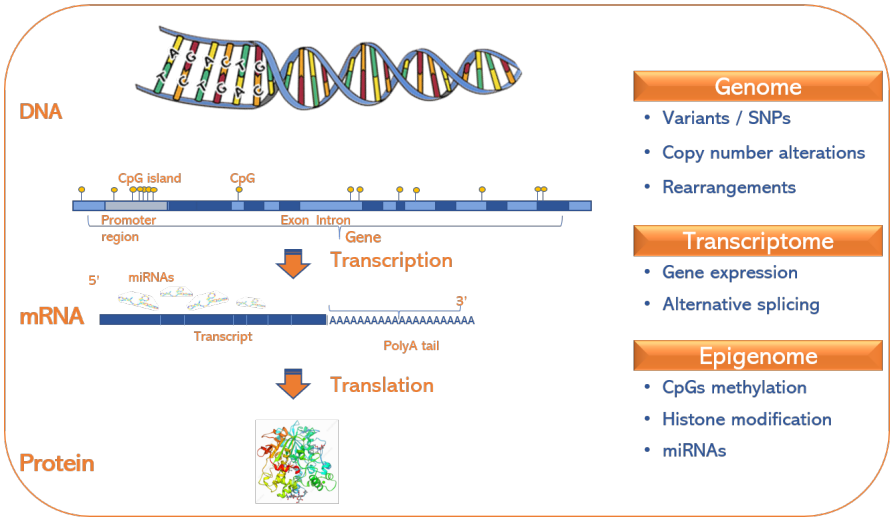


Figure 1.1: *Omics* related to the central dogma of molecular biology. At each step of the DNA transformation to protein, *omics* and their components are depicted. Applications in the most relevant *omics* data are also listed.

The following subsections contain a description of the relevant *omes* to this thesis. I will give details on the genome, transcriptome, exposome and epigenome. Particularly on the latest with the expanded section on the methylome, since its data are closely studied in this work. After that, I will introduce the experiments where the *omics* data are obtained through high-throughput technologies and preprocessed. Next, subsections dedicated to model these data and their association with phenotype will follow, with some statistical details. To conclude the introduction, a few remarks on biological knowledge and *omics* interdependences.

1.1.1 The genome

The *genome* is the complete sequence of DNA (Deoxyribonucleic acid), composed of only four nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Every living organism has a genome, usually confined in the nucleus of the cells, which can be frightfully long. For instance, the human genome contains more than 3 billion (10^9) nucleotides and is longer than 2 meters (Figure 1.2).

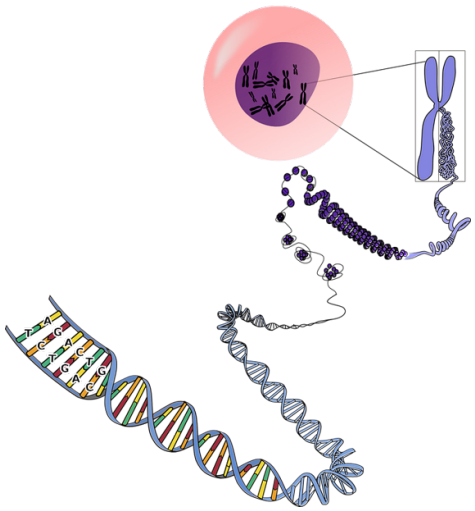


Figure 1.2: DNA location and composition. DNA is located in the nucleus of the cell and arranged in chromosomes.

Genes, included in the DNA sequence, are the basic unit of heredity in a living organism and are composed of coding regions (exons) and non-coding regions (introns). All living things depend on genes. To pick a formal definition of a gene, it is the fragment of genetic information corresponding to one protein with a specific function [Alberts et al., 2002]. Gene structure is depicted in Figure 1.3. At present, sixteen years after the human genome was sequenced and despite the ongoing ENCODE and GENCODE projects [Consortium et al., 2012, Davis et al., 2017b], there is still controversy surrounding the number of genes it contains,

which seems to be between 19,000 and 22,000 [Ezkurdia et al., 2014, Willyard, 2018].

The human genome is organized into 22 paired chromosomes, plus the sexual pairs composed of two X chromosomes in females and one X and one Y chromosomes in males. Each gene has a specific location in this chromosomal structure, known as locus.

Some failures can occur in the DNA, such as small mutations, copy number (CN) alterations, translocations, inversions or loss of heterozygosity (LOH). These errors can be produced during DNA replication or caused by some external factors such as radiation or other chemicals and have a direct impact on disease. An important concept in this field is the single nucleotide polymorphism (SNP), which is an alteration of just one nucleotide in the DNA sequence present in at least 1% of the population. Genome wide association studies (GWAS) have associated many of these SNPs with disease ([catalog, 2019]). To study genomic alterations several studies can be designed, like case-control, trios, population studies or GWAS; among others.

1.1.2 The transcriptome

The *transcriptome* comprises the complete set of transcripts, i.e. genes after transcription. As the central dogma of molecular biology declares, the transcription step starts by selecting from the genome the exons of a gene and copies them into the mRNA transcripts in the so called splicing process, with the help of the RNA polymerase enzyme. This enzyme, together with one or more transcription factors, binds to the promoter region, and starts the process at the transcription start site (TSS) located at the beginning of the 5' start of the gene (Figure 1.3). It then reads the DNA sequence until the stop codon, located at the 3' UTR boundary, producing a complementary, antiparallel RNA strand called a primary transcript. In this process, the base uracil (U) replaces thymine.

After transcription, the region comprised between the 5' UTR start and the 3' UTR end is called coding sequence (CDS). The promoter region is therefore a critical component of the gene. As obtained from the ENCODE

project, only a small portion from the complete human genome represents protein-coding genes or exons (1-3%, [Consortium et al., 2012]).

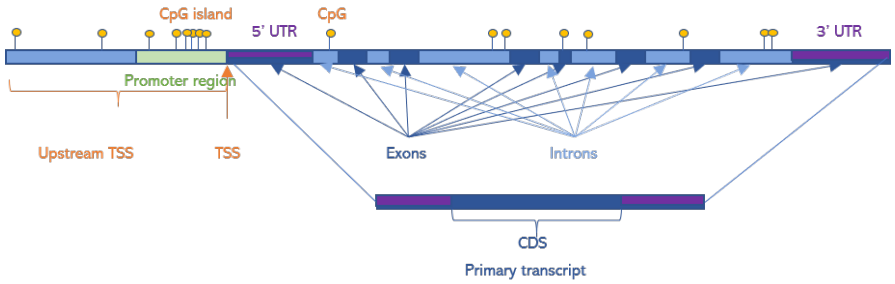


Figure 1.3: Components of a gene. Gene is arranged in exons and introns, with the transcription starting site and the promoter region. CpGs can be found at any genomic position. After transcription, only exons are selected, composing the primary transcript.

The transcriptome reflects the genes that are being actively expressed under specific conditions. Therefore the transcriptome is the set of all mRNA molecules in the assessed sample, which is in fact composed of a mixture of the cell transcriptomes that comprise the sample. This might mask individual cell transcription, as cells are heterogeneous. Single cell analysis can be in this sense a way to overcome this issue. This technique can also be applied to other *omics*.

Alternative splicing is a regulated process during gene expression where there is a selection of the available exons to be included in the final processed transcript in the mRNA. That results in a single gene having different transcripts coding for multiple proteins, which can in turn have different functions [Matlin et al., 2005].

In general, transcriptomic studies measure the expression of transcripts, genes or exons and compare these measures between conditions. The typical experimental design is a case-control study, where the gene expression of cases are compared against control samples to elucidate which genes in the cases are upregulated or downregulated versus controls, allowing this way the discovery of biomarkers. The decision of what kind of control sam-

ples to choose or how many samples to process are part of the experimental design, one of the key issues in transcriptomics studies [Draghici, 2003].

1.1.3 The epigenome

While the *epigenome* is the set of reversible alterations that affect the expression but without altering the DNA sequence, epigenomics is a wide field that studies several alterations. These include:

Histone modification: Histones are proteins found in the cell nucleus that pack and order the DNA in structural units called nucleosomes. A histone modification is a post-translational modification which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation. These modifications can impact gene expression by altering chromatin structure or recruiting histone modifiers [Arivaradarajan and Misra, 2018].

miRNAs: miRNAs are small non-coding RNA molecules (containing around 22 nucleotides) highly conserved across species, that act at a transcriptional and post transcriptional level. miRNA bind to mRNA fragments by complementarity, silencing this way mRNA expression so they can not be translated into proteins by ribosomes. miRNAs can also be considered a part of the transcriptome as they are in fact being transcribed at the same time as the rest of the transcripts.

DNA methylation: The set of methylated DNA is known as the methylome and will be remarkably detailed in the following section.

Although these alterations may sound terrifying for a non-expert, the fact is that they are essential for the normal development and regulation of gene expression in mammals [Esteller, 2011].

1.1.4 The methylome

The *methylome*, as part of the epigenome, has important roles in genetic regulation and is composed of the set of methylated DNA. Methylation

involve the addition of a methyl group to the 5^{th} carbon of a cytosine. Methylated sites are found primarily at CpG dinucleotides but are also found at non-CpG sites (CpA, CpT, and CpC). There are $\sim$ 28 million CpG sites in the human genome, and 70–80% are methylated in normal, healthy cells [Nair et al., 2018].

CpG islands are regions with a high frequency of CpG sites. Although there is no consensus about the definition for CpG islands, they are usually defined as regions of larger than 200-500 bp that have guanine cytosine content greater than 50-55%. Many genes in mammalian genomes have CpG islands placed in their promoter regions, located at the 5' start of the genes (see Figure 1.3). Up to 60% of CpG islands are in these promoter regions and about 70% of proximal promoters contain a CpG island. However, CpG islands that are not in promoter regions can also be found within coding regions and noncoding regions of genes, which may be targets for *de novo* methylation in cancer and aging [Nair et al., 2018]. Humans have around 27,000 CpG islands, most of them close to promoter regions [Saxonov et al., 2006]. In cancer, CpG island promoters are prone to hypermethylation, silencing in consequence associated genes. In contrast, the rest of the genome in cancer is subject to hypomethylation and gene activation of cancer-associated oncogenes [Nair et al., 2018].

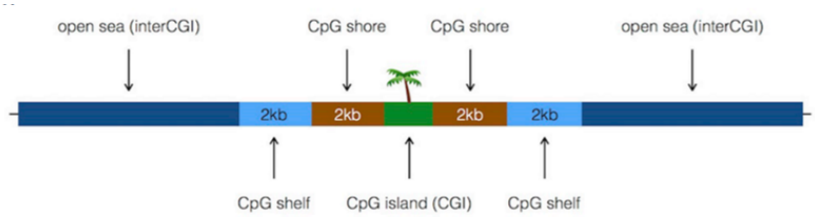


Figure 1.4: Scheme of the UCSC CpG annotations as described in the R package *annotatr*: Associating genomic regions with genomic annotations. Position of islands, shores, shelves and open sea, relative to the CpG.

Outer CpG island boundaries, other position relative definitions are given (Figure 1.4). CpG shores are defined at 2 kb extension upstream and downstream of the CpG island, less any CpG islands. The CpG shelves are

a further 2 kb extension upstream and downstream of the furthest upstream and downstream boundaries of the CpG shores, less any CpG island and shore annotations. The complement of the CpG islands, shores, and shelves make up the open sea annotation [Cavalcante and Sartor, 2016].

Methylation is assessed usually using case-control design studies, where some affected samples are compared against some controls. As in transcriptomics, the decisions taken during the experimental design are crucial for the validity of the results.

1.1.5 The exposome

The *exposome* is the measurement of the set of exposures that a human undergoes during a lifespan. It was first proposed by [Wild, 2005] to encompass the totality of human environmental (i.e., nongenetic) exposures from conception onward, complementing the genome; and was developed to obtain a more holistic picture. The exposome contains several overlapping domains of nongenetic factors contributing to disease risk, including a general external domain (social, urban environment, climate factors), a specific external domain (specific contaminants, lifestyle factors, tobacco, occupation), and an internal environment (metabolism, gut microflora, inflammation, oxidative stress) [Wild, 2012, Vrijheid et al., 2014].

Measurement of the environmental exposures is important for determining whether an environmental agent causes actual harm. Tools for measuring the exposome are aimed at assessing exposures that include external measures, biomonitoring, and measurements of biological effect. It might be hard to quantify the exposome which can retrieve variables very distinct in nature, categorical or continuous; possibly depicting many different distributions.

1.1.6 Other *omes*

In addition to the preceding defined terminologies, there are other terms in this world of *omes*. I will just mention some of them, that can directly or indirectly be related to the *omics* deconstruction performed in this work.

The *phenome* is the set of all phenotypes expressed by a cell, tissue, organ, organism, or species; where the phenotype is defined as any of the observable characteristics or traits [Braun et al., 2019]. This is relevant in the framework of this thesis as we will care about association between phenotype and the different *omics* under different circumstances.

The *proteome* is the entire set of proteins that is produced or modified by an organism. Proteomics covers the exploration of proteomes from the overall level of protein composition, structure, and activity. It is an interdisciplinary domain that has benefited greatly from the genetic information of various genome projects, including the Human Genome Project. Proteomics generally refers to the large-scale experimental analysis of proteins and proteomes, but often is used specifically to refer to protein purification and mass spectrometry. It is an important component of functional genomics [Arivaradarajan and Misra, 2018].

The *metabolome* refers to the complete set of small molecules (metabolites) and products of metabolism found in a biological sample, i.e. cell, tissue, organ or organism. The metabolites can be endogenous (produced by the organism such as amino acids, organic acids, sugars, vitamins, antibiotics, etc) or exogenous (such as drugs, contaminants or additives). Metabolomics encompasses the scientific study of chemical processes involving metabolites [Arivaradarajan and Misra, 2018]. Whereas transcriptomic and proteomic analyses reveal the set of gene products being produced in the cell, metabolic profiling can give an instantaneous snapshot of the physiology of that cell.

The *microbiome* comprises all of the genetic material within a microbiota (the entire collection of microorganisms in a specific niche, such as the human gut). This can also be referred to as the metagenome of the microbiota [www.nature.com, 2019].

One of the challenges of systems biology and functional genomics is to integrate different *omics* information to provide a better understanding of cellular biology. This is sometimes known as *integromics*, the last of the *omics* cited in this introduction [Van Steen and Malats, 2015].

1.2 *Omics* experiments

Once the *omes* and *omics* are clarified, let us follow an experiment to assess any of such data. To study any of the *omics* in a specific application, such as transcriptomics differential expression, there are several decisions to be made in the different phases that compose the study (Figure 1.5). The first step and probably the most critical one is the experimental design. At this step, the hypothesis needs to be clearly established. Besides, several decisions related to: sample size, control election, sample processing technology and data analysis strategy have to be decided. Each of the choices will condition the final results.

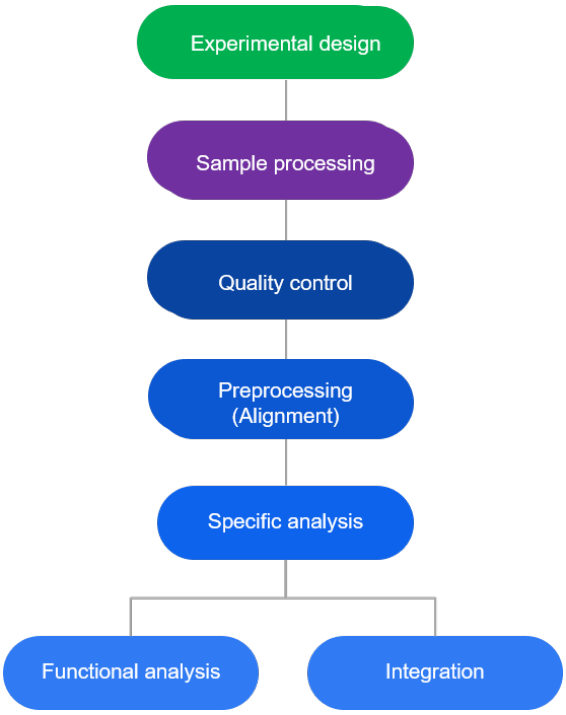


Figure 1.5: Generic workflow for *omics* processing. In green steps previous to sample processing. In lilac steps related to sample processing. In blue steps related to data handling.

1.2.1 *Omics* technologies

Each *omic* can be measured through different technologies, some of them are low-throughput such as the real time reverse transcription polymerase chain reaction (RT-PCR), fluorescence in situ hybridization (FISH) or multiplex ligation-dependent probe amplification (MLPA); but I will focus on the high-throughput, concordant with the mass measurements defined by the *omics* suffix.

Since the end of the Human Genome project, most of high-throughput technologies acquire data from the whole genome. The choice of the technology to measure specific *omics* data depends on the information required, that is, the application. There are several applications of each technology. I will briefly introduce the microarrays and the next generation sequencing (NGS) and finally give a list of the applications. I do not intend to exhaustively explain them but rather to explain the essentials for the understanding of this thesis.

Microarrays

Here we get to the field I know the best. After several years, I have formulated the microarray definition as follows: A microarray is a collection of biomolecules (or probes) in micro format, arranged in rows and columns on a solid support. The nature of these biomolecules define the type of microarray. In this sense, some examples are: the DNA microarrays, to assess the transcriptomics or genomics; the array comparative genomic hybridization (aCGH); the protein arrays or the methylation arrays.

Although there are a myriad of different DNA array technologies, I will explain briefly the distinction between one colour and two colour array designs. In the one colour array each sample is hybridized in an array and in the two colour array, two samples are hybridized together, thus in a competitive manner. The sample for analysis is usually dyed in red (Cy5 dye) whereas the control sample (or a pool of control samples) is dyed in green (Cy3 dye). Experimental designs for the two colour technology are usually more complex and the so-called dye swap design is often applied.

The typical sample preparation steps before it is hybridized onto the

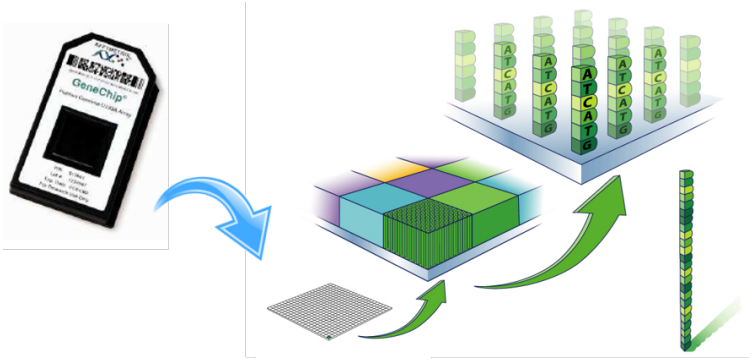


Figure 1.6: Affymetrix microarray structure. Content of the microarray. Each spot contains thousands of pre-spoted probes. Adapted from the original Affymetrix catalog information.

array are: amplification, fragmentation and labelling with a fluorochrome. The microarray market leader in transcriptomics is Affymetrix (now Thermo Fisher Scientific), with the old 3' family of arrays, where location of probes where at the 3' end of the gene; or the whole transcript designs, the Gene 1.0 and 2.0 ST series, the Transcriptome arrays or the latest Clariom developments, that enable the alternative splicing analysis. For genomic microarrays, even though Affymetrix has some hybrid arrays with probes to detect copy number and probes to measure SNPs, Illumina leads the market with its SNP arrays. In the methylomics field and on account of the sodium bisulfite treatment, Illumina is number-one best-seller with the Infinium MethylationEPIC (EPIC) array, containing around 850,000 probes; the popular Infinium HumanMethylation450 (450K) and the previous Infinium HumanMethylation27 BeadChip array (27K). The Illumina methylation arrays use two different types of two colour probes:

Type I, where unmethylated and methylated signals are measured by different beads in the same colour channel (green and red). This is original from the 27k array.

Type II, where the unmethylated (red) and methylated (green) signals are measured by a single bead. Added to 450K and EPIC designs.

In general, there is a huge difference in length, location and density of probes in the arrays. Measurements obtained from microarrays are continuous, corresponding to the intensity of signal of the stimulated fluorochrome through the scanner. These data are then preprocessed. Proportions are usually taken (SNPs or methylation) and often log_2 transformed to obtain the final values for posterior analyses.

Next generation sequencing

Like in microarrays, there are many different technologies comprised in the term NGS. The common objective of all of them is to obtain the sequence of (A, T/U, C, G) nucleotides of the chosen fragment of DNA or RNA to examine. Sanger sequencing was the founding method of DNA sequencing, which enabled the sequence of the human genome in the early 2000's. In 2005 the first sequencer, Roche's 454 appeared to produce high-throughput sequences. Since then, the technology has rapidly evolved into the so-called second and third generation sequencers, with their singularities [Kchouk et al., 2017]. Currently, the market leader is no doubt Illumina, with the largest number of samples processed by their platforms, from the old HiSeq 2000 to the newest NovaSeq 6000. And now, with the recent acquisition of Pacific Biosciencies, Illumina adds long read sequencing to the range of applications offered. NGS can interrogate genomics, transcriptomics, epigenomics and metagenomics in different applications. In general, the NGS sample processing involves the following steps: library preparation, amplification and sequencing.

To assess genomic alterations, the common approaches are: whole genome sequencing (WGS), whole exome sequencing (WES) or targeted panel sequencing. The three strategies enable the following applications: variant calling, SNP genotyping, study of rearrangements and copy number variants, among others. The most popular NGS transcriptome measurements are: RNA-seq, small RNA-seq or targeted panel sequencing; to study expression profiles, differential expression between genes, transcripts or smallRNA or alternative splicing. Sodium bisulfite treatment is the basis of methylation microarrays but also of the two popular methods to assess

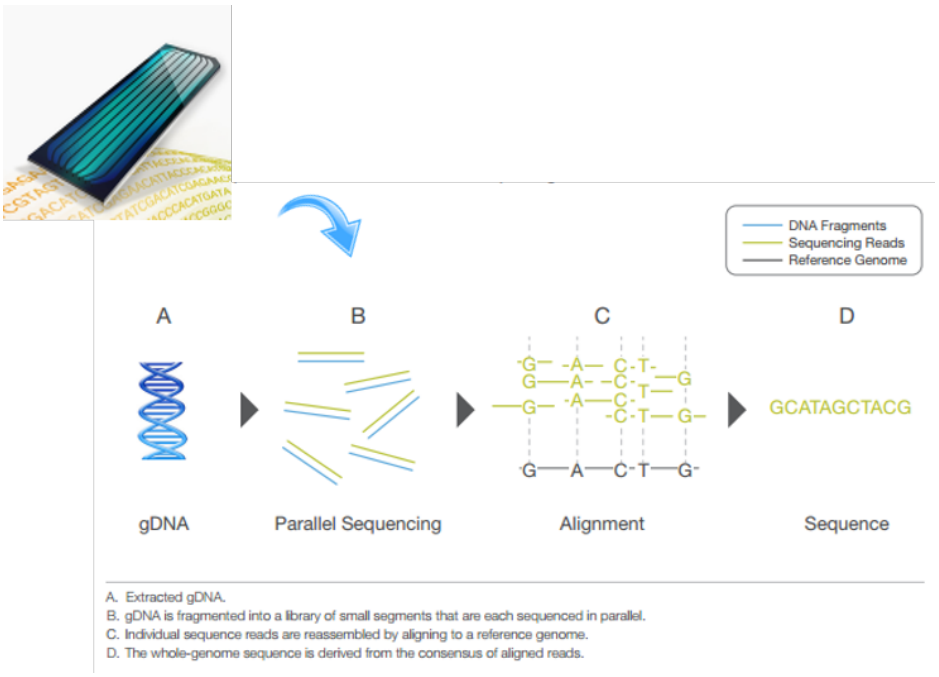


Figure 1.7: Scheme of genomic Illumina NGS. DNA is sequenced and aligned to obtain the fragments nucleotide sequence; using the Illumina flow cells. Adapted from www.illumina.com.

methylation, mainly on CpGs, in NGS: whole genome bisulfite sequencing (WGBS), considered the gold-standard for assaying DNA methylation and reduced representation bisulfite sequencing (RRBS), which combines digestion of genomic DNA with restriction enzymes and sequencing with bisulfite treatment in order to enrich for areas with high CpG content. RRBS only interrogates 6 - 12% of the human CpGs. Some other technologies have been developed such as the methylation-sensitive restriction enzyme bisulfite sequencing, which has the reduced sequencing requirements of RRBS, but significantly expands the coverage of CpG sites in the genome. [Bonora et al., 2019].

***Omics* applications**

	Application	Technology	
		Microarrays	NGS
Genome	SNP calling	√	√
	Variant calling	-	√
	Copy number variation	√	√
	Rearrangements	-	√
Transcriptome	Expression profiling	√	√
	Differential expression	√	√
	Alternative splicing	√	√
	small RNA detection	√	√
Methylome	CpG measurement	√	√

Table 1.1: Applications of microarrays and NGS according to each *omic* and technology.

The most common applications for microarrays and NGS are listed in Table 1.1. Even though it shows that most of the applications can be performed with microarrays, current NGS applications overcome those offered by microarrays in their detail offered, as NGS reports single nucleotide measurements, does not have probe selection or cross-hybridization bias and because it is in continuous evolution [Han et al., 2015]. In contrast, there is not much novel development done for microarray technology at the moment.

1.2.2 *Omics* data quality control and preprocessing

We have previously seen that each *omic* type of information can be assessed by different technologies. Each *omic* information, measured by a certain

technology, has a specific workflow to obtain the final data that will be analysed (Figure 1.5). This process is performed over the set of data samples one wants to analyse, once the data has been obtained from the technical platform. There are of course some technical parameters given by the instruments that allow also the evaluation of the data quality before any analytical step is performed.

Once data has been obtained from the platform, quality control of these data is required. Although methods applied in each situation may be very different, the initial quality control step has the objective to guarantee that the data have enough quality and specifically that there are no outliers among the samples. Several measures and graphics are standard to check for quality control in each analysis application such as box plots or MA plots for microarrays or coverage, the Phred quality score distribution or the total number of duplications for NGS.

The following step is the preprocessing, which encompasses several proceedings with the objective of bringing the data to comparable measures in the set of assessed samples. In the specific case of NGS, data are usually aligned to a reference genome or assembled. Preprocessing includes normalization, data transformations and unwanted variation adjustment, among others. There are different methods to normalize the data which include within sample and between samples methods. Variable transformations are typical in *omics* analysis being the most typical in this context the logarithm transformation. Once data are normalized, it is typical (and advised!) to examine how it aggregates. Hierarchical clustering, principal component analysis and multidimensional scaling are the most typical methods, which allow to detect batch effects and other kinds of unwanted variation. If detected, it can be adjusted in the analysis and even corrected using specific methods. I have to confess that, due to my professional bias, I am tempted here to go into the details, but I understand they are beyond the scope of this thesis. The reader can consult [Draghici, 2003, Hahne et al., 2010, Han et al., 2015] for more information. The important message at this point is that to achieve the final data for each combination of *omics* and technology, several manipulations on the original data are performed, conditioning their final shape and

structure.

1.3 *Omics* data and their distribution

After preprocessing steps, it is time to perform the data analysis, according to the settled hypothesis. Each *omic* data obtained by a certain technical platform and application has a specific distribution. In formal statistics, distribution functions are defined over random variables, and the distribution functions are in fact probability distribution functions. In the *omics* world and in the concrete context of this thesis, features measured across samples are considered as random variables. The nature of the distribution function depends on the numerical space where the random variable is defined. If the space is finite, then distribution will be considered as discrete, otherwise it will be continuous.

Assumptions made about the distribution of an *omic* random variable will condition the parametric methods that can be applied in subsequent analyses, such as in regression models fitting. Wrong assumptions can lead to unreliable results. To overcome this issue, one could think about using non parametric methods, which do not require distribution assumptions. However, they present other sorts of limitations ([Lumley et al., 2002]).

This is the list of the typical distributions obtained from *omics* data and some of the uses in the analysis of the *omes* world:

Continuous:

- Normal: the most common, used in transcriptome and exposome
- Log-normal: used in exposome and metabolome
- Logistic: very popular in generalized linear models, when the response variable is binary
- Beta: for proportional data, popular in the methylome data analysis
- Simplex: for proportional data, could be used also in the methylome data analysis

- Exponential: the natural distribution of microarray data with many low and some extreme values. It is usually log transformed

Discrete:

- Poisson: some of the earliest methods developed for NGS transcriptomics were based on this distribution
- Binomial: employed in the genome and exposome data analysis
- Beta-binomial: used in the NGS methylome analysis
- Negative binomial: used in transcriptome data analysis

Table 1.2 shows, based on Table 1.1, the type of data obtained in each *omic*. The main difference between microarray and NGS data is that microarrays measures continuous values whereas NGS returns reads, i.e. fragments of sequences that are usually counted by genomic regions.

1.3.1 Probability distributions

I will give details on some distributions that are subject of study in my thesis, related to the methylation data. Methylation beta-values are defined as proportions (methylated reads or probes versus total reads or probes) and therefore defined in the (0,1) interval. Their distribution can be skewed, often bimodal and sometimes have extreme peaks at 0 and/or 1. Usually, these data are analysed under the assumption that they follow a beta or a beta-binomial distribution. The simplex is however another option, for being defined for proportional data too.

Beta distribution

The beta distribution is a family of continuous probability distributions defined on the interval (0,1). It is the most popular from the distributions to adjust proportions [Kieschnick and McCullough, 2003] and is at present the natural statistical distribution model for microarray DNA methylation

	Application	Technology	
		Microarrays	NGS
Genome	SNP calling	Categorical	Count
	Variant calling	-	Count
	Copy number variation	[0,inf)	[0,inf)
	Rearrangements	-	Count
Transcriptome	Expression profiling	[0,inf)	Count
	Differential expression	[0,inf)	Count
	Alternative splicing	[0,inf)	Count
	small RNA detection	[0,inf)	Count
Methylome	CpG Beta-values	[0,1]	[0,1]
	CpG M-values	[0,inf)	[0,inf)

Table 1.2: *Omics* application measurements. For each *omics* and technology, the type of data generated to study each application are given.

measures [Teschendorff and Relton, 2018]. It belongs to the exponential family and its density function, with parameters μ and ϕ is:

$$p(x, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)(\Gamma(1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{1.1}$$

where Γ denotes the gamma function [Ferrari and Cribari-Neto, 2004].

Simplex distribution

The simplex distribution is defined in the (0,1) interval and belongs to the family of dispersion models. Considering the normal distribution with mean μ and variance σ^2 with the following probability density function:

$$p(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 \big/ 2\sigma^2} \tag{1.2}$$

Then the simplex distribution with location parameter μ and dispersion parameter σ^2 is defined as follows:

$$p(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi x^3(1-x)^3}} e^{-d(x;\mu) \big/ 2\sigma^2}, \tag{1.3}$$

where $d(x;\mu) = \frac{(x-\mu)^2}{x(1-x)\mu^2(1-\mu)^2}$ and $x \in (0,1), \mu \in (0,1)$ [Song, 2007].

Beta-binomial distribution

The binomial distribution is a discrete distribution, mixture of the binomial and the beta distribution, $Binom(p, n)$ where the binomial probability of success of each trial p follows a beta distribution $p \sim Beta(\alpha, \beta)$, for some shape parameters α and β. The mixture can be obtained by multiplying the two distributions and has the following density function:

$$p(m|n, \alpha, \beta) = \binom{n}{m} \frac{B(m+\alpha, n-m+\beta)}{B(\alpha,\beta)}, \tag{1.4}$$

where B is the beta function. For $\alpha = \beta = 1$ it is the discrete uniform distribution from 0 to n [Dolzhenko and Smith, 2014].

There is another parametrization of the beta-binomial distribution based on π and γ where $\pi = \frac{\alpha}{(\alpha+\beta)}$ and $\gamma = \frac{1}{(\alpha+\beta+1)}$

The parameter π is the equivalent to the binomial probability of success parameter, which can be interpreted as the average methylation level of a set of replicate samples. The parameter γ is called the dispersion parameter. Observe that the binomial distribution is a special case of beta-binomial distribution with $\gamma = 0$.

1.3.2 Distribution parameter estimation

Omics features can be regarded as random variables that follow a probability distribution. Distribution probability functions are defined by one or more parameters. To formulate the distribution formula, we will have to estimate its parameters. This concept, parameter estimation, refers to the process of using sample data to estimate the parameters, considered fixed and unknown, of the selected distribution. Several parameter estimation methods are available. The most popular methods used in life data analysis are least squares, method of moments (MOM) and maximum likelihood estimation (MLE). There are also Bayesian estimation methods that consider the parameter to estimate as a random variable and uses *a priori* knowledge to obtain the estimations. Bayesian estimation will be explained later in this essay, in the hierarchical regression model subsection. There are several properties of the estimators such as unbiasedness, efficiency, sufficiency and consistency that can be useful to decide which estimator better adjusts the data.

Method of moments

In a distribution, the moments are defined as the expected values of powers of the random variable under consideration. In this sense, the k order moment of a random variable X is $\mathbb{E}(X^k)$. Parameters can be estimated using these moments as they can be expressed as functions of the parameters of interest in a system of equations whose solutions are the estimates of those parameters.

Maximum likelihood estimator

The likelihood function is the probability function applied to a given set of observations as function of the parameters. MLE method obtains the maximum of the likelihood function, where the parameter of the distribution is treated as a random variable and is adjusted by the sample data. Usually the log function is applied to the likelihood function to transform the products into sums, which are easier to derive.

1.3.3 Overdispersion

When dealing with *omics* data, particularly obtained from NGS, where measurements are counts; overdispersion is a well-known concept. The intuitive idea behind is that dispersion of coverage (reads captured at a certain location) might be very big, affecting the measures obtained [Zhou et al., 2011]. Overdispersion basically violates the mean-variance relation induced from a proper probability model, as variance of sequence counts tends to be greater than would be expected, which prevents investigators from using a specific parametric distribution for the data. Overdispersion may also emerge from other data collection procedures, one of which is that the response variable is recorded as an aggregation of dependent variables [Song, 2007]. Methods to analyse *omics* data, particularly those dealing with NGS data, should take into account overdispersion. In fact, the main reason to use a negative binomial model in RNA-seq data analysis is because this essentially corresponds to an overdispersed Poisson model.

1.4 Modelling *omics* data

1.4.1 Association with phenotype

I have previously defined the phenotype as any observable trait in an individual. This includes different conditions and information on the characteristics and natural history of a disease or other clinical outcomes

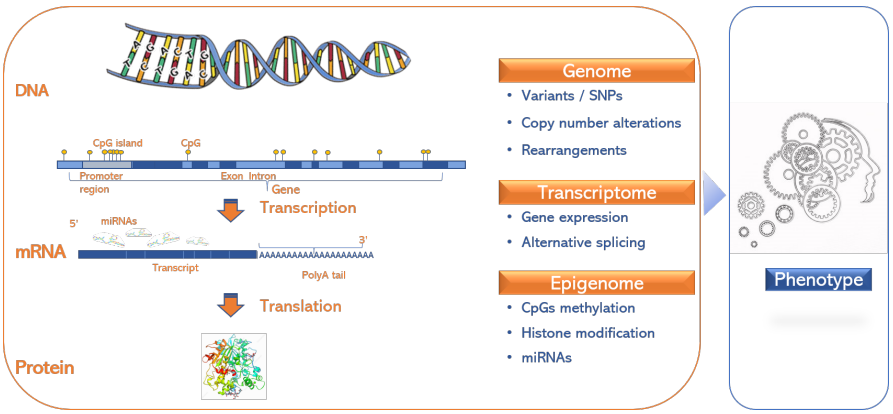


Figure 1.8: *Omics*, related to the central dogma of molecular biology, and the association with the phenotype.

like survival. *Omics* studies are ultimately focused in explaining these characteristics, where each of the different *omics* provide data that can be translated into clinically relevant information contributing in such wise to the identification of biomarkers. This can directly benefit in the management of human health which includes diagnosis, treatment, and prevention of disease (Figure 1.8). Ideally, a holistic picture of an individual can be obtained by analyzing data from multi-platform *omics* experiments combined with patients' clinical outcomes. Integromics can help us understand the complex biological processes that characterize a disease, as well as how these processes relate to the development of the disease.

In this framework, *omics* features such as gene expression, are treated as variables. Further than the nature of *omics* variables, we have to consider the nature of the condition variable -or response or outcome- we want to explain; and also the relationships that are rendered among variables. Such relationships depend on the sort of variables that are being studied. When the response variable is categorical, association with the features are usually assessed with a χ^2 test or Fisher's exact test if features are categorical. In continuous *omics* settings, Student's t-test, analysis of variance or covariance, or some non-parametric tests are used to test

association between features and a categorical phenotype. When the outcome is continuous correlations and regressions -linear models- are commonly employed. In this setup, relationships are usually assumed to be linear but other shapes of associations can be given and are in fact more realistic [Gasparrini et al., 2015, Xiao et al., 2017]. U-shape, J-shape and r-shape are different shapes that can adopt the associations between two variables [May and Bigelow, 2005]. There are several methods that cope with non-linear associations such as exposome-wide association study with natural cubic splines [Patel et al., 2010], multivariable fractional polynomial models [Royston et al., 1999], generalized additive models [Hastie and Tibshirani, 1986, Rigby and Stasinopoulos, 2005], regression trees [Breiman et al., 1984], random forest [Breiman, 2001] or neural networks [Venables and Ripley, 2002].

In the association of *omics* data with phenotype, regression models can be very helpful, as other covariates can be incorporated. This enables the adjustment of association coefficients in unbalanced designs but also for different sources of unwanted variation; leading to more accurate results. The simplest is the linear model, which can be extended to the generalized linear model. Other more sophisticated options include the hierarchical models.

1.4.2 The linear regression model

The linear regression model has the following formulation:

$$y_{ij} = \mu_i + \alpha_i x_{ij} + \sum_{k=1}^{K} \beta_{ik} z_{ijk} + \epsilon_{ij}, \tag{1.5}$$

where y is the is the outcome or response variable,μ is the mean of y, x is the condition to study, α is the condition coefficient, z are the covariates, β are the estimated coefficients in the K groups and ϵ is the error term. *i* represents the *omics* feature measurement, *j* the subject and *k* the covariates. In general, in our *omics* world, the *omics* feature is taken as the outcome and the condition is the phenotype but these roles are sometimes interchanged.

1.4.3 Generalized linear models

Generalized linear models (GLM) allow the response variable Y of the linear regression model to adopt distributions belonging to the exponential family distribution. In GLM, it is necessary to specify a so-called link function that describes how the mean of the response and the linear predictors are related. For instance, when the response variable is discrete, the logit transformation and the binomial link is used in a logistic regression. In this context, the beta regression model and the simplex regression model, that have two parameters; are fitted with two link functions, the first to link the mean and the responses whereas the second links the precision (in beta regression) or dispersion (in simplex regression) parameter to other regressors [Faraway, 2016].

1.4.4 Quantile regression models

Quantile regression was introduced in 1978 [Koenker and Bassett Jr, 1978] to expand the potential of linear models, that compare the mean among different groups, and focus in the quantile sample distribution. The regression coefficients are computed by minimizing the sum of weighted absolute residuals [Beyerlein, 2014]. Quantile regression fits specified percentiles of the response, to accommodate the different distribution shapes and can potentially describe the entire conditional distribution of the response. This model can be used when the response variable is skewed or asymmetric.

1.5 Biological knowledge

Biological knowledge has been summarized over the years by uncountable researchers and is accessible through a number of different sources, most of them publicly available. This knowledge can be used in the analysis of *omics* data to obtain meaningful results, such as in functional analysis, that helps to draw conclusions. The main resources used in this thesis are summarized in the following list:

- Genome browsers:
  - University of california, Santa Cruz (UCSC, [ucsc, 2019])
  - Ensembl ([ensembl, 2019])
- Relevant projects:
  - ENCODE and GENCODE ([Davis et al., 2017b])
  - The cancer genome atlas (TCGA, [research network, 2019])
  - BLUEPRINT ([consortium, 2019])
- Functional databases:
  - Gene ontology (GO, [GO, 2019])
  - Molecular signature database (MSigDB, [mSigDB, 2019])
  - Kyoto encyclopedia of genes and genomes (KEGG, [KEGG, 2019])
  - Comparative toxicogenomics database (CTD, [CTD, 2019])

The biological knowledge is usually incorporated in several steps of the *omics* data analysis. In this regard, genome browsers are often accessed in the initial phases, the alignment and also the integration analysis. Project data are used to test and validate biological hypotheses but are also an important source of data. Functional databases are generally exploited at the end of the analysis to give a biological sense to the *omics* analysis results.

1.6 *Omics* interdependences

As we have previously seen, *omics* features are not independent, as genes can be coexpressed or interacting in epistasis to impact the phenotype. Moreover, these dependencies can be present among different *omics* features. For instance, a mutation produced in a locus of a gene or a methylated CpG located in its promoter can affect the expression of the gene. Standard uni-*omic* models may be not accurate enough and a multi-*omics* approach are preferred [Wu et al., 2019]. In this sense, the incorporation of biological knowledge can help to elucidate the different processes. One possibility to incorporate such knowledge is the use of hierarchical models, detailed in the following subsections.

1.6.1 Hierarchical regression models

A hierarchical model is one that is written in terms of stages or sub-models. This are sometimes called multilevel models and also regarded as mixed-effect models [Niemi, 2016].

Bayesian versus frequentist approach

As we have previously seen, in "standard" statistics everything not observed is treated as fixed but unknown (variables, parameters, etc.) and will be estimated from the sample data from a population with no previous assumptions. This is, in proper statistical language, the frequentist approach.

In contrast, in the Bayesian approach, parameters are treated as random variables and uses prior information about the parameters (prior distribution) to modulate distribution and obtain final values (posterior distribution). To that end, we can use the Bayes theorem to update a prior distribution with previous knowledge such as data observations and obtain the posterior distribution (Posterior $\propto$ Likelihood x Prior). Bayesian inference can be used for parameter estimation, where the estimation is asymptotically equivalent to MLE estimates, for prediction and for hypothesis

testing. The Bayesian assumptions are more natural than the frequentist approach for hierarchical models because they include prior information in the estimations [Gelman et al., 2013, Grosskopf, 2019, Niemi, 2016].

A remarkable difference in the interpretation of statistical significance between the frequentist and Bayesian approach is that the later does not use p-values in the hypothesis testing and returns the credible interval, which is equivalent to the frequentist confidence interval, although interpretation is different. The interpretation of a parameter estimation 95% credible interval in the Bayesian approach is that given our observed data, there is a 95% probability that the true value of the parameter falls in the credible interval. In frequentism, on the other hand, as the parameter is considered a fixed value and the data (including the bounds of the confidence interval) are random variables, the frequentist confidence interval is equivalent to saying that here is a 95% probability that the calculated confidence interval from experiments of this sort of data, will contain the true parameter value.

A hierarchical Bayesian model is really the combination of two things: a model written in hierarchical form that is estimated using Bayesian methods. The sub-models combine to form the hierarchical model, and the Bayes theorem is used to integrate the pieces together and account for all the uncertainty that is present [Allenby et al., 2005]. There are several programs that enable the Bayesian approach such as BUGS, WinBUGS, OpenBUGS , JAGS [Lunn et al., 2009, Plummer et al., 2003], or stan [Carpenter et al., 2017]. We decided to use JAGS as the bayesian programming environment but they are all very similar.

JAGS

JAGS (Just another Gibbs sampler) is a program developed by Martyn Plummer [Plummer et al., 2003] that implements Bayesian inference based on Gibbs sampling. Gibbs sampling is a Bayesian inference technique that uses experimental data to generate samples from a certain posterior probability density function. In this sense, JAGS implements a Markov Chain Monte Carlo (MCMC) approach to estimate the parameters. MCMC is a generic term indicating an algorithm that samples

from probability distributions (sampler) by constructing a Markov chain (which is a model for a sequence of random variables) that has the desired probability density as its equilibrium distribution. The construction of the Markov chain is combined with Monte Carlo integration, a technique to approximate the expected value of a function of a random variable integral by means of the samples. The approximation becomes more accurate as soon as the number of samples increases. JAGS uses hierarchical models to instruct the sampler and models are written using a dialect of BUGS [Coro, 2017, Niemi, 2016]. A typical Bayesian application has the following steps:

1. Define the model with a likelihood and the priors
2. Choose the number of Markov chains that will be fitted and the initial values
3. Update of the model through the burn in phase, with a specific number of iterations
4. Estimate model coefficients obtained for the posterior distribution by sampling
5. Check the convergence through the behaviour of chains and effective samples

Chapter 2

HYPOTHESES

In the evolving world of *omics*, where different measurements and technologies are involved, there are still several scientific open questions related to statistical data modelling that should be addressed to get accurate results. While an uncountable number of methods have been developed to analyse *omics* data, often the underlying nature of the data and the suitability of modelling methods are overlooked. In this sense, we have focused on three topics related to *omics* data distribution and their association with phenotype. These are the initial hypotheses proposed in this thesis:

- Association between exposome and continuous *omics* features are assumed to be linear, conditioning posterior analyses. However, non-linear associations are more realistic in some scenarios making necessary the use of advanced statistical methods that can cope with these types of relationships.
- Methylation beta-values are analysed assuming standard distributions but the use of inappropriate data models can lead to biased results. The simplex distribution, that accommodates proportions, is suitable to model methylation beta-values. Besides, the simplex regression using generalized linear models can be assessed in this context to associate beta-values with phenotype.

- Regression models are broadly used to assess the association of *omics* data with phenotype. Biological knowledge, summarised through the years in different public databases, can be incorporated into regression models to improve results in *omics* association studies.

Chapter 3

OBJECTIVES

The global aim of this thesis is to study how advanced statistical methods can improve the results obtained in *omics* association analyses. To this end, we study the use of new distributions and models than can accommodate real data shapes and relationships. For that, we defined three main objectives, that address our hypotheses:

- Objective 1: to study non-linear association between the exposome and other *omics* data
  - to study several methods that can cope with non-linear relationships
  - to benchmark these methods
  - to develop a tool for the research community to address the non-linearity issue in *omics* settings
- Objective 2: to study how methylome data distribute and which is the best regression model to assess the association with phenotype
  - to analyse the distribution of methylation beta-values from different real data sets, obtained from microarrays and NGS
  - to assess the suitability of using different regression models including beta, simplex, linear and quantile models for the association with phenotype

- Objective 3: to develop a package implementing Bayesian hierarchical models which allow the incorporation of prior biological knowledge and the integration of different omics data.
  - to develop an R package that can be applied in diverse situations
  - to compare results obtained through Bayesian hierarchical regression models with standard methods

The list of objectives related to the *omics* and the association with phenotype are depicted in Figure 3.1.

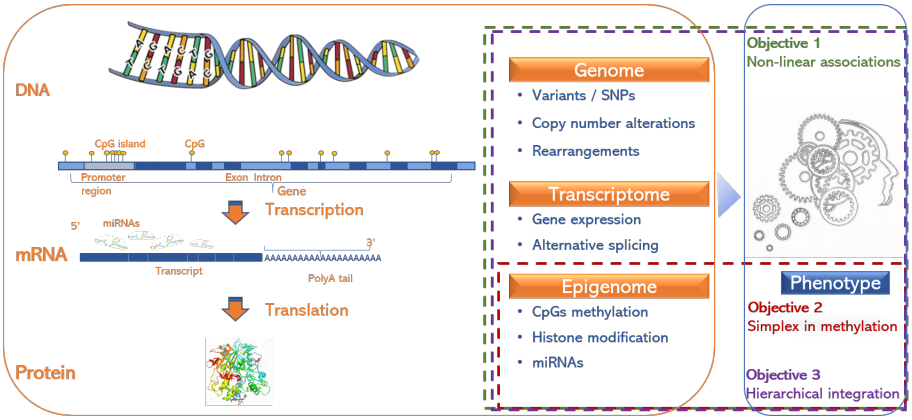


Figure 3.1: Objectives of the thesis in relation to *omics* and phenotype. The first objective is the study of non-linear associations between *omics* data and continuous phenotype in the context of environmental exposures. The second objective is related to epigenomics, where methylomics data are studied and simplex generalized linear models formulated in the association with phenotype. The third objective involves the integration of previous knowledge to any *omics* data in the association with phenotype, using Bayesian hierarchical models.

Chapter 4

DATA SETS

Before getting into the results, I would like to summarize the data sets that have been used in this thesis, most of them of public repositories. Some of the data sets are already detailed in the articles written.

4.1 HELIX-INMA

In the context of the european project Human Early Life eXposome (HELIX) [Vrijheid et al., 2014, Maitre et al., 2018], the INMA (Infancia y Medio Ambiente) mother-child cohort [Guxens et al., 2012] has been utilized in the *Non-linear models to link exposome with omic data* article to generate simulations. In this data set a total of 237 environmental factors were assessed in 122 mothers during pregnancy through questionnaires, geospatial modelling, and biological monitoring. The INMA transcriptome and methylome data sets were used to illustrate the non-linear association. INMA transcriptome was obtained from 308 children using the Human Transcriptome Array (HTA) 2.0 (Affymetrix, USA) whereas the INMA methylome obtained with the 450k methylation array (Illumina, USA), measured in the same 308 individuals.

4.2 GEO data sets

The Gene Expression Omnibus (GEO, [Barrett et al., 2012]) contains thousands of data sets related to high-throughput experiments. It was originally created to serve as a microarray data repository under the MIAME (minimum information about a microarray experiment, [Brazma et al., 2001]) standards but now it contains also NGS and other forms of genomic data. Data sets are accessible through a unique Series identifier and for the studies that contain more than one platform SuperSeries are create containing as many Series of data as platforms considered. The data sets of this repository analysed at any step during this thesis are, by objective:

4.2.1 Second objective: Methylation data analysis

GSE50660

Epigenome-wide microarray association study in peripheral-blood DNA in 464 individuals who were current (n = 22), former (n = 263) and never smokers (n = 179). This research was performed on the Illumina 450k microarray.

GSE116339

Epigenome-wide microarray association study performed with the Illumina EPIC array. In this study, DNA from the blood of 659 individuals who were exposed to polybrominated biphenyl (PBB) in the 1970's from the Michigan PBB Registry.

4.2.2 Third objective: Bayesian hierarchical model

GSE77269

This Series contains methylation data of 60 samples of patients suffering from hepatocellular carcinoma with portal vein tumor thrombosis (PVTT). The study contains matched adjacent normal (Normal), primary tumor

(PT) and PVTT samples from 20 HCC patients obtained with the Illumina 450k microarray. It belongs to the SuperSeries GSE77276, containing RNA-seq data and copy number information as well.

GSE117931

This SuperSeries contains two data sets of 37 peripheral blood mononuclear cell samples obtained from systemic sclerosis (N = 18) and normal controls (N = 19). Series GSE117929 contains the methylation data ExpressionSet profiled with the Illumina HumanMethylation450 BeadChip array whereas the GSE117928 Series encloses the expression profile obtained with the Illumina HumanHT-12 microarray. Both data sets are used in the vignette of the *HOmics* package to show the integration between CpGs methylation and the closest genes.

4.3 Methylome resource

The methylome resource [rnbeads.org, 2018] was established after applying the method *RnBeads* [Müller et al., 2019] to some of the largest public reference data sets that are currently available for WGBS, RRBS and for the Illumina methylation microarrays. Two data sets were downloaded from this resource, described below:

4.3.1 RRBS data of 188 cases suffering Ewing Sarcoma

This study assesses DNA methylation associated with Ewing sarcoma, a bone cancer primarily affecting children and young adults, using reduced representation bisulfite sequencing. In addition to tissue samples, healthy mesenchymal stem cells (MSCs), MSCs affected with Ewing sarcoma and Ewing cell lines were also included in the study. A total of 188 RRBS samples were included in the study.

4.3.2 WGBS data of 81 blood sample methylomes from the BLUEPRINT project

Whole genome bisulfite sequencing profiles were generated for 81 different cell types obtained from blood samples in the BLUEPRINT project framework, which was created with the aim to further understand how genes are activated or repressed in both healthy and diseased human cells. Among others, primary monocyte and neutrophil cell samples from healthy donors were profiled.

Part II

Results

Chapter 5

NON-LINEAR MODELS TO LINK EXPOSOME WITH *OMIC* DATA

Non-linear models to link exposome with *omic* data
Lara Nonell, Xavier Basagaña, Lydiane Agier, Martine Vrijheid, Rémy Slama and Juan R González
Submitted to BMC Bionformatics the 20th February 2019, under second revision

5.1 Abstract

Background

There is a growing interest in elucidating the relationships between the exposome and global molecular profiles obtained from different *omic* technologies. Current exposome-wide association studies (ExWAS) assume linearity but association between *omic* data and environmental exposures require methods to deal with possible non-linear relationships along with the multivariate nature of the exposome. Here, we systematically assess existing methods to address both issues. These include: exposome-wide association study with natural cubic splines, multivariable fractional polynomial models, generalized additive splines models, generalized additive models using boosting, regression trees models using the deletion substitution addition algorithm, random forest and neural networks. We evaluated the performance of the proposed methods under two different situations. First, a comprehensive simulation study was conducted to measure models' performance on different scenarios, varying the type of non-linear relationships, the correlation among exposures and the effect size. Then, two real data sets were analysed to evaluate their behaviour when assessing exposome-*omic* association with non-linear relationships.

Results

Our results show that multivariable fractional polynomials had the best performance in the explored scenarios. This is robust across the different shapes of associations, including linearity. In real data tests however, there was a large variability obtained by each model regarding our evaluation measurements. Therefore, in order to facilitate real data analyses, we have implemented all the studied methods in an R package, `nlOmicAssoc`, that allows standard Bioconductor data type objects to deal with omic data (i.e. ExpressionSet or SummarizedExperiment).

Conclusions

Multivariable fractional polynomials show good performance when modelling non-linear associations between *omic* data and multiple exposures. This methodology can also be applied to analyse other studies having multiple continuous predictors such as the case of endophenotypes or multiple intermediate risk factors.

5.2 Background

Over the last decade, epidemiological studies have started investigating the health effects of the exposome, the set of environmental exposures experienced over the human lifespan [Wild, 2005, Maitre et al., 2018]. For example, studies linked the exposome with complex traits such as type 2 diabetes mellitus [Patel et al., 2010], blood pressure [McGinnis et al., 2016] or lung function [Agier et al., 2019]. Recent European projects such as EXPOsOMICs [Vineis et al., 2017] or HELIX [Vrijheid et al., 2014] are interested in the search for the relationships between exposures and global profiles of molecular features obtained from different *omic* technologies including transcriptomics, epigenomics, proteomics and metabolomics.

Omic data analyses typically compare feature profiles between two or more groups of subjects. For instance, one may be interested in finding genes that are differentially expressed or methylated between exposed and non-exposed individuals. The statistical methods used in such situations are often based on the t-test or on linear models [Smyth, 2004]. Generalizations based on regression methods have been developed for the case where one aims to link omic data with a continuous variable. As an example, elucidating the relationship between copy number alterations and the transcriptome enables the identification of regulatory mechanisms leading to abnormal gene expression [Solvang et al., 2011]. In such cases, one should pay attention to the shape of associations between variables, as non-linear relationships are common in biological processes. Therefore, techniques that rely on the linearity assumption may produce invalid results [May and Bigelow, 2005]. Vandenberg et al. [Vandenberg et al., 2012]

provide several examples of well-understood biological mechanisms explaining how hormones and endocrine disrupting chemicals can produce non-linear relationships between dose and response in cells and tissues. Considering environmental exposures, temperature was associated with the incidence of childhood hand, foot and mouth disease with non-linear relationship [Xiao et al., 2017]. Actually, temperature is known to affect mortality according to a U-shaped relationship [Gasparrini et al., 2015].

Due to the dimension of the exposome it is necessary to use more complex models than those used to analyse associations between *omic* features and a single exposure. Fave et al. have used the coinertia analysis (CIA), a multivariate method to analyse two tables, to link environmental exposures and gene expression levels in about 1000 individuals of Quebec [Fave et al., 2018]. As in the case of principal component analysis, CIA is a descriptive methodology that does not provide a list of statistically significant exposures linked to *omic* features and does not easily allow to control for potential confounding factors. An alternative is to use regression-based methods that can incorporate variable selection procedures and also adjust for confounding factors. This has already been performed in association studies linking the exposome to a single health outcome [Agier et al., 2016]. In consequence, linking the exposome with *omic* data would benefit from methods that accommodate both non-linear relationships and variable selection procedures.

Methods that admit non-linear associations have already been proposed in different settings. In the particular field of microarrays, Qu and Xu [Qu and Xu, 2006] recommended a method that searches for clusters based on orthogonal polynomials under a multivariate Gaussian mixture model. This was modified later by a method that uses a stochastic expectation-maximization algorithm for detecting quantitative trait-associated genes [Zhan et al., 2011]. Natural cubic splines have also been used to capture time course significance [Storey et al., 2005] and linear models for microarray data (limma, [Ritchie et al., 2015]), probably the most extended method to analyse microarray and RNA-seq data, also enables the use of splines when dealing with a continuous outcome. Other methods to detect non-linear relationships between microarray data and continuous

variables include non-parametric non-linear correlation [Chen et al., 2010] or sinusoidal models [Li-Ping et al., 2014], among others.

Generalization to a scenario of non-linear association between *omic* data and a large number of continuous variables (such as the exposome) can be decoded by using variable selection algorithms which deal with problems associated to high dimensionality (e.g. multiple testing). There are several statistical methods to address this issue. Nonetheless, their performance has not been systematically assessed in the context of exposome-*omic* data analysis. We therefore conducted a comprehensive simulation study based on realistic assumptions to compare different methods that have been proposed for capturing multivariate non-linear associations.

In this paper we evaluated different methods by extensive simulation studies and in real data sets. Simulations were conducted under several settings, considering different numbers of true predictors, shapes of variable associations (including linear associations as well), correlation levels of true predictors (i.e. exposures truly affecting the outcome) and coefficients of determination. In addition, real data sets from two different frameworks were assessed to study the effect of the exposome in two *omic* data sets: exposome-transcriptomics and exposome-methylation from the HELIX project [Vrijheid et al., 2014]. We have developed an R package, `nlOmicAssoc`, that empowers the user to apply any of the assessed methods. The package combines the ability to model non-linear variable association to the typical omic continuous outcome and graphic tools. It allows dealing with missing data and provides a common interface to the most popular methods available in R to assess non-linear association at large scale, also using standard Bioconductor data types to encapsulate omic data (e.g. ExposomeSet).

5.3 Methods

Methods to assess association between omics and multiple continuous exposures

We propose to assess the association between a continuous variable Y (e.g. one omic feature) and multidimensional X (e.g. exposome) by using several statistical methods. These methods can capture linear and non-linear effects, all of them enclose an algorithm to select variables. All analyses were performed in R (version 3.4.3). For each method, a brief description, the R package used are given. Methods will be hereinafter referred to as the given acronym (in parentheses).

Exposome-wide association study with Natural Cubic Splines regressions (ExWASsp)

The Exposome-wide association study (ExWAS) [Patel et al., 2010] usually relies on linear regression models fitted independently for each covariate with correction for multiple comparisons. Here, we allowed for the use of natural cubic splines (ExWASsp). For simulation purposes, natural cubic splines with 3 segments were handled, using the R package `splines` [R Core Team, 2016]. In this particular case, a natural cubic spline is a segmented curve in k polynomials of order 3. For a fixed spline with $k = 3$, we have followed the algorithm: 1) adjust univariate models with cubic splines for each predictor; 2) select significant variables using the likelihood ratio test (LRT) applying Benjamini-Hochberg (BH) multiple comparisons to adjust p-values ([Benjamini and Hochberg, 1995]) and 3) adjust a multivariate cubic splines model with selected variables.

Regression Trees model using the algorithm Deletion Substitution Addition (partDSA)

Classification and Regression Trees (CART) [Breiman et al., 1984], a binary recursive partitioning algorithm, allows the exploration of the individual contribution of various covariates as well as their interactions for

the purpose of predicting outcomes. Each node is split using the best split among all variables and the end product is a decision tool allowing to group individuals based on their variable values. The `partDSA` package [Molinaro et al., 2010] applies the deletion-substitution-addition algorithm to extend the CART [Molinaro and Lostritto, 2010]. `partDSA` provides a recursive partitioning tool for prediction when numerous variables jointly affect the outcome. Besides generating *and* statements, partDSA explores and chooses the best among all possible *or* statements, typical from CART models, empowering the method to build a parsimonious model with both conjunctions. The method uses the minimum percent difference to choose the best partition at each step. Cross-validation is employed in order to select the best model.

Multivariable Fractional Polynomial (MFP) model using stepwise

Fractional polynomials (FP) models, as introduced by Royston & Altman [Royston and Altman, 1994] and modified by Royston, Ambler & Sauerbrei (1999) [Royston et al., 1999], are useful to preserve the continuous nature of the covariates in a regression model, but having some or all non-linear relationships. Multivariable fractional polynomials (MFP) adjust a fractional polynomials (FP) model by combining backward elimination of covariates that are not statistically significant and iterative examination of the polynomial form of all continuous covariates. `mfp` package [Ambler and modified by Benner, 2015] was used, which works as follows. At each step, it on one side interrogates the significance for each variable using the closed test and on the other side it constructs the most complex permitted FP model for each continuous covariate while fixing the current functional forms of the other covariates. The algorithm attempts to simplify it by reducing the number of variables. The best-fitting FP is obtained for each covariate and the algorithm terminates when no more covariates are excluded and the functional forms of the continuous covariates do not change anymore. The closed test procedure is applied as a sequence of tests in each of which the correct Type I error rate for each component test is approximately maintained. For simulation purposes, two

models were fitted, one with 4 df (MFP) one with just one (MFP1df).

Generalized Additive Splines model using backfitting (GAM)

Generalized Additive models (GAM) [Hastie and Tibshirani, 1986] were introduced as a natural extension of generalized linear models (GLM), by replacing the linear form by a sum of smooth functions. This non-parametric function can be estimated in a flexible manner using a cubic spline smoother by applying the backfitting algorithm, which was specifically developed along with GAM. `mda` package [Leisch et al., 2016] was used for simulation purposes. However, for the real data analyses and in the implementation of the `nlOmicAssoc` package, the `gam` package [Hastie, 2018] was used to fit a multiresponse additive model, fitted by adaptive backfitting using smoothing splines. The procedure fits n additive models, but the same amount of smoothing (df) is used for each term. Then the method chooses between omitting the term (df=0), or including it either as a linear term (df=1) or as a term fitted by smoothing spline ($|\text{df}| > 0$). The model selection is based on an approximation to the generalized cross-validation criterion, which is used at each step of the backfitting procedure. Once the selection process stops, the model is backfitted using the chosen amount of smoothing.

Generalized Additive Model using Boosting (GAMboost)

A Generalized Additive Model (GAM) is fitted using a boosting algorithm based on component-wise univariate base-learners, which are simple regression estimators with a fixed set of input variables and a univariate response. Base-learners used in this case were P-splines with a B-spline basis. Package `mboost` [Buehlmann and Hothorn, 2007] was used to fit GAM models using boosting.

Random Forest using an implemented variable selection step (RF)

Breiman proposed random forests [Breiman, 2001], a technique related to CART but with a different construction. In a random forest, in contrast to

traditional regression trees generation, each node is split using the best split among a subset of predictors randomly chosen at that node. This strategy, which is robust against overfitting, is applied by using only two parameters, the number of variables in the random subset for each node and the number of trees in the forest; and is usually not very sensitive to their values. The `randomForest` package [Liaw and Wiener, 2002] provides an R interface to the Fortran programs by Breiman and Cutler. The algorithm adjusts a random forest model departing from the p predictor variables and by fixing the number of trees to B. For each tree $i = 1, \ldots, B$, a bootstrapped sample of size n is generated and then a regression tree is fitted by selecting $m \leq p$ random predictors. The result is a forest of B random trees. Variable selection has been implemented through a stepwise selection procedure based on the out-of-bag error as previously described [Diaz-Uriarte and Alvarez de Andres, 2006].

Neural Networks using an implemented variable selection step (NNet)

An Artificial Neural Network (ANN) is an information processing paradigm in which layers of neuron-like (units) nodes mimic how human brains analyse information. These structures have received a lot of attention for their abilities to 'learn' about relationships among variables and to approximate any continuous function. There are input layers and output layers, and a hidden layer between them where artificial neurons take in a set of weighted inputs and produce an output through an activation function. The R package `nnet` [Venables and Ripley, 2002] was used to adjust a neural network with two units in the hidden layer. Neural networks are sometimes qualified as "black box" given the difficulties to establish the relationship between explanatory variables and dependent variables. We have to that end, implemented a variable selection algorithm similar to the one developed for random forest, based on the connection weights [Olden et al., 2004] following a stepwise selection. At each step, the coefficient of determination was computed and variable importance was measured [Gevrey et al., 2003].

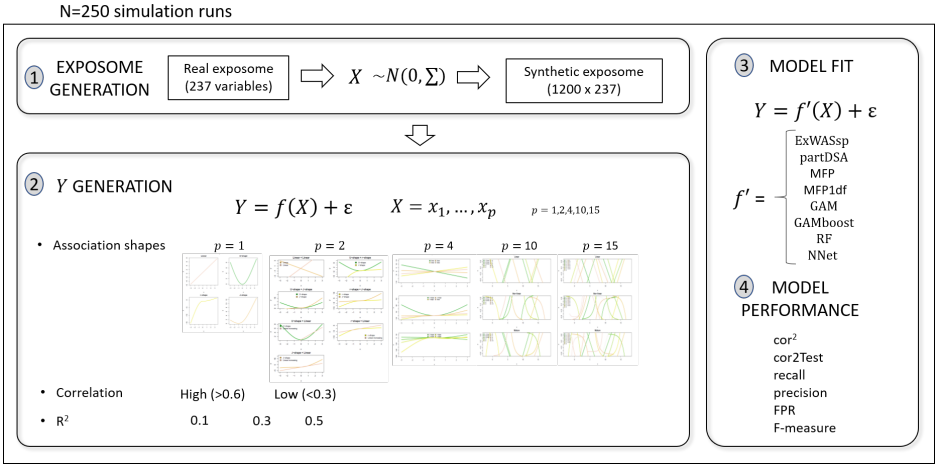


Figure 5.1: For each of the 250 simulation runs, a synthetic exposome of sample size 1200 was generated. 120 scenarios were generated for Y with various association shapes, correlation level of true predictors and R^2. Methods were evaluated to test associations in terms of cor^2, cor2Test, recall, precision, FPR and F-measure.

Data simulation based on real exposome data

Synthetic data were generated using exposome data from a real study and are based on similar procedures as described in [Agier et al., 2016]. Figure 5.1 summarizes the main steps we have performed to simulate the data. Exposure variables were simulated with a realistic correlation structure using data from the existing INMA (Infancia y Medio Ambiente) mother-child cohort [Guxens et al., 2012], in which a total of 237 environmental factors were assessed in 122 mothers during pregnancy through questionnaires, geospatial modeling, and biological monitoring. The closest positive definite matrix of the pairwise correlations was used as our benchmark exposures correlation matrix Σ. Using this matrix, synthetic exposomes, X, were generated for a sample size equal to 1,200 in order to reproduce a real situation in an ongoing European expo-

some project having INMA as one of the participating cohorts, HELIX [Vrijheid et al., 2014, Maitre et al., 2018]. Simulated exposomes were randomly generated from a mean centered multivariate normal distribution and Σ as the covariance matrix: $X \sim N(0, \Sigma)$. Multiple response variables (e.g. *omic* features), Y, were simulated according to different types of non-linear relationships with the exposome as described in [May and Bigelow, 2005]. These include: linear, U-shape, J-shape and r-shape (Supplementary figure 5.5). A total of 120 different scenarios were generated according to: the number exposures truly associated with Y (p = 1, 2, 4, 10 or 15), the correlation level between true predictors (low or high), the proportion of variance explained by the true predictors (coefficient of determination, R^2 = 0.1, 0.3 or 0.5) and different combination of shapes of association (4 for 1 predictor, 7 for 2 predictors, 3 for 4 predictors, 3 for 10 and 3 for 15 predictors, Supplementary figures 5.5 - 5.9). Correlation levels were obtained by assessing the 95th percentile of correlation distribution for each variable with the rest of the exposome. Those exposures having a 95th percentile lower than 0.3 were classified as having "low correlation" and those larger than 0.6 were considered as having "high correlation".

The outcome Y, a set of continuous variables, was generated as a prediction of non-linear models with natural splines obtained by the original databases using p randomly simulated exposures X:

$$Y = f(X) + \epsilon \tag{5.1}$$

where

$$\epsilon \sim N(0, \sigma^2_\epsilon)$$

$$\sigma^2_\epsilon = \frac{(1–R^2)Var\Big(f\big(X\big)\Big)}{R^2}, \quad R^2 = 0.1, 0.3, 0.5$$

$$X = x_1, \ldots, x_p, \quad p = 1, 2, 4, 10, 15$$

250 simulation runs were generated for each scenario. Each proposed statistical method was applied to assess the association between simulated Y and X data set.

Measures of performance

Six measures were considered to assess performance of the different methods in the simulations. Pseudo-coefficient of determination was taken as the square of correlation coefficient between the predictions and the observed response:

$$\text{cor}^2 = \text{pseudo-R}^2 = Cor(\hat{y}, y)^2 \tag{5.2}$$

To control for feasible inflation of cor^2 due to overfitting, this measure was also evaluated in the test set of a 4-fold cross-validation procedure (cor2Test).

The other evaluation measures are based on the true positive (TP), number of true predictors; false positive (FP), number of false positive predictors that the model returns; true negative (TN), number of false predictors that the model rejects and false negative (FN), number of true predictors rejected by the model. According to these definitions, recall is defined as the proportion of true predictors which the model selects with respect to all the real predictors:

$$\text{recall} = \frac{TP}{TP + FN} \tag{5.3}$$

Recall measure, or true positive rate, is called sensitivity in case of binary classification.

Precision, sometimes called positive predictive value, is defined as the proportion of true predictors with respect to all the variables selected by the model:

$$\text{precision} = \frac{TP}{TP + FP} \tag{5.4}$$

The false positive rate (FPR), Type I error or also (1 - specificity), is defined as the proportion of false predictors which the model selects with respect to all the false predictors:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5.5}$$

The F-measure is defined as a combination of precision and recall:

$$\mathrm{F} = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.6}$$

By this definitions, the higher the cor^2, cor2Test, recall, precision and F-measure, the better; the lower the FPR the better.

Real data: INMA data sets

The INMA project also has information on different *omic* data. In this paper we will illustrate the analysis of exposome-transcriptomic and exposome-epigenomic data. The INMA transcriptome was obtained on 308 children using the HTA 2.0 array (Affymetrix, USA), which provides gene expression for 67,528 transcript clusters. The INMA epigenomic methylome consists of 476,946 features obtained with the 450k methylation array (Illumina, USA) which were measured in the same 308 individuals by means of their beta ratios. The exposome data set used in this illustrative example includes 23 exposure variables measured in the 308 mothers during their pregnancy, mainly in the third trimester (Supplementary table 5.2).

INMA transcriptomic data set was filtered retaining those features with a mean expression above 3 followed by a filter based on the standard deviation (sd) of the remaining features above 0.3; obtaining a total of 7,629 features. INMA epigenomic methylation data was filtered using exclusively a filter based on the sd, selecting those CpGs with sd $>$0.1; obtaining a total of 7,320 CpGs. The exposome-transcriptome and exposome-methylome analyses include 100 individuals with complete data. The association analyses between INMA's exposome and the two omic data sets will be referred as INMA_transcriptomics and INMA_methylomics, respectively.

Real data analysis

For each data set, every analysed omic feature was evaluated by three measures based on the correlation test between real and predicted data:

cor^2, its p-value and the false discovery rate BH adjusted p-value. In addition, AIC was also obtained for those methods that allow its computation. To compare variables among methods, a scoring measure was defined as the total number of times the variable was selected by a method divided by the number of total feature tests performed. This enables to measure the relevance of a variable in the analysed data set across the different methods.

Since we had no gold standard to compare the results obtained in the non-linear multivariate analysis, we performed an *in silico* analysis to provide evidences that our results may have any real impact. To this end, an accuracy analysis between the results obtained from each INMA data analysis and genes obtained from The Comparative Toxicogenomics Database (CTD) [Davis et al., 2017a] was carried out as follows. For each significant gene, `CTDquerier` R package [Hernandez-Ferrer and Gonzalez, 2018] was used to obtain related chemicals and was compared to the method's selected variables, which are chemical exposures as well. Then, sensitivity, specificity, precision and the F-measure were obtained from these comparisons.

INMA_transcriptomics data set was annotated with the Affymetrix files (v.na36), extracting gene symbols. R package IIlluminaHumanMethylation450kanno.ilmn12.hg19 was used annotate the INMA_methylomics data set, where CpGs were annotated to the belonging gene or to the closest gene.

5.4 Results

Simulation study

Overall, multivariable fractional polynomial was the best method when considering all scenarios (Table 5.1, Figure 5.2 and Supplementary figure 5.10). Our simulation results also indicate that neural networks are the least convenient. We have organized the main results to facilitate the reader's model evaluation based on his/her interests (e.g. favouring high specificity or low false discovery rate).

Model		cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp		0.40 [0.1,0.69]	0.21 [0.02,0.47]	0.80 [0.13,1]	0.17 [0.01,1]	0.19 [0.03,0.57]	0.2 [0,0.48]	48.54 [2,116]
partDSA		0.39 [0.24,0.57]	0.18 [0.01,0.43]	0.63 [0.07,1]	0.17 [0.07,0.36]	0.23 [0.08,0.44]	0.04 [0.02,0.06]	11.22 [8,14]
MFP		0.29 [0.07,0.52]	0.27 [0.04,0.53]	0.55 [0,1]	0.77 [0,1]	0.58 [0,1]	0 [0,0.01]	3.02 [1,6]
MFP1df		0.21 [0.03,0.5]	0.20 [0.02,0.5]	0.46 [0,1]	0.69 [0,1]	0.50 [0,1]	0 [0,0.01]	1.77 [1,4]
GAM		0.30 [0.07,0.53]	0.27 [0.04,0.53]	0.52 [0,1]	0.48 [0,1]	0.42 [0,1]	0.01 [0,0.03]	3.83 [1,9]
GAMboost		0.32 [0.13,0.53]	0.27 [0.04,0.53]	0.71 [0.13,1]	0.16 [0.04,0.44]	0.21 [0.07,0.46]	0.07 [0.02,0.11]	18.41 [7,28]
RF		0.23 [0.04,0.46]	0.23 [0.02,0.48]	0.69 [0.1,1]	0.26 [0.04,0.67]	0.3 [0.06,0.8]	0.06 [0,0.23]	16.56 [2,57]
NNet		0.38 [0.22,0.59]	0.11 [0,0.42]	0.72 [0,1]	0.06 [0,0.25]	0.08 [0,0.23]	0.47 [0.02,1]	111.32 [5,237]
Model	R2	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	0.1	0.19 [0.08,0.37]	0.06 [0.01,0.13]	0.73 [0.1,1]	0.27 [0.02,1]	0.26 [0.04,0.8]	0.11 [0,0.34]	28.67 [1,83]
partDSA	0.1	0.26 [0.23,0.29]	0.03 [0,0.06]	0.53 [0,1]	0.11 [0,0.23]	0.16 [0,0.29]	0.05 [0.04,0.06]	12.95 [11,14]
MFP	0.1	0.09 [0.06,0.13]	0.08 [0.02,0.14]	0.43 [0,1]	0.72 [0,1]	0.47 [0,1]	0 [0,0.01]	2.12 [1,4]
MFP1df	0.1	0.07 [0.02,0.12]	0.07 [0.01,0.13]	0.37 [0,1]	0.61 [0,1]	0.42 [0,1]	0 [0,0.01]	1.43 [1,3]
GAM	0.1	0.10 [0.06,0.14]	0.08 [0.02,0.14]	0.41 [0,1]	0.47 [0,1]	0.37 [0,1]	0.01 [0,0.02]	2.62 [1,5]
GAMboost	0.1	0.15 [0.12,0.18]	0.08 [0.03,0.14]	0.66 [0.1,1]	0.09 [0.04,0.2]	0.14 [0.06,0.27]	0.09 [0.06,0.12]	22.77 [17,29]
RF	0.1	0.06 [0.03,0.1]	0.05 [0.01,0.11]	0.67 [0.1,1]	0.11 [0.02,0.29]	0.16 [0.03,0.34]	0.12 [0.02,0.34]	28.98 [7,81]
NNet	0.1	0.30 [0.2,0.4]	0.02 [0,0.05]	0.73 [0,1]	0.03 [0,0.07]	0.05 [0,0.13]	0.56 [0.24,1]	133.42 [57,237]
ExWASsp	0.3	0.42 [0.28,0.58]	0.19 [0.08,0.32]	0.82 [0.2,1]	0.14 [0.01,0.5]	0.17 [0.02,0.46]	0.21 [0.01,0.47]	51.91 [3,113]
partDSA	0.3	0.39 [0.34,0.44]	0.17 [0.08,0.25]	0.66 [0.1,1]	0.17 [0.08,0.33]	0.23 [0.09,0.4]	0.04 [0.03,0.05]	11.23 [9,13]
MFP	0.3	0.29 [0.24,0.34]	0.27 [0.16,0.36]	0.58 [0,1]	0.80 [0,1]	0.60 [0,1]	0 [0,0.01]	3.13 [1,6]
MFP1df	0.3	0.20 [0.05,0.32]	0.20 [0.03,0.34]	0.48 [0,1]	0.72 [0,1]	0.53 [0,1]	0 [0,0.01]	1.81 [1,4]
GAM	0.3	0.30 [0.25,0.35]	0.27 [0.17,0.36]	0.54 [0,1]	0.50 [0,1]	0.44 [0,1]	0.01 [0,0.03]	3.87 [1,8]
GAMboost	0.3	0.32 [0.25,0.37]	0.27 [0.17,0.36]	0.73 [0.2,1]	0.15 [0.04,0.36]	0.21 [0.08,0.4]	0.07 [0.03,0.11]	18.1 [10,27]
RF	0.3	0.23 [0.16,0.28]	0.23 [0.13,0.32]	0.71 [0.1,1]	0.25 [0.05,0.67]	0.31 [0.08,0.73]	0.05 [0,0.16]	13.66 [3,40]
NNet	0.3	0.38 [0.24,0.52]	0.07 [0,0.18]	0.74 [0.1,1]	0.04 [0,0.09]	0.06 [0.01,0.15]	0.50 [0.11,1]	118.29 [28,237]
ExWASsp	0.5	0.60 [0.47,0.72]	0.37 [0.21,0.51]	0.86 [0.2,1]	0.11 [0.01,0.36]	0.14 [0.02,0.36]	0.26 [0.02,0.53]	64.96 [6,127]
partDSA	0.5	0.53 [0.46,0.59]	0.36 [0.23,0.46]	0.69 [0.1,1]	0.22 [0.09,0.44]	0.29 [0.11,0.5]	0.03 [0.02,0.05]	9.48 [7,12]
MFP	0.5	0.49 [0.44,0.53]	0.46 [0.33,0.56]	0.64 [0.1,1]	0.80 [0.25,1]	0.65 [0.15,1]	0 [0,0.01]	3.81 [1,8]
MFP1df	0.5	0.34 [0.08,0.52]	0.34 [0.06,0.53]	0.51 [0,1]	0.74 [0,1]	0.56 [0,1]	0 [0,0.01]	2.06 [1,5]
GAM	0.5	0.50 [0.45,0.54]	0.47 [0.36,0.56]	0.61 [0,1]	0.48 [0,1]	0.46 [0,1]	0.01 [0,0.04]	4.99 [1,10]
GAMboost	0.5	0.50 [0.4,0.55]	0.47 [0.34,0.56]	0.75 [0.2,1]	0.23 [0.05,0.67]	0.28 [0.08,0.62]	0.05 [0.01,0.1]	14.35 [5,25]
RF	0.5	0.41 [0.32,0.47]	0.41 [0.29,0.51]	0.70 [0.1,1]	0.42 [0.1,1]	0.45 [0.13,1]	0.02 [0,0.07]	7.05 [2,20]
NNet	0.5	0.47 [0.26,0.63]	0.23 [0,0.49]	0.68 [0,1]	0.12 [0,0.5]	0.13 [0,0.57]	0.34 [0,1]	82.26 [2,237]
Model	Cor.var	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	high	0.47 [0.17,0.71]	0.17 [0.01,0.4]	0.88 [0.4,1]	0.07 [0.01,0.21]	0.11 [0.02,0.31]	0.31 [0.08,0.52]	75.73 [21,125]
partDSA	high	0.40 [0.24,0.57]	0.18 [0.01,0.43]	0.59 [0.07,1]	0.15 [0.07,0.3]	0.21 [0.07,0.4]	0.04 [0.03,0.06]	11.58 [8,14]
MFP	high	0.29 [0.07,0.52]	0.27 [0.04,0.53]	0.48 [0,1]	0.67 [0,1]	0.50 [0,1]	0 [0,0.01]	3.02 [1,6]
MFP1df	high	0.21 [0.03,0.5]	0.20 [0.01,0.49]	0.36 [0,1]	0.56 [0,1]	0.40 [0,1]	0 [0,0.01]	1.78 [1,4]
GAM	high	0.30 [0.07,0.53]	0.27 [0.04,0.53]	0.44 [0,1]	0.34 [0,1]	0.31 [0,1]	0.01 [0,0.03]	4.5 [1,9.55]
GAMboost	high	0.32 [0.13,0.53]	0.27 [0.04,0.53]	0.67 [0.1,1]	0.13 [0.04,0.36]	0.18 [0.06,0.4]	0.07 [0.02,0.11]	19.29 [8,29]
RF	high	0.24 [0.04,0.46]	0.24 [0.03,0.48]	0.71 [0.1,1]	0.22 [0.02,0.67]	0.27 [0.05,0.67]	0.08 [0,0.24]	20.7 [3,57]
NNet	high	0.38 [0.21,0.59]	0.11 [0,0.42]	0.66 [0,1]	0.06 [0,0.21]	0.07 [0,0.22]	0.45 [0.02,1]	107.62 [5,237]
ExWASsp	low	0.34 [0.09,0.61]	0.24 [0.04,0.5]	0.73 [0.1,1]	0.27 [0.03,1]	0.27 [0.06,0.8]	0.08 [0,0.27]	21.36 [1,65]
partDSA	low	0.39 [0.24,0.57]	0.18 [0.01,0.43]	0.66 [0.1,1]	0.19 [0.08,0.4]	0.25 [0.09,0.5]	0.04 [0.02,0.05]	10.86 [7,14]
MFP	low	0.29 [0.07,0.52]	0.27 [0.04,0.53]	0.62 [0.07,1]	0.87 [0.33,1]	0.65 [0.12,1]	0 [0,0]	3.02 [1,6]
MFP1df	low	0.21 [0.03,0.5]	0.20 [0.02,0.5]	0.55 [0,1]	0.83 [0,1]	0.61 [0,1]	0 [0,0.01]	1.77 [1,4]
GAM	low	0.30 [0.08,0.53]	0.28 [0.04,0.53]	0.60 [0,1]	0.62 [0,1]	0.54 [0,1]	0.01 [0,0.02]	3.15 [1,7]
GAMboost	low	0.32 [0.13,0.53]	0.27 [0.04,0.53]	0.75 [0.2,1]	0.18 [0.04,0.53]	0.24 [0.08,0.5]	0.06 [0.01,0.1]	17.53 [7,26]
RF	low	0.23 [0.04,0.45]	0.23 [0.02,0.47]	0.67 [0.1,1]	0.30 [0.04,1]	0.34 [0.07,0.8]	0.05 [0,0.16]	12.43 [2,40]
NNet	low	0.39 [0.23,0.59]	0.10 [0,0.42]	0.77 [0.1,1]	0.06 [0.01,0.29]	0.08 [0.01,0.24]	0.48 [0.02,1]	115.02 [5,237]

Table 5.1: Performance measures of each model for different scenarios in terms of the evaluation measures: cor^2, cor2Test, recall, precision, FPR, F-measure and Nvar (number of selected variables). Brackets indicate confidence interval based on 5% and 95% percentiles from simulation results for each measure. The first block contains averaged results across all scenarios and the subsequent blocks contain averaged results across the different assumptions for R^2 and Cor.var (correlation levels between true predictors).

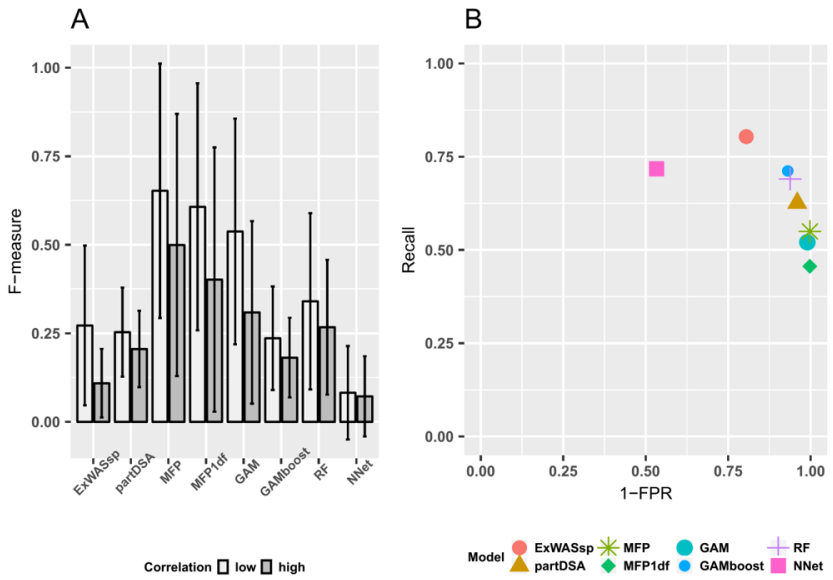


Figure 5.2: Results for all tested methods in the 250 simulation runs. A. Bar plots depicting the mean F-measure, which is related to the precision and recall measures and defined as the ratio of their product divided by their sum. Results are averaged by high or low correlated variables. Error bars depict corresponding standard deviation. B. Recall (sensitivity) versus 1-FPR (specificity) averaged measures across the 250 simulation runs.

Results based on F-measure, recall, precision, FPR and NVar

MFP presented the best F-measure (that considers both recall and precision, the higher the better) over all scenarios (on average 0.58, ranging from 0.47 to 0.65 across simulation scenarios), despite it recorded rather low recall values. It also displayed close to null FPR values.

MFP1df and GAM were the second and third best methods in terms of F-measure (on average 0.50 and 0.42, respectively). In all scenarios, criteria values were comparable (yet slightly lower) between MFP1df and MFP; in comparison, GAM mainly displayed a marginally lower precision.

GAMboost, partDSA and RF obtained comparable results, with average F-measure values in the 0.21-0.30 range. They selected a larger number of variables (average values > 10 whilst it was on average < 3 in all MFP, MFP1df and GAM methods), leading to large recall values, but at the cost of a low precision. FPR values were also larger.

ExWASsp and NNet selected by far the largest number of variables, such that their precision and more global F-measure values were low, and their FPR was high (despite their recall values were overall the highest).

As expected, all methods performances improved with increasing explanatory power of variables (i.e. R^2) and with decreased correlation amongst true predictors, except for ExWASsp.

Results based on cor^2 and cor2Test measures

Despite ExWASsp, partDSA and NNet showed the highest cor^2 (on average 0.4, 0.39 and 0.38 respectively, ranging from 0.19 to 0.6 in the different scenarios), these methods seem to overfit the data, as can be seen from their cor2Test values. In this regard MFP, GAM and GAMboost showed the highest cor2Test value (0.27, ranging from 0.08 to 0.47).

Evaluation of the pseudo-coefficient of determination, obtained directly and particularly through the cross-validation procedure, was in general close to the simulated coefficient of determination (0.1, 0.3 or 0.5), being again MFP, GAM and GAMboost the closest methods. In this sense, RF did not improve these values much whereas NNet seemed to be very much overfitted.

Results by type of association

Performance measures obtained in simulated scenarios assuming exclusively linear associations showed MFP1df was the best (F-measure of 0.51, range 0.41-0.6, Supplementary table 5.3), slightly outperforming in this case MFP (0.48, range 0.38-0.57). Except for MFP1df, results evaluated on linear models only had slightly lower F-measures and higher FPR.

Scenarios for one true predictor considering U-shape, r-shape and J-shape showed an even better adjustment for MFP with 4 df; with an

F-measure of ranging from 0.84 to 0.88. Not surprisingly, in these scenarios, MFP1df obtained very bad results for U-shapes (F-measure of 0.04) whereas presented just a reasonable drop for J-shape and r-shape (F-measure of 0.78 and 0.76 respectively) (Supplementary tables 5.4-5.6).

GAM obtained acceptable results in all shapes of associations whilst partDSA, RF, GAMboost and NNet presented poor results.

Real data analyses

In order to provide a more comprehensive benchmarking, we analysed the INMA real data sets with all the proposed methods using an FDR adjusted p-value < 0.05 and a $cor^2 > 0.33$ to indicate that a model was statistically significant. Results obtained for INMA_transcriptomics showed a large variability among methods in terms of the number of significant probes (Figure 5.3 and Supplementary table 5.7). partDSA returned 7,624 features; GAMboost 7,306; RF 6,977; MFP 3,759; ExWASsp 2,367; GAM 1,719 and NNet 1,554 probes.

99 common probes were found across all methods, belonging to 31 genes. The highest cor^2 and lowest p-values were obtained for GAM. The top scoring exposure variables, defined as the proportion of times the variable was selected by a method, were phtalates for ExWASsp, MFP and RF (0.75, 0.65 and 0.45 respectively), and metals for other methods: caesium for partDSA (0.49), arsenic for GAM (0.38), mercury for GAMboost (0.77) and zinc for NNet (0.89). Accuracy analysis, obtaining real data from The Comparative Toxicogenomics database, revealed that GAMboost was the method with the highest F-measure (mean=0.57, P05=0.40, P95=0.67) and GAM had the lowest (mean=0.36, P05=0.14, P95=0.67) (Supplementary figure 5.11) whereas MFP was the method with the best balance between sensitivity (mean=0.50, P05=0.17, P95=1) and specificity (mean=0.85, P05=0.74, P95=0.94). For MFP, the contribution of the variables is very different in patterns as shown in Figure 5.4. Some of these interactions have already been described in the CTD, as the *S100A12*-Zn or the *IFI27*-BPA.

INMA_methylomics analysis also showed different results depending

on the method (Figure 5.3 and Supplementary table 5.7). GAMboost returned 6,978 CpGs; partDSA 6,785; RF 6,544; MFP 4,033; NNet 2,457; GAM 2,216 and ExWASsp 1,493. 188 common CpGs were obtained across those methods, located in (or close to) 126 genes. Top scoring exposures variables were PCB 138 for MFP and RF (0.39 and 0.34 respectively), PCB 180 for ExWASsp (0.33), PCB 118 for partDSA (0.40), caesium for GAM (0.25), bisphenol A for GAMboost (0.76) and zinc for NNet (0.61). Accuracy analysis performed in this data set illustrates a similar behaviour compared with that of INMA_transcriptomics, with the highest F-measures being observed for GAMboost (mean=0.56, P05=0.40, P95=0.67) and GAM (mean=0.40, P05=0.22, P95=0.50). partDSA and MFP were in this data set the methods with the most stable measure of sensitivity (mean=0.51 for both) and specificity (mean=0.84 and 0.83, respectively). Details are shown in Supplementary figure 5.11.

nlOmicAssoc package

All methods assessed in this article have been encapsulated in the R package called `nlOmicAssoc`. The package includes a common interface for all methods, imputation capabilities, and graphical enhancements. It also includes functions to filter and obtain variable scores for each method as well and a comparison tool for the obtained results. Furthermore, parallelization is enabled. `nlOmicAssoc` is available at *https://github.com/isglobal-brge/nlOmicAssoc*.

5.5 Discussion

We have for the first time evaluated and compared seven different methods that accommodate non-linear associations between multiple predictors and multiple variables in the specific context of exposome and omic data. The methods were appraised in simulated data and also in real data.

Simulated data were evaluated by means of six different performance measures and we took into account different scenarios with different

correlation levels of true predictors or with different exposures explanatory power. We have simulated in our study some highly correlated variables, which configure the usual exposome structure, but also nearly uncorrelated variables. In this context, MFP was the most efficient method. ExWASsp presented a high recall but a low precision while GAM offered a good balance between recall and precision, outperformed though by MFP in the F-measure. NNet certainly did not perform well in the context of non-linear association. GAMboost, RF and partDSA performed similarly, underperforming ExWASsp and MFP.

MFP was able to capture linear associations and performed even better in the specific non-linear context. Accordingly, it achieved a good F-measure in the tested mixed scenarios. As expected, for exclusive linear associations MFP1df, had a better performance. Of note, DSA, previously reported by Agier et al. [Agier et al., 2016] as a good method to capture linearity but also reported by Barrera et al. for two-way interactions [Barrera-Gómez et al., 2017]; is only slightly related to partDSA, which had in our simulations unremarkable results. Both algorithms rely on the deletion-substitution-addition algorithm but different implementations were used.

Generalized additive models, either using a backfitting for each continuous covariate or boosting; and the multivariable fractional polynomial procedure provide two strategies to address the common problem of model-building by selection of variables and functional forms for continuous covariates. Consideration needs to be made on methods based on splines, MFP, ExWASsp, GAM and GAMboost. Although splines are used by all those methods to supplant the usual linear association assumption and were also used to generate synthetic variables in the case of simulations, results are very different among them.

When applying these methods to real data sets, high variability was observed among them. Transcriptomic data set (INMA_transcriptomics) showed a higher cor^2 value and significance in the GAM method. In both data sets, GAMboost and partDSA selected the highest number of probes, which may be suspected to be false positives. In this sense ExWASsp and MFP are more parsimonious, with a lower risk for false positive find-

ings. MFP, the method that better performed for simulations, presented consistent sensitivity and specificity as of compared to the CTD data, supporting the simulation results. GAMboost was the one having the highest F-measure. It has to be taken into account that CTD associations between chemicals and genes are obtained from the literature, which assesses their relationships using linear models. Besides and in contrast to our simulations design, 23 variables (exposome) were studied for association versus thousands of probes (*omic*) measured in 100 individuals; which could also affect the results. Therefore, more investigations should be performed on the data and their biological implications to reach a final conclusion.

The exposome, of continuous nature, is growingly considered in the epidemiological literature, which is why it is important to provide an overview of the most efficient methods allowing to characterize its association with health and biological parameters or with the different omics. Most of omic data such as transcriptomic or methylation data are continuous or can easily be transformed to be continuous. For instance, count data obtained from RNA-seq experiments can be transformed into continuous data using the voom mean-variance trend modelling method [Law et al., 2014].

Despite the fact that we tried to cover many scenarios in our analyses, some limitations need to be mentioned. We have selected several methods that have variable selection procedures and cope non-linear associations. Specifically, for random forest and neural networks, other variable selection approaches could have been selected. Alternative methods such as Xboost, deep learning and other machine learning methods could have been considered. Besides, all methods were applied with the default parameters, and we may not have shown the optimal capacities of the different methods. Typically, the selection of degrees of freedom or knots in case of methods based on splines, the number of trees in a random forest or the size in neural networks could affect the given results. The scenarios examined in this study are limited in terms of the type of associations that exist between the exposure variables and omic measurements; more complex shapes may exist, which methods might not be capable of coping with. Besides, given the correlated nature of variables in the exposome it is important to take into account variable correlations. These two facts

could partly explain the variation in real data sets results. Another point to consider is the computing duration of the different methods. The analyses conducted here have been performed probe by probe to detect associations of each gene or CpG to environmental variables. Depending on the number of probes to analyse and the variables to assess, computational time can be an issue, which can be mitigated using parallel procedures. Filtering the data is also advised before applying any of the proposed methods. This includes removing the outliers, as most of assessed methods are sensitive to outliers. Furthermore, a multiblock approach could be conducted to assess association on the two original data sets. Examples of such strategies are canonical correlation analysis (CCA) or cointertia analysis (CIA) which has recently been used to analyse *omic* data [Meng et al., 2016]. However, those are just descriptive procedures, as previously stated.

5.6 Conclusion

Based on our simulation and real data assessment results, we identified MFP as generally showing the best performance. In real case analyses, methodological choices should also be guided by computational complexity and flexibility considerations such as the ability to accommodate for confounders or interaction terms. Given this large variability among methods, one could combine different approaches by considering either those common variables across methods or the union of all significant variables. These types of solutions have already been proposed in other settings with no available standard methods, as the case of copy number detection [Medvedev et al., 2010].

We have implemented a new package, called `nlOmicAssoc`, that is designed to analyse the association between continuous data such as the exposome and omic data. Moreover, it can also be used to study the association of several exposures to a specific outcome such as cholesterol. The package contains different functions to fit each of the aforementioned methods and is programmed to work with the standard R objects such as *matrix* or `data.frame` but also with `ExpressionSet`

or `SummarizedExperiment`, enabling the interaction with other Bioconductor packages.

5.7 Back matter

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JRG, XB, RS and MV proposed the idea of using multivariate methods to assess association between linear outcomes and the exposome. JRG, LA, and RS designed simulation studies. LN wrote the manuscript, carried out simulation studies, performed data analyses and programmed the R package. JRG helped in creating the R package, writing the first version of the manuscript and oversaw the project. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge the input of HELIX - Exposomics statistical working group. Details on the HELIX project can be found at www.projecthelix.eu, and on the EXPOsOMICS project at http://www.exposomicsproject.eu. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya. We thank Caterina Alba Beltran for her initial contribution to this work by programming a first version of the simulations. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013 (grants number 308333, HELIX). This research has received funding from the Ministerio de Economía y Competitividad y Fondo Europeo de Desarrollo (MTM2015-68140-R).

Supplementary material

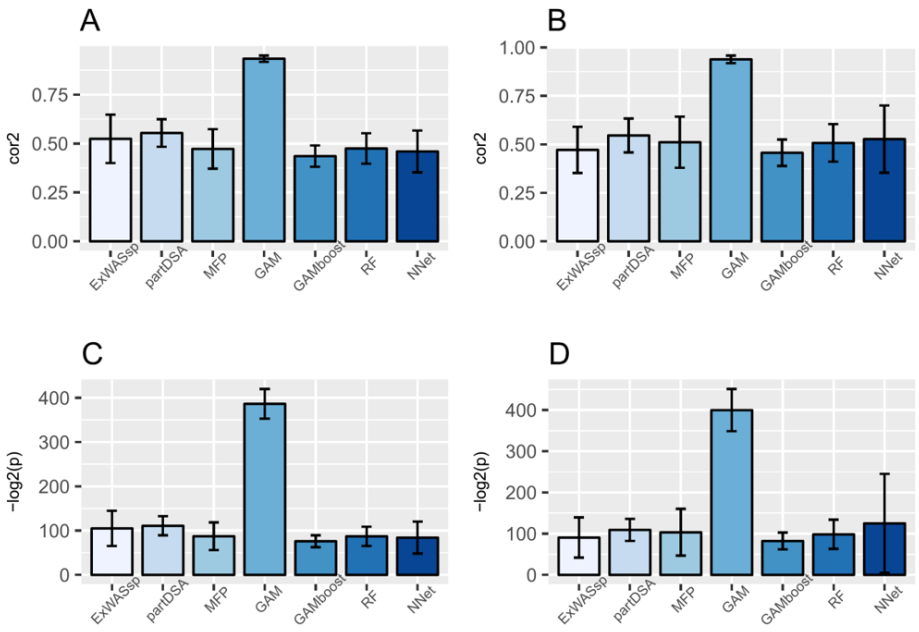


Figure 5.3: A. Bar plots of the mean cor^2 for real data set INMA_transcriptomics. B. Bar plots of the mean cor^2 for real data set INMA_methylomics. C. Bar plots of the mean -($\log_2$(p)) for real data set INMA_transcriptomics. D. Bar plots of the mean -($\log_2$(p)) for real data set INMA_methylomics. Error bars depict corresponding standard deviation.

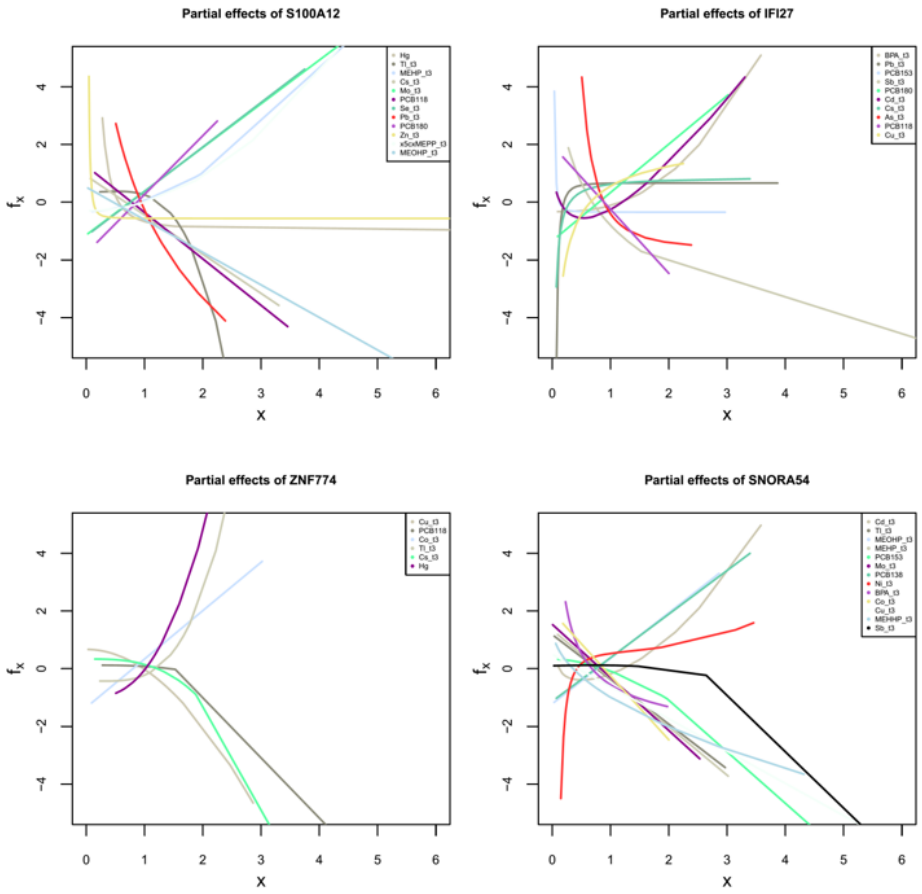
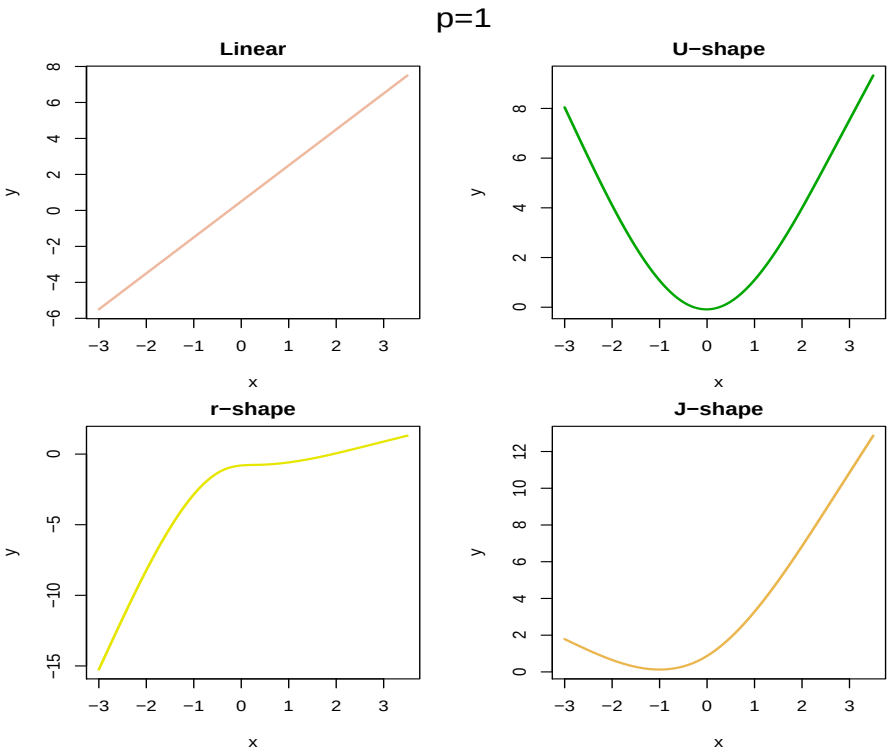
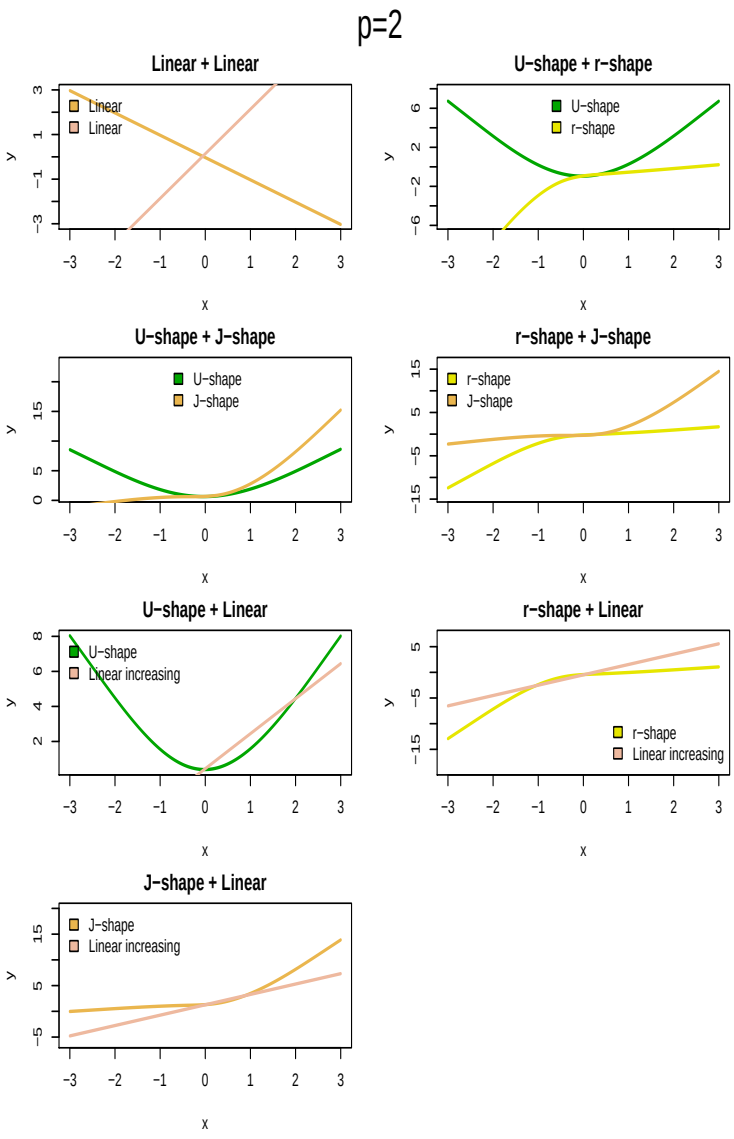


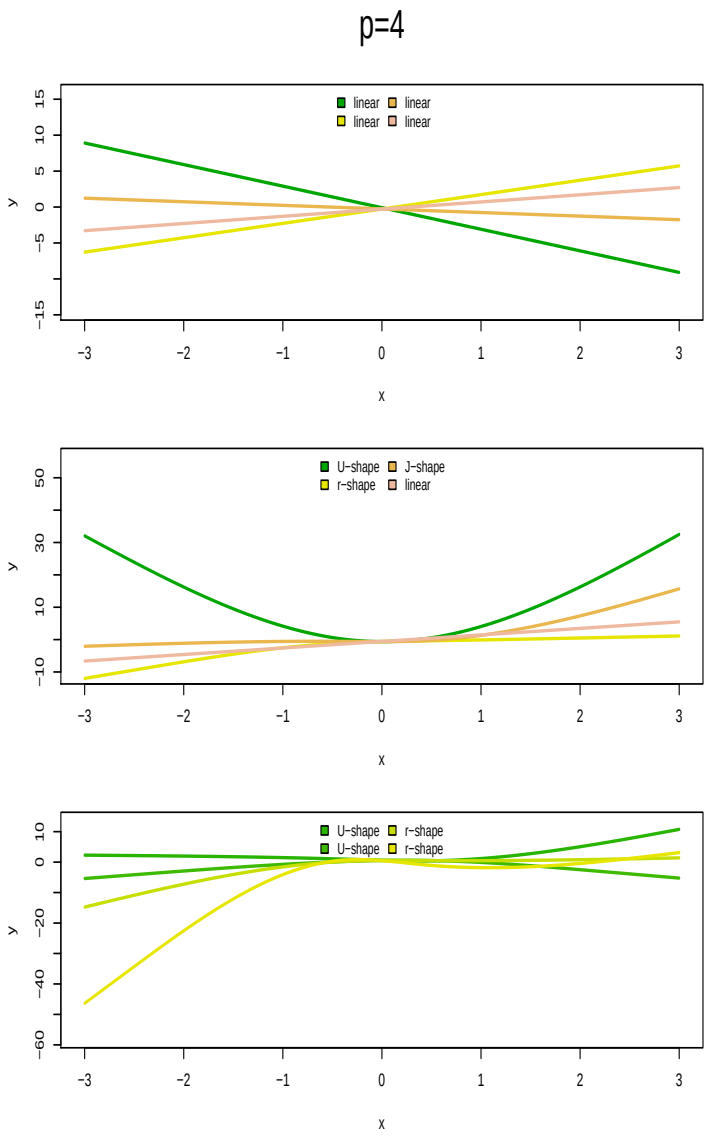
Figure 5.4: Partial effect plots for four of the top gene features selected by the model MFP with an adjusted p-value<0.05 in the INMA_transcriptomics data set. For each gene, measures were scaled and the effect of the final selected variables were standardized and plotted.



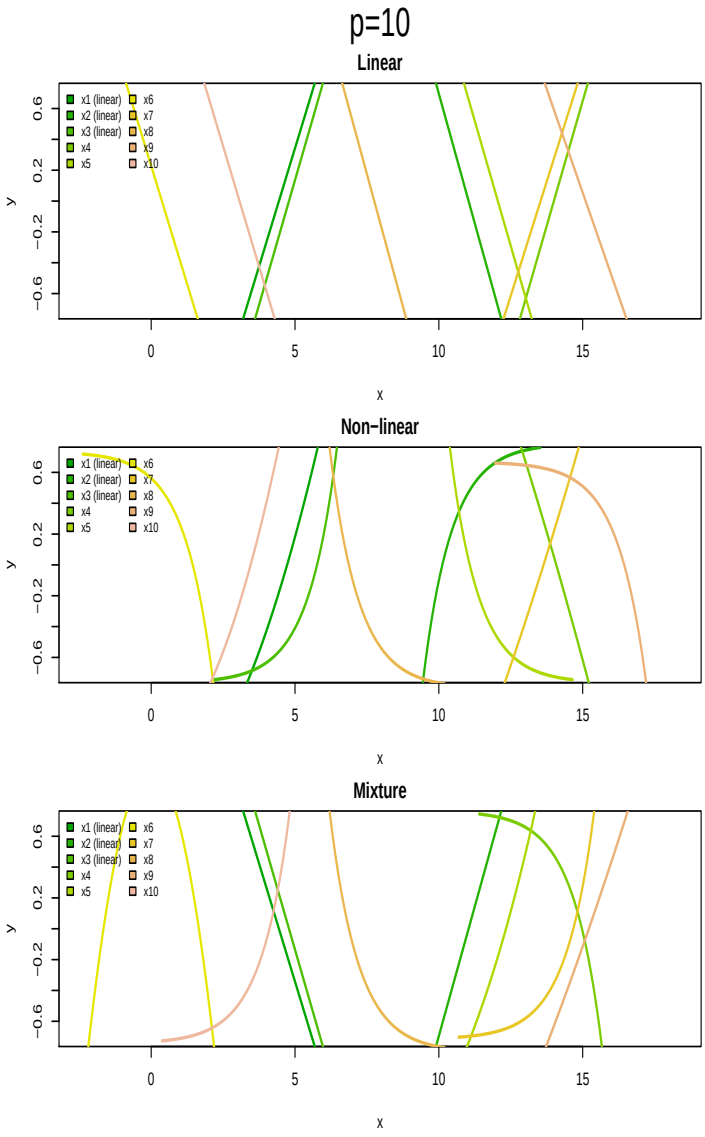
Supplementary figure 5.5: Shapes of simulated associations between two continuous variables: linear, U-shape, J-shape and r-shape.



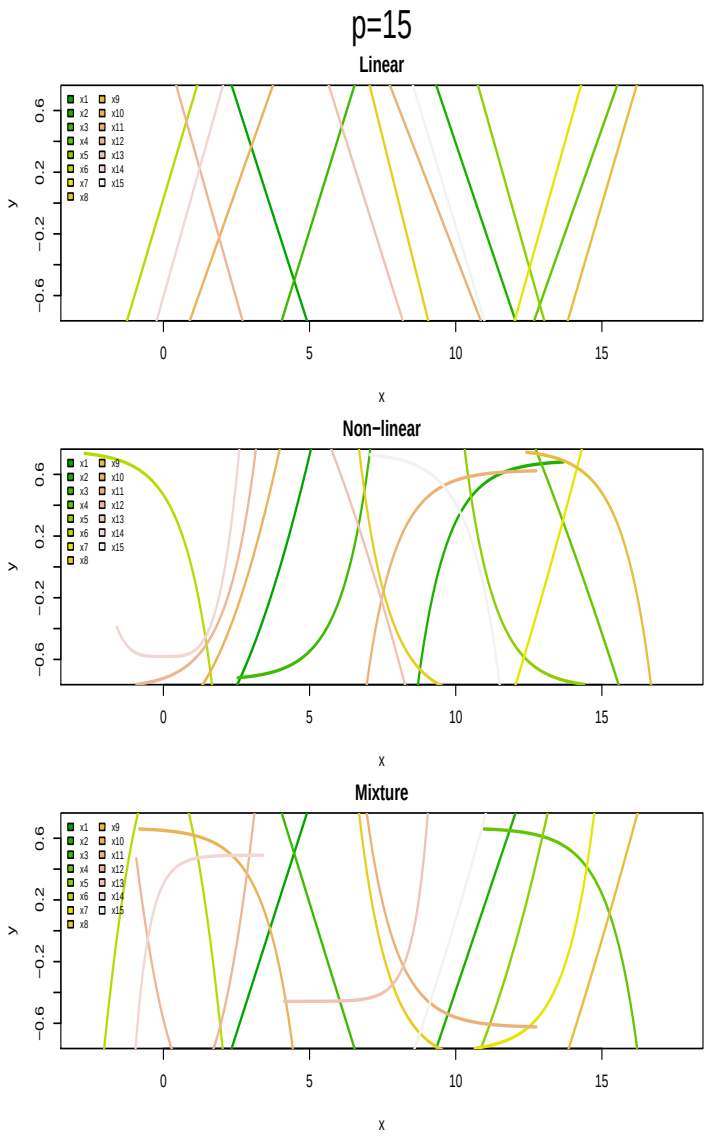
Supplementary figure 5.6: Simulated associations for two true predictors, combining the shapes linear, U-shape, J-shape and r-shape.



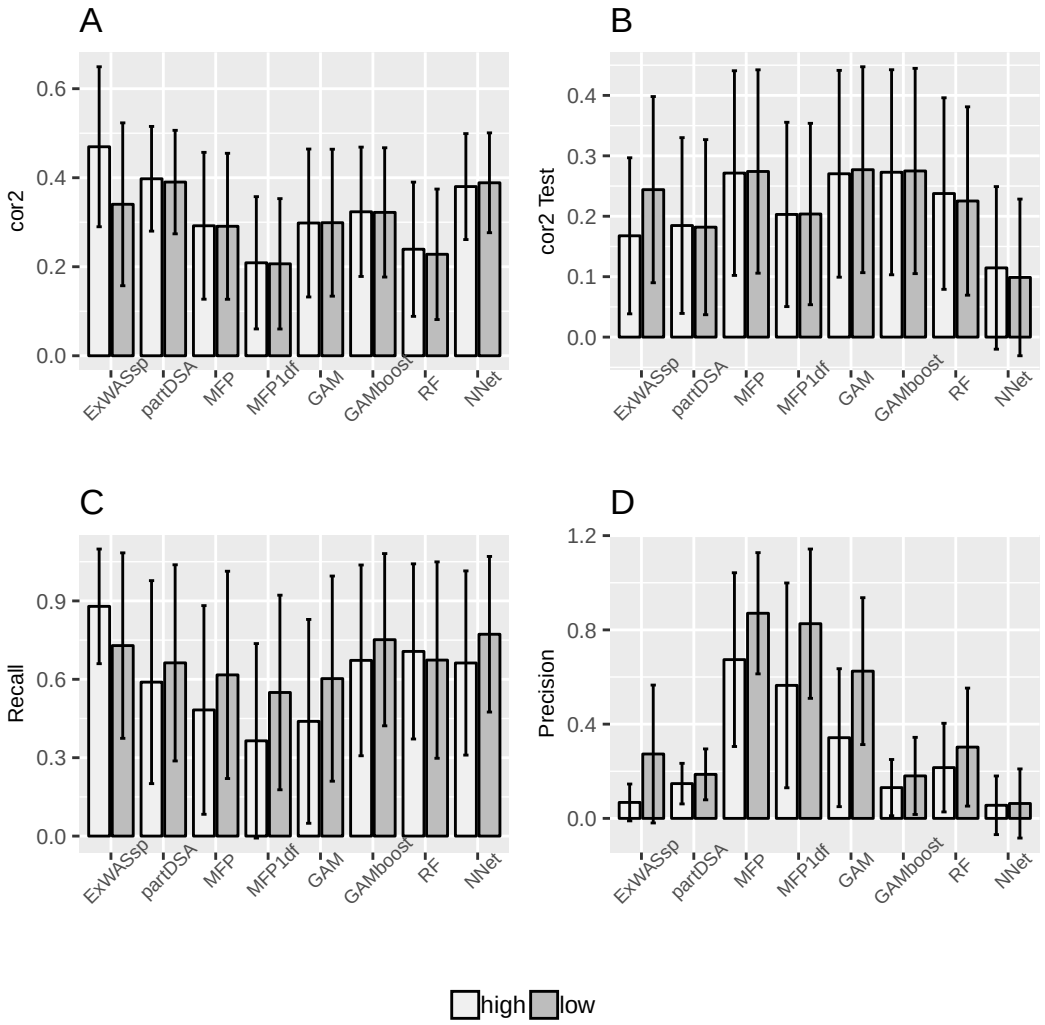
Supplementary figure 5.7: Simulated associations for four true predictors, combining the shapes linear, U-shape, J-shape and r-shape.



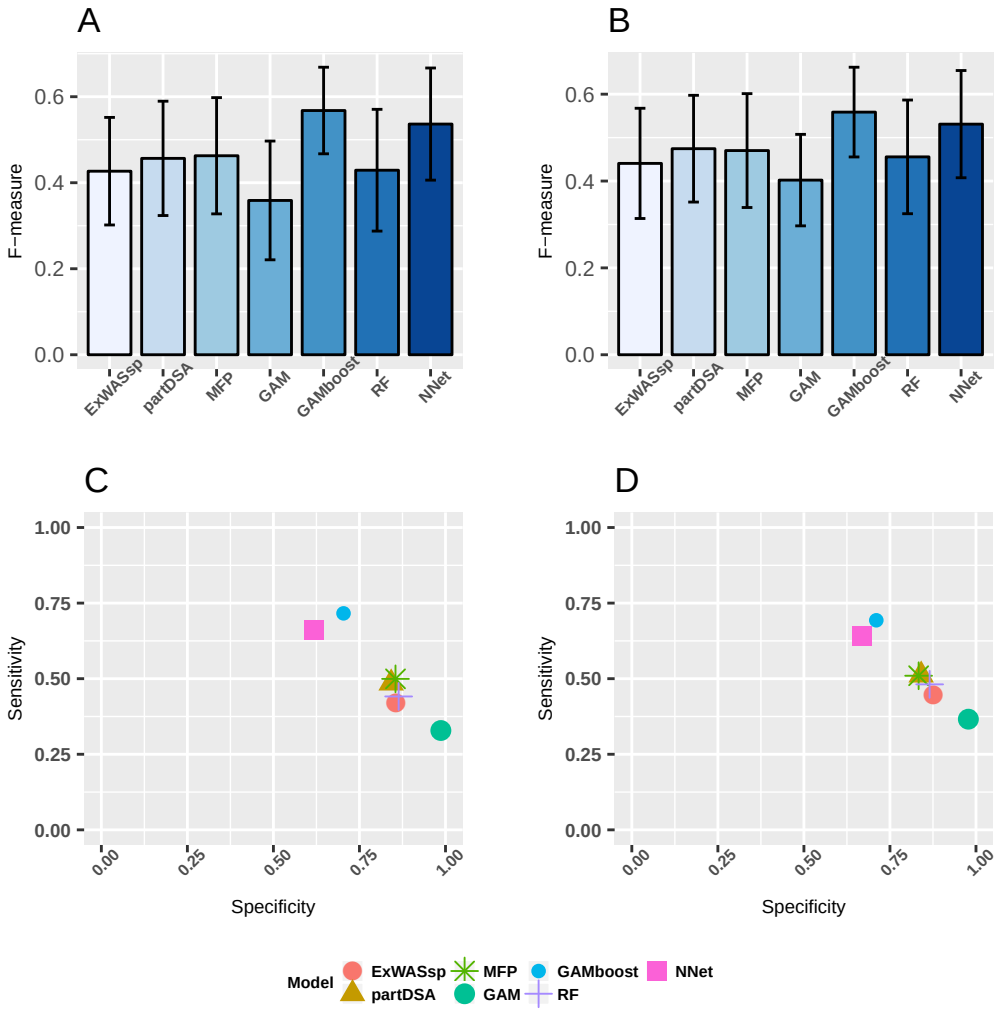
Supplementary figure 5.8: Simulated associations for 10 true predictors, combining the shapes linear, U-shape, J-shape and r-shape. They are combined in three scenarios: Linear, non-linear and a mixture.



Supplementary figure 5.9: Simulated associations for 15 true predictors, combining the shapes linear, U-shape, J-shape and r-shape. They are combined in three scenarios: Linear, non-linear and a mixture.



Supplementary figure 5.10: Performance measures for tested methods averaged over the 250 simulation runs in high/low correlated true predictor scenarios. A: Bar plots of the mean cor^2. B: Bar plots of the mean cor2Test. C: Bar plots of the mean Recall. D: Bar plots of the mean Precision. Error bars depict corresponding standard deviation.



Supplementary figure 5.11: Accuracy analysis performed on INMA data sets in terms of F-measure and sensitivity versus specificity for tested methods. A and C: INMA_transcriptomics. B. and D: INMA_methylomics.

Exposure	Family	Matrix	TimePoint	Type	LOD	LODUnit	Description	CTD chemical term
As_t3	Metals	urine	T3	numeric	0.2	ng/mL	Arsenic (ng/g creatinine adjusted)	Arsenic
BPA_t3	BPA	urine	T3	numeric	NA	NA	BPA (µg/g creatine)	bisphenol A
Cd_t3	Metals	urine	T3	numeric	0.2	ng/mL	Cadmium (ng/g creatinine adjusted)	Cadmium
Co_t3	Metals	urine	T3	numeric	0.2	ng/mL	Cobalt (ng/g creatinine adjusted)	Cobalt
Cs_t3	Metals	urine	T3	numeric	0.2	ng/mL	Caesium (ng/g creatinine adjusted)	Cesium
Cu_t3	Metals	urine	T3	numeric	0.2	ng/mL	Cooper (ng/g creatinine adjusted)	Copper
Hg	Metals	cord blood	NA	numeric	2	µg/l	Mercury (µg/l)	Mercury
MBzP_t3	Phtalates	urine	T3	numeric	0.5	ng/mL	MBzP (ug/g creatinine adjusted)	mono-benzyl phthalate
MEHHP_t3	Phtalates	urine	T3	numeric	0.5	ng/mL	MEHHP (ug/g creatinine adjusted)	mono(2-ethyl-5-hydroxyhexyl) phthalate
MEHP_t3	Phtalates	urine	T3	numeric	1	ng/mL	MEHP (ug/g creatinine adjusted)	mono-(2-ethylhexyl)phthalate
MEOHP_t3	Phtalates	urine	T3	numeric	0.5	ng/mL	MEOHP (ug/g creatinine adjusted)	mono(2-ethyl-5-oxohexyl)phthalate
Mo_t3	Metals	urine	T3	numeric	0.2	ng/mL	Molybdenum (ng/g creatinine adjusted)	Molybdenum
Ni_t3	Metals	urine	T3	numeric	0.2	ng/mL	Nickel (ng/g creatinine adjusted)	Nickel
Pb_t3	Metals	urine	T3	numeric	0.2	ng/mL	Lead (ng/g creatinine adjusted)	Lead
PCB118	PCBs	serum	NA	numeric	NA	NA	PCB 118 (ng/g lipid adjusted)	PCB 118
PCB138	PCBs	serum	NA	numeric	NA	NA	PCB 138 (ng/g lipid adjusted)	PCB 138
PCB153	PCBs	serum	NA	numeric	NA	NA	PCB 153 (ng/g lipid adjusted)	PCB 152
PCB180	PCBs	serum	NA	numeric	NA	NA	PCB 180 (ng/g lipid adjusted)	PCB 180
Sb_t3	Metals	urine	T3	numeric	0.2	ng/mL	Antimony (ng/g creatinine adjusted)	Antimony
Se_t3	Metals	urine	T3	numeric	0.2	ng/mL	Selenium (ng/g creatinine adjusted)	Selenium
Tl_t3	Metals	urine	T3	numeric	0.2	ng/mL	Thallium (ng/g creatinine adjusted)	Thallium
x5cxMEPP_t3	Phtalates	urine	T3	numeric	1	ng/mL	5cxMEPP (ug/g creatinine adjusted)	2-ethyl-5-carboxypentyl phthalate
Zn_t3	Metals	urine	T3	numeric	0.2	ng/mL	Zinc (ng/g creatinine adjusted)	Zinc

Supplementary table 5.2: Variable description of the INMA exposome variables used in the real data analyses.

Model		cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp		0.43 [0.11,0.71]	0.20 [0.02,0.44]	0.75 [0.2,1]	0.12 [0.01,0.4]	0.16 [0.02,0.38]	0.23 [0.01,0.51]	58.21 [4,124]
partDSA		0.40 [0.24,0.58]	0.19 [0.01,0.43]	0.53 [0.07,1]	0.17 [0.07,0.4]	0.22 [0.07,0.44]	0.04 [0.02,0.06]	11.19 [8,14]
MFP		0.29 [0.07,0.52]	0.29 [0.05,0.53]	0.44 [0,1]	0.73 [0,1]	0.48 [0,1]	0 [0,0.01]	2.14 [1,5]
MFP1df		0.30 [0.07,0.52]	0.29 [0.05,0.53]	0.47 [0,1]	0.73 [0,1]	0.51 [0,1]	0 [0,0.01]	2.35 [1,6]
GAM		0.30 [0.08,0.53]	0.28 [0.05,0.53]	0.44 [0,1]	0.41 [0,1]	0.36 [0,1]	0.01 [0,0.04]	5.11 [1,11]
GAMboost		0.34 [0.14,0.54]	0.29 [0.05,0.53]	0.64 [0.1,1]	0.12 [0.04,0.32]	0.18 [0.06,0.4]	0.09 [0.05,0.12]	23.34 [16,30]
RF		0.24 [0.04,0.46]	0.24 [0.03,0.48]	0.61 [0.1,1]	0.22 [0.04,0.67]	0.26 [0.06,0.57]	0.07 [0,0.23]	19.57 [3,57]
NNet		0.45 [0.25,0.64]	0.14 [0.01,0.35]	0.82 [0.33,1]	0.04 [0.01,0.1]	0.07 [0.01,0.17]	0.52 [0.15,1]	124.06 [40,237]
Model	R^2	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	0.1	0.21 [0.08,0.39]	0.05 [0.01,0.11]	0.64 [0.1,1]	0.20 [0.01,1]	0.22 [0.03,0.67]	0.13 [0,0.35]	32.65 [2,87]
partDSA	0.1	0.26 [0.23,0.29]	0.03 [0,0.06]	0.44 [0,1]	0.11 [0,0.23]	0.15 [0,0.29]	0.05 [0.04,0.06]	12.99 [11,14]
MFP	0.1	0.09 [0.06,0.13]	0.09 [0.03,0.15]	0.33 [0,1]	0.67 [0,1]	0.38 [0,1]	0 [0,0.01]	1.51 [1,3]
MFP1df	0.1	0.10 [0.06,0.13]	0.09 [0.04,0.15]	0.37 [0,1]	0.69 [0,1]	0.41 [0,1]	0 [0,0.01]	1.69 [1,3]
GAM	0.1	0.10 [0.07,0.14]	0.08 [0.03,0.14]	0.34 [0,1]	0.42 [0,1]	0.32 [0,1]	0.01 [0,0.02]	3.17 [1,6]
GAMboost	0.1	0.16 [0.13,0.19]	0.08 [0.04,0.14]	0.57 [0.1,1]	0.09 [0.03,0.21]	0.14 [0.05,0.28]	0.10 [0.08,0.12]	25.31 [20,31]
RF	0.1	0.06 [0.03,0.1]	0.06 [0.02,0.11]	0.60 [0.1,1]	0.11 [0.02,0.29]	0.16 [0.03,0.33]	0.13 [0.02,0.34]	33.13 [7,81]
NNet	0.1	0.31 [0.21,0.42]	0.03 [0,0.07]	0.80 [0.27,1]	0.03 [0,0.09]	0.06 [0.01,0.15]	0.56 [0.23,1]	133.26 [57,237]
ExWASsp	0.3	0.45 [0.32,0.6]	0.18 [0.09,0.29]	0.78 [0.2,1]	0.09 [0.01,0.24]	0.14 [0.02,0.33]	0.25 [0.04,0.49]	62.15 [12,120]
partDSA	0.3	0.40 [0.34,0.44]	0.17 [0.09,0.25]	0.56 [0.1,1]	0.17 [0.08,0.36]	0.22 [0.09,0.4]	0.04 [0.03,0.05]	11.31 [9,13]
MFP	0.3	0.29 [0.25,0.33]	0.29 [0.21,0.36]	0.46 [0,1]	0.75 [0,1]	0.51 [0,1]	0 [0,0.01]	2.17 [1,4]
MFP1df	0.3	0.30 [0.25,0.34]	0.29 [0.22,0.37]	0.49 [0,1]	0.74 [0,1]	0.53 [0,1]	0 [0,0.01]	2.4 [1,5]
GAM	0.3	0.30 [0.26,0.35]	0.28 [0.21,0.36]	0.46 [0,1]	0.42 [0,1]	0.37 [0,1]	0.01 [0,0.04]	5.15 [1,10]
GAMboost	0.3	0.34 [0.3,0.37]	0.29 [0.21,0.36]	0.65 [0.1,1]	0.12 [0.04,0.3]	0.18 [0.06,0.38]	0.09 [0.06,0.12]	23.24 [17,29]
RF	0.3	0.24 [0.19,0.28]	0.24 [0.17,0.32]	0.62 [0.1,1]	0.22 [0.05,0.5]	0.27 [0.07,0.55]	0.06 [0.01,0.17]	16.39 [3,40]
NNet	0.3	0.46 [0.35,0.56]	0.12 [0.05,0.2]	0.83 [0.33,1]	0.04 [0.01,0.1]	0.06 [0.01,0.17]	0.53 [0.17,1]	126.26 [40,237]
ExWASsp	0.5	0.64 [0.53,0.74]	0.36 [0.24,0.48]	0.83 [0.33,1]	0.07 [0.01,0.16]	0.12 [0.02,0.25]	0.33 [0.09,0.57]	79.8 [24,137]
partDSA	0.5	0.54 [0.46,0.59]	0.36 [0.24,0.46]	0.59 [0.1,1]	0.23 [0.09,0.5]	0.28 [0.1,0.55]	0.03 [0.02,0.04]	9.28 [6,12]
MFP	0.5	0.50 [0.46,0.53]	0.49 [0.42,0.55]	0.52 [0,1]	0.76 [0,1]	0.56 [0,1]	0 [0,0.01]	2.72 [1,6]
MFP1df	0.5	0.50 [0.46,0.53]	0.49 [0.42,0.56]	0.54 [0,1]	0.75 [0,1]	0.58 [0,1]	0 [0,0.01]	2.95 [1,7]
GAM	0.5	0.50 [0.46,0.54]	0.48 [0.41,0.55]	0.53 [0,1]	0.39 [0,1]	0.39 [0,1]	0.02 [0,0.05]	7 [2,13]
GAMboost	0.5	0.52 [0.48,0.56]	0.49 [0.42,0.55]	0.69 [0.2,1]	0.15 [0.04,0.43]	0.22 [0.07,0.48]	0.08 [0.04,0.11]	21.46 [14,29]
RF	0.5	0.43 [0.36,0.48]	0.43 [0.34,0.5]	0.60 [0.1,1]	0.34 [0.1,0.67]	0.36 [0.13,0.8]	0.03 [0,0.09]	9.18 [2,28]
NNet	0.5	0.57 [0.46,0.67]	0.27 [0.15,0.45]	0.83 [0.33,1]	0.05 [0.01,0.14]	0.08 [0.01,0.22]	0.47 [0.08,1]	112.72 [20,237]
Model	Cor.var	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	high	0.49 [0.18,0.73]	0.17 [0.01,0.38]	0.84 [0.5,1]	0.07 [0.01,0.16]	0.12 [0.02,0.27]	0.34 [0.1,0.56]	83.31 [26,135]
partDSA	high	0.40 [0.25,0.58]	0.19 [0.01,0.43]	0.50 [0.07,1]	0.15 [0.07,0.33]	0.20 [0.07,0.4]	0.04 [0.03,0.06]	11.5 [8,14]
MFP	high	0.29 [0.07,0.52]	0.29 [0.05,0.53]	0.36 [0,1]	0.61 [0,1]	0.40 [0,1]	0 [0,0.01]	2.1 [1,4]
MFP1df	high	0.30 [0.08,0.52]	0.29 [0.05,0.53]	0.39 [0,1]	0.61 [0,1]	0.42 [0,1]	0 [0,0.01]	2.3 [1,5]
GAM	high	0.30 [0.08,0.53]	0.28 [0.04,0.52]	0.35 [0,1]	0.24 [0,0.67]	0.23 [0,0.67]	0.02 [0,0.05]	6.12 [2,12]
GAMboost	high	0.34 [0.14,0.54]	0.28 [0.05,0.53]	0.57 [0.1,1]	0.09 [0.03,0.21]	0.14 [0.05,0.3]	0.10 [0.06,0.12]	24.1 [16,31]
RF	high	0.25 [0.05,0.47]	0.25 [0.03,0.48]	0.63 [0.1,1]	0.19 [0.02,0.5]	0.23 [0.05,0.5]	0.09 [0.01,0.24]	24.34 [3,57]
NNet	high	0.44 [0.24,0.64]	0.14 [0.01,0.35]	0.78 [0.25,1]	0.04 [0.01,0.09]	0.06 [0.01,0.15]	0.52 [0.16,1]	124.39 [40,237]
ExWASsp	low	0.37 [0.09,0.64]	0.23 [0.03,0.47]	0.67 [0.1,1]	0.18 [0.02,0.67]	0.20 [0.04,0.5]	0.13 [0,0.34]	33.13 [2,82.55]
partDSA	low	0.39 [0.24,0.58]	0.18 [0.01,0.43]	0.56 [0.1,1]	0.19 [0.08,0.43]	0.24 [0.09,0.5]	0.04 [0.02,0.05]	10.88 [7,14]
MFP	low	0.29 [0.07,0.52]	0.29 [0.05,0.53]	0.51 [0,1]	0.85 [0,1]	0.57 [0,1]	0 [0,0.01]	2.17 [1,5.55]
MFP1df	low	0.30 [0.07,0.52]	0.29 [0.06,0.53]	0.54 [0.07,1]	0.84 [0.33,1]	0.60 [0.11,1]	0 [0,0.01]	2.4 [1,6]
GAM	low	0.30 [0.08,0.53]	0.29 [0.05,0.53]	0.53 [0,1]	0.57 [0,1]	0.49 [0,1]	0.01 [0,0.03]	4.09 [1,9]
GAMboost	low	0.34 [0.14,0.54]	0.29 [0.05,0.53]	0.70 [0.2,1]	0.15 [0.04,0.4]	0.21 [0.07,0.45]	0.08 [0.05,0.11]	22.57 [17,28]
RF	low	0.24 [0.04,0.45]	0.24 [0.03,0.47]	0.58 [0.1,1]	0.26 [0.04,0.67]	0.29 [0.07,0.67]	0.05 [0,0.17]	14.79 [3,40]
NNet	low	0.45 [0.25,0.64]	0.14 [0.01,0.35]	0.86 [0.47,1]	0.04 [0.01,0.11]	0.07 [0.01,0.19]	0.52 [0.15,1]	123.74 [40,237]

Supplementary table 5.3: Performance measures of each model for different scenarios with linear associations in terms of the evaluation measures: cor^2, cor2Test, recall, precision, FPR, F-measure and Nvar (number of selected variables). Brackets indicate confidence interval based on 5% and 95% percentiles from simulation results for each measure. The first block contains averaged results across all scenarios and the subsequent blocks contain averaged results across the different assumptions for R^2 and Cor.var (correlation levels between true predictors). ExWASsp: exposome-wide association study with natural cubic splines regressions, partDSA: regression trees model using the algorithm deletion substitution addition, MFP: multivariable fractional polynomial model using stepwise, MFP1df: MFP with one degree of freedom, GAM: generalized additive splines model using backfitting, GAMboost: generalized additive model using boosting, RF: random forest using an implemented variable selection step and NNet: neural network with an implemented variable selection step.

Model		cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp		0.35 [0.1,0.6]	0.26 [0.04,0.52]	1 [1,1]	0.26 [0.02,1]	0.32 [0.03,1]	0.09 [0,0.26]	21.22 [1,62]
partDSA		0.40 [0.24,0.57]	0.20 [0.01,0.43]	0.97 [1,1]	0.09 [0.07,0.12]	0.17 [0.13,0.22]	0.04 [0.03,0.05]	10.66 [8,13]
MFP		0.30 [0.08,0.52]	0.29 [0.06,0.54]	0.91 [0,1]	0.87 [0,1]	0.88 [0,1]	0 [0,0]	2.13 [2,3]
MFP1df		0.02 [0.02,0.02]	0 [0,0.02]	0.04 [0,0]	0.04 [0,0]	0.04 [0,0]	0 [0,0]	1.01 [1,1]
GAM		0.31 [0.09,0.53]	0.29 [0.06,0.54]	0.85 [0,1]	0.61 [0,1]	0.68 [0,1]	0 [0,0.01]	1.77 [1,4]
GAMboost		0.34 [0.14,0.54]	0.29 [0.05,0.54]	0.97 [1,1]	0.06 [0.04,0.08]	0.11 [0.08,0.14]	0.07 [0.05,0.1]	17.63 [13,24]
RF		0.24 [0.04,0.46]	0.24 [0.03,0.48]	1 [1,1]	0.19 [0.02,0.5]	0.30 [0.05,0.67]	0.05 [0,0.17]	12.01 [2,40]
NNet		0.30 [0.19,0.47]	0.05 [0,0.44]	0.54 [0,1]	0.04 [0,0.5]	0.05 [0,0.67]	0.47 [0,1]	112.01 [2,237]
Model	R^2	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	0.1	0.15 [0.08,0.23]	0.08 [0.02,0.15]	1 [1,1]	0.35 [0.03,1]	0.41 [0.06,1]	0.05 [0,0.14]	12.67 [1,35]
partDSA	0.1	0.26 [0.23,0.29]	0.03 [0,0.07]	0.95 [0,1]	0.08 [0,0.09]	0.14 [0,0.17]	0.05 [0.04,0.06]	12.31 [11,14]
MFP	0.1	0.10 [0.07,0.14]	0.10 [0.04,0.16]	0.86 [0,1]	0.81 [0,1]	0.83 [0,1]	0 [0,0.01]	2.11 [2,3]
MFP1df	0.1	0.02 [0.02,0.02]	0 [0,0.01]	0 [0,0]	0 [0,0]	0 [0,0]	0 [0,0]	1 [1,1]
GAM	0.1	0.11 [0.08,0.15]	0.09 [0.04,0.15]	0.8 [0,1]	0.59 [0,1]	0.65 [0,1]	0 [0,0.01]	1.66 [1,3]
GAMboost	0.1	0.16 [0.13,0.19]	0.09 [0.04,0.15]	0.95 [1,1]	0.05 [0.04,0.06]	0.09 [0.07,0.12]	0.08 [0.06,0.1]	20.47 [16,25]
RF	0.1	0.07 [0.03,0.11]	0.06 [0.02,0.1]	1 [1,1]	0.08 [0.02,0.2]	0.14 [0.03,0.33]	0.09 [0.02,0.24]	22.16 [5,57]
NNet	0.1	0.28 [0.2,0.37]	0 [0,0.01]	0.54 [0,1]	0 [0,0.01]	0.01 [0,0.02]	0.57 [0.24,1]	135.04 [57,237]
ExWASsp	0.3	0.36 [0.28,0.46]	0.25 [0.14,0.36]	1 [1,1]	0.23 [0.02,1]	0.29 [0.03,1]	0.09 [0,0.25]	22.74 [1,59]
partDSA	0.3	0.40 [0.37,0.44]	0.19 [0.12,0.26]	0.98 [1,1]	0.10 [0.08,0.12]	0.18 [0.15,0.22]	0.04 [0.03,0.05]	9.99 [8,12]
MFP	0.3	0.30 [0.25,0.34]	0.29 [0.2,0.38]	0.94 [0,1]	0.90 [0,1]	0.91 [0,1]	0 [0,0]	2.11 [2,3]
MFP1df	0.3	0.02 [0.02,0.02]	0.01 [0,0.01]	0.03 [0,0]	0.03 [0,0]	0.03 [0,0]	0 [0,0]	1 [1,1]
GAM	0.3	0.31 [0.26,0.35]	0.29 [0.2,0.37]	0.86 [0,1]	0.62 [0,1]	0.69 [0,1]	0 [0,0.01]	1.77 [1,4]
GAMboost	0.3	0.33 [0.29,0.37]	0.29 [0.2,0.37]	0.98 [1,1]	0.06 [0.05,0.07]	0.11 [0.09,0.13]	0.07 [0.05,0.08]	17.18 [13,21]
RF	0.3	0.23 [0.18,0.28]	0.23 [0.15,0.32]	1 [1,1]	0.18 [0.04,0.5]	0.29 [0.07,0.67]	0.04 [0,0.11]	9.68 [2,28]
NNet	0.3	0.28 [0.19,0.37]	0.02 [0,0.22]	0.57 [0,1]	0.03 [0,0.33]	0.05 [0,0.5]	0.50 [0.01,1]	119.18 [2.95,237]
ExWASsp	0.5	0.55 [0.48,0.63]	0.45 [0.32,0.56]	1 [1,1]	0.20 [0.01,1]	0.26 [0.03,1]	0.12 [0,0.29]	28.25 [1,69]
partDSA	0.5	0.54 [0.5,0.58]	0.37 [0.27,0.47]	0.99 [1,1]	0.10 [0.08,0.12]	0.19 [0.15,0.22]	0.04 [0.03,0.05]	9.68 [8,12]
MFP	0.5	0.50 [0.46,0.53]	0.49 [0.4,0.57]	0.94 [0,1]	0.89 [0,1]	0.91 [0,1]	0 [0,0]	2.17 [2,4]
MFP1df	0.5	0.02 [0.02,0.02]	0 [0,0.02]	0.08 [0,1]	0.08 [0,1]	0.08 [0,1]	0 [0,0]	1.03 [1,1]
GAM	0.5	0.51 [0.46,0.55]	0.49 [0.4,0.58]	0.89 [0,1]	0.61 [0,1]	0.69 [0,1]	0 [0,0.01]	1.88 [1,4]
GAMboost	0.5	0.52 [0.48,0.56]	0.49 [0.4,0.58]	0.99 [1,1]	0.07 [0.05,0.08]	0.12 [0.1,0.15]	0.06 [0.05,0.08]	15.25 [12,19]
RF	0.5	0.43 [0.37,0.47]	0.42 [0.33,0.52]	1 [1,1]	0.33 [0.1,0.5]	0.47 [0.18,0.67]	0.01 [0,0.04]	4.2 [2,10]
NNet	0.5	0.32 [0.18,0.52]	0.13 [0,0.52]	0.50 [0,1]	0.08 [0,0.5]	0.10 [0,0.67]	0.34 [0,1]	81.81 [2,237]
Model	Cor.var	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	high	0.39 [0.14,0.62]	0.22 [0.03,0.48]	1 [1,1]	0.03 [0.01,0.07]	0.07 [0.03,0.13]	0.16 [0.06,0.28]	37.92 [14,67]
partDSA	high	0.40 [0.25,0.57]	0.19 [0.01,0.43]	0.97 [1,1]	0.09 [0.07,0.12]	0.17 [0.13,0.22]	0.04 [0.03,0.06]	10.86 [8,13.55]
MFP	high	0.30 [0.08,0.52]	0.29 [0.05,0.54]	0.87 [0,1]	0.81 [0,1]	0.83 [0,1]	0 [0,0.01]	2.18 [2,4]
MFP1df	high	0.02 [0.02,0.02]	0 [0,0.01]	0.05 [0,0.2]	0.05 [0,0.2]	0.05 [0,0.2]	0 [0,0]	1.02 [1,1]
GAM	high	0.31 [0.09,0.53]	0.29 [0.05,0.54]	0.81 [0,1]	0.55 [0,1]	0.63 [0,1]	0 [0,0.01]	1.91 [1,4]
GAMboost	high	0.34 [0.14,0.54]	0.29 [0.05,0.55]	0.96 [1,1]	0.06 [0.04,0.08]	0.11 [0.07,0.14]	0.07 [0.05,0.1]	17.9 [13,24]
RF	high	0.25 [0.05,0.46]	0.24 [0.03,0.49]	1 [1,1]	0.16 [0.02,0.5]	0.26 [0.05,0.67]	0.06 [0,0.17]	15.25 [2,40]
NNet	high	0.30 [0.18,0.49]	0.07 [0,0.46]	0.46 [0,1]	0.04 [0,0.5]	0.06 [0,0.67]	0.43 [0,1]	100.78 [2,237]
ExWASsp	low	0.31 [0.09,0.54]	0.29 [0.06,0.54]	1 [1,1]	0.49 [0.08,1]	0.57 [0.14,1]	0.01 [0,0.05]	4.52 [1,13]
partDSA	low	0.40 [0.24,0.57]	0.20 [0.01,0.43]	0.97 [1,1]	0.10 [0.08,0.12]	0.17 [0.14,0.22]	0.04 [0.03,0.05]	10.46 [8,13]
MFP	low	0.30 [0.08,0.52]	0.30 [0.07,0.54]	0.95 [1,1]	0.93 [0.25,1]	0.94 [0.4,1]	0 [0,0]	2.07 [2,2]
MFP1df	low	0.02 [0.02,0.02]	0.01 [0,0.02]	0.03 [0,0]	0.03 [0,0]	0.03 [0,0]	0 [0,0]	1 [1,1]
GAM	low	0.31 [0.09,0.53]	0.29 [0.06,0.54]	0.89 [0,1]	0.66 [0,1]	0.73 [0,1]	0 [0,0.01]	1.63 [1,3]
GAMboost	low	0.34 [0.14,0.54]	0.29 [0.06,0.54]	0.98 [1,1]	0.06 [0.04,0.08]	0.11 [0.08,0.15]	0.07 [0.05,0.09]	17.37 [12,23]
RF	low	0.24 [0.04,0.45]	0.23 [0.03,0.47]	1 [1,1]	0.23 [0.04,0.5]	0.34 [0.07,0.67]	0.03 [0,0.11]	8.78 [2,28]
NNet	low	0.30 [0.2,0.42]	0.04 [0,0.38]	0.62 [0,1]	0.03 [0,0.33]	0.05 [0,0.5]	0.52 [0.01,1]	123.23 [2,237]

Supplementary table 5.4: Performance measures of each model for different scenarios with U-shape associations in terms of the evaluation measures: cor^2, cor2Test, recall, precision, FPR, F-measure and Nvar (number of selected variables). Brackets indicate confidence interval based on 5% and 95% percentiles from simulation results for each measure. The first block contains averaged results across all scenarios and the subsequent blocks contain averaged results across the different assumptions for R^2 and Cor.var (correlation levels between true predictors). ExWASsp: exposome-wide association study with natural cubic splines regressions, partDSA: regression trees model using the algorithm deletion substitution addition, MFP: multivariable fractional polynomial model using stepwise, MFP1df: MFP with one degree of freedom, GAM: generalized additive splines model using backfitting, GAMboost: generalized additive model using boosting, RF: random forest using an implemented variable selection step and NNet: neural network with an implemented variable selection step.

Model		cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp		0.41 [0.11,0.68]	0.22 [0.02,0.5]	1 [1,1]	0.11 [0.01,1]	0.15 [0.02,1]	0.20 [0,0.45]	47.43 [1,108]
partDSA		0.42 [0.25,0.59]	0.21 [0.01,0.46]	0.95 [1,1]	0.08 [0.07,0.11]	0.15 [0.13,0.2]	0.05 [0.03,0.06]	11.67 [9,14]
MFP		0.29 [0.08,0.51]	0.27 [0.05,0.52]	0.9 [0,1]	0.81 [0,1]	0.84 [0,1]	0 [0,0.01]	2.3 [2,4]
MFP1df		0.19 [0.05,0.35]	0.19 [0.03,0.37]	0.77 [0,1]	0.75 [0,1]	0.76 [0,1]	0 [0,0.01]	1.13 [1,2]
GAM		0.30 [0.08,0.53]	0.29 [0.05,0.55]	0.81 [0,1]	0.44 [0,1]	0.52 [0,1]	0.01 [0,0.03]	2.96 [1,7]
GAMboost		0.33 [0.13,0.53]	0.29 [0.05,0.55]	0.97 [1,1]	0.06 [0.04,0.09]	0.11 [0.07,0.17]	0.07 [0.04,0.1]	17.87 [11,25]
RF		0.25 [0.05,0.46]	0.25 [0.03,0.5]	1 [1,1]	0.16 [0.02,0.5]	0.25 [0.05,0.67]	0.06 [0,0.17]	15.65 [2,40]
NNet		0.41 [0.26,0.56]	0.15 [0,0.5]	0.94 [0,1]	0.03 [0,0.1]	0.05 [0,0.18]	0.39 [0.04,1]	93.84 [10,237]
Model	R^2	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	0.1	0.19 [0.09,0.34]	0.06 [0.01,0.13]	1 [1,1]	0.20 [0.01,1]	0.26 [0.03,1]	0.11 [0,0.3]	28.12 [1,72]
partDSA	0.1	0.27 [0.24,0.3]	0.03 [0,0.07]	0.9 [0,1]	0.07 [0,0.08]	0.13 [0,0.15]	0.05 [0.05,0.06]	12.91 [12,14]
MFP	0.1	0.10 [0.07,0.13]	0.09 [0.03,0.15]	0.81 [0,1]	0.78 [0,1]	0.79 [0,1]	0 [0,0]	1.94 [1,3]
MFP1df	0.1	0.06 [0.04,0.09]	0.06 [0.02,0.11]	0.67 [0,1]	0.66 [0,1]	0.66 [0,1]	0 [0,0.01]	1.08 [1,2]
GAM	0.1	0.10 [0.06,0.14]	0.09 [0.04,0.15]	0.71 [0,1]	0.46 [0,1]	0.52 [0,1]	0.01 [0,0.02]	2.17 [1,5]
GAMboost	0.1	0.15 [0.12,0.18]	0.09 [0.03,0.14]	0.96 [1,1]	0.04 [0.03,0.06]	0.08 [0.07,0.11]	0.09 [0.07,0.11]	22.21 [18,27]
RF	0.1	0.07 [0.04,0.1]	0.06 [0.02,0.11]	1 [1,1]	0.06 [0.01,0.14]	0.11 [0.02,0.25]	0.12 [0.03,0.34]	28.42 [7,81]
NNet	0.1	0.31 [0.23,0.38]	0.02 [0,0.05]	0.90 [0,1]	0.01 [0,0.01]	0.02 [0,0.02]	0.54 [0.24,1]	129.29 [57,237]
ExWASsp	0.3	0.42 [0.29,0.56]	0.21 [0.09,0.34]	1 [1,1]	0.08 [0.01,0.33]	0.12 [0.02,0.5]	0.21 [0.01,0.43]	50.5 [3,103]
partDSA	0.3	0.42 [0.38,0.45]	0.19 [0.11,0.28]	0.97 [1,1]	0.08 [0.08,0.1]	0.15 [0.14,0.18]	0.05 [0.04,0.05]	11.66 [10,13]
MFP	0.3	0.29 [0.24,0.33]	0.27 [0.17,0.35]	0.93 [0,1]	0.86 [0,1]	0.88 [0,1]	0 [0,0.01]	2.3 [2,4]
MFP1df	0.3	0.19 [0.15,0.23]	0.19 [0.13,0.27]	0.8 [0,1]	0.78 [0,1]	0.79 [0,1]	0 [0,0.01]	1.16 [1,2]
GAM	0.3	0.30 [0.26,0.35]	0.29 [0.2,0.39]	0.84 [0,1]	0.46 [0,1]	0.55 [0,1]	0.01 [0,0.03]	2.86 [1,6]
GAMboost	0.3	0.32 [0.28,0.36]	0.29 [0.2,0.38]	0.97 [1,1]	0.06 [0.04,0.08]	0.11 [0.08,0.14]	0.07 [0.05,0.09]	17.88 [13,22]
RF	0.3	0.24 [0.19,0.29]	0.24 [0.16,0.34]	1 [1,1]	0.14 [0.04,0.33]	0.23 [0.07,0.5]	0.05 [0.01,0.11]	13.01 [3,28]
NNet	0.3	0.41 [0.31,0.49]	0.08 [0.02,0.16]	0.96 [1,1]	0.01 [0,0.02]	0.02 [0.01,0.05]	0.47 [0.17,1]	112.55 [40,237]
ExWASsp	0.5	0.61 [0.5,0.71]	0.39 [0.23,0.54]	1 [1,1]	0.05 [0.01,0.14]	0.08 [0.02,0.25]	0.27 [0.03,0.49]	63.66 [6.95,117]
partDSA	0.5	0.57 [0.54,0.61]	0.40 [0.3,0.5]	0.98 [1,1]	0.10 [0.08,0.11]	0.17 [0.15,0.2]	0.04 [0.03,0.05]	10.44 [9,12]
MFP	0.5	0.48 [0.44,0.52]	0.46 [0.35,0.56]	0.96 [1,1]	0.79 [0.25,1]	0.84 [0.4,1]	0 [0,0.01]	2.66 [2,5]
MFP1df	0.5	0.32 [0.28,0.36]	0.32 [0.24,0.4]	0.84 [0,1]	0.82 [0,1]	0.83 [0,1]	0 [0,0.01]	1.13 [1,2]
GAM	0.5	0.50 [0.46,0.54]	0.49 [0.38,0.58]	0.88 [0,1]	0.39 [0,1]	0.48 [0,1]	0.01 [0,0.03]	3.85 [1,8.05]
GAMboost	0.5	0.51 [0.47,0.54]	0.49 [0.39,0.58]	0.97 [1,1]	0.07 [0.06,0.11]	0.14 [0.11,0.2]	0.05 [0.03,0.07]	13.52 [9,17]
RF	0.5	0.43 [0.38,0.48]	0.44 [0.33,0.53]	1 [1,1]	0.28 [0.07,0.5]	0.42 [0.13,0.67]	0.02 [0,0.06]	5.53 [2,14]
NNet	0.5	0.52 [0.44,0.58]	0.36 [0.13,0.54]	0.96 [1,1]	0.06 [0.01,0.2]	0.11 [0.01,0.33]	0.16 [0.02,0.49]	39.68 [5,116]
Model	Cor.var	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	high	0.47 [0.19,0.7]	0.17 [0.01,0.41]	1 [1,1]	0.02 [0.01,0.03]	0.03 [0.02,0.06]	0.32 [0.13,0.47]	75.6 [32,112.55]
partDSA	high	0.42 [0.25,0.6]	0.21 [0.01,0.45]	0.94 [0,1]	0.08 [0,0.1]	0.15 [0,0.18]	0.05 [0.04,0.06]	11.89 [10,14]
MFP	high	0.29 [0.08,0.51]	0.27 [0.04,0.52]	0.86 [0,1]	0.72 [0,1]	0.76 [0,1]	0 [0,0.01]	2.49 [2,5]
MFP1df	high	0.19 [0.04,0.35]	0.19 [0.03,0.37]	0.63 [0,1]	0.62 [0,1]	0.62 [0,1]	0 [0,0.01]	1.2 [1,2]
GAM	high	0.30 [0.08,0.53]	0.28 [0.04,0.55]	0.77 [0,1]	0.29 [0,1]	0.39 [0,1]	0.01 [0,0.03]	3.85 [1,8]
GAMboost	high	0.33 [0.13,0.53]	0.28 [0.04,0.54]	0.95 [1,1]	0.06 [0.03,0.09]	0.11 [0.06,0.17]	0.07 [0.04,0.11]	18.12 [11,26]
RF	high	0.25 [0.05,0.47]	0.25 [0.03,0.5]	1 [1,1]	0.13 [0.02,0.42]	0.22 [0.03,0.59]	0.08 [0.01,0.24]	19.42 [2,57]
NNet	high	0.41 [0.25,0.56]	0.16 [0,0.5]	0.90 [0,1]	0.03 [0,0.1]	0.05 [0,0.18]	0.39 [0.04,1]	93.18 [10,237]
ExWASsp	low	0.34 [0.09,0.6]	0.26 [0.05,0.52]	1 [1,1]	0.21 [0.02,1]	0.27 [0.03,1]	0.08 [0,0.24]	19.26 [1,58.55]
partDSA	low	0.42 [0.25,0.59]	0.21 [0.01,0.47]	0.97 [1,1]	0.09 [0.07,0.11]	0.16 [0.13,0.2]	0.04 [0.03,0.06]	11.45 [9,14]
MFP	low	0.29 [0.08,0.51]	0.27 [0.05,0.53]	0.94 [0,1]	0.90 [0,1]	0.91 [0,1]	0 [0,0]	2.1 [1,4]
MFP1df	low	0.19 [0.05,0.34]	0.20 [0.04,0.38]	0.91 [0,1]	0.89 [0,1]	0.89 [0,1]	0 [0,0]	1.05 [1,1]
GAM	low	0.30 [0.08,0.52]	0.29 [0.06,0.56]	0.86 [0,1]	0.58 [0,1]	0.65 [0,1]	0.01 [0,0.02]	2.07 [1,5]
GAMboost	low	0.33 [0.14,0.53]	0.29 [0.05,0.55]	0.98 [1,1]	0.06 [0.04,0.09]	0.11 [0.08,0.17]	0.07 [0.04,0.1]	17.63 [11,25]
RF	low	0.24 [0.05,0.46]	0.24 [0.03,0.49]	1 [1,1]	0.19 [0.02,0.5]	0.29 [0.05,0.67]	0.05 [0,0.17]	11.88 [2,40]
NNet	low	0.42 [0.27,0.56]	0.15 [0,0.5]	0.99 [1,1]	0.03 [0,0.1]	0.05 [0.01,0.18]	0.40 [0.04,1]	94.51 [10,237]

Supplementary table 5.5: Performance measures of each model for different scenarios with r-shape associations in terms of the evaluation measures: cor^2, cor2Test, recall, precision, FPR, F-measure and Nvar (number of selected variables). Brackets indicate confidence interval based on 5% and 95% percentiles from simulation results for each measure. The first block contains averaged results across all scenarios and the subsequent blocks contain averaged results across the different assumptions for R^2 and Cor.var (correlation levels between true predictors). ExWASsp: exposome-wide association study with natural cubic splines regressions, partDSA: regression trees model using the algorithm deletion substitution addition, MFP: multivariable fractional polynomial model using stepwise, MFP1df: MFP with one degree of freedom, GAM: generalized additive splines model using backfitting, GAMboost: generalized additive model using boosting, RF: random forest using an implemented variable selection step and NNet: neural network with an implemented variable selection step.

Model		cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp		0.42 [0.11,0.69]	0.22 [0.02,0.48]	1 [1,1]	0.11 [0.01,1]	0.14 [0.02,1]	0.21 [0,0.46]	49.58 [1,110]
partDSA		0.42 [0.25,0.59]	0.21 [0.01,0.46]	0.94 [0,1]	0.08 [0,0.11]	0.15 [0,0.2]	0.05 [0.03,0.06]	11.65 [9,14]
MFP		0.30 [0.08,0.52]	0.30 [0.06,0.54]	0.87 [0,1]	0.84 [0,1]	0.85 [0,1]	0 [0,0]	1.82 [1,3]
MFP1df		0.22 [0.06,0.39]	0.22 [0.04,0.41]	0.79 [0,1]	0.77 [0,1]	0.78 [0,1]	0 [0,0.01]	1.13 [1,2]
GAM		0.31 [0.08,0.53]	0.29 [0.05,0.54]	0.81 [0,1]	0.42 [0,1]	0.50 [0,1]	0.01 [0,0.03]	3.14 [1,7]
GAMboost		0.34 [0.14,0.54]	0.29 [0.06,0.54]	0.97 [1,1]	0.05 [0.04,0.06]	0.09 [0.07,0.12]	0.08 [0.06,0.11]	20.74 [16,26]
RF		0.25 [0.05,0.47]	0.25 [0.03,0.49]	1 [1,1]	0.16 [0.02,0.5]	0.25 [0.03,0.67]	0.07 [0,0.24]	16.41 [2,57]
NNet		0.44 [0.27,0.61]	0.14 [0,0.4]	0.96 [1,1]	0.01 [0,0.04]	0.02 [0.01,0.07]	0.46 [0.11,1]	108.87 [28,237]
Model	R^2	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	0.1	0.20 [0.09,0.37]	0.06 [0.01,0.13]	1 [1,1]	0.19 [0.01,1]	0.24 [0.03,1]	0.12 [0,0.32]	30.23 [1,76.05]
partDSA	0.1	0.27 [0.24,0.3]	0.04 [0.01,0.08]	0.89 [0,1]	0.07 [0,0.08]	0.13 [0,0.15]	0.05 [0.05,0.06]	12.95 [12,14]
MFP	0.1	0.10 [0.07,0.13]	0.10 [0.04,0.16]	0.80 [0,1]	0.80 [0,1]	0.80 [0,1]	0 [0,0]	1.38 [1,2]
MFP1df	0.1	0.07 [0.05,0.1]	0.07 [0.03,0.13]	0.72 [0,1]	0.71 [0,1]	0.71 [0,1]	0 [0,0.01]	1.09 [1,2]
GAM	0.1	0.11 [0.07,0.14]	0.09 [0.04,0.15]	0.70 [0,1]	0.44 [0,1]	0.51 [0,1]	0.01 [0,0.02]	2.27 [1,5]
GAMboost	0.1	0.16 [0.13,0.19]	0.09 [0.04,0.15]	0.95 [0,1]	0.04 [0,0.05]	0.08 [0,0.1]	0.09 [0.08,0.11]	23.29 [19,28]
RF	0.1	0.07 [0.04,0.11]	0.06 [0.02,0.11]	1 [1,1]	0.06 [0.01,0.14]	0.11 [0.02,0.25]	0.12 [0.03,0.34]	28.99 [7,81]
NNet	0.1	0.33 [0.24,0.42]	0.02 [0,0.06]	0.91 [0,1]	0.01 [0,0.02]	0.02 [0,0.03]	0.55 [0.24,1]	130.25 [57,237]
ExWASsp	0.3	0.43 [0.3,0.57]	0.20 [0.09,0.32]	1 [1,1]	0.08 [0.01,0.33]	0.11 [0.02,0.5]	0.22 [0.01,0.45]	53.12 [3,107.05]
partDSA	0.3	0.42 [0.39,0.46]	0.20 [0.12,0.27]	0.96 [1,1]	0.08 [0.08,0.1]	0.15 [0.14,0.18]	0.05 [0.04,0.05]	11.7 [10,13]
MFP	0.3	0.30 [0.26,0.34]	0.30 [0.22,0.37]	0.88 [0,1]	0.85 [0,1]	0.86 [0,1]	0 [0,0.01]	1.93 [1,3]
MFP1df	0.3	0.22 [0.18,0.26]	0.22 [0.15,0.3]	0.80 [0,1]	0.80 [0,1]	0.80 [0,1]	0 [0,0.01]	1.13 [1,2]
GAM	0.3	0.31 [0.26,0.35]	0.29 [0.21,0.37]	0.85 [0,1]	0.43 [0,1]	0.52 [0,1]	0.01 [0,0.03]	3.11 [1,7]
GAMboost	0.3	0.33 [0.3,0.38]	0.29 [0.22,0.37]	0.97 [1,1]	0.05 [0.04,0.06]	0.09 [0.07,0.12]	0.08 [0.06,0.1]	20.46 [16,25]
RF	0.3	0.25 [0.21,0.3]	0.25 [0.17,0.33]	1 [1,1]	0.14 [0.02,0.33]	0.23 [0.05,0.5]	0.06 [0.01,0.17]	14.48 [3,40]
NNet	0.3	0.44 [0.34,0.53]	0.11 [0.04,0.19]	0.98 [1,1]	0.01 [0,0.02]	0.02 [0.01,0.05]	0.48 [0.17,1]	113.94 [40,237]
ExWASsp	0.5	0.61 [0.51,0.71]	0.38 [0.25,0.53]	1 [1,1]	0.05 [0.01,0.13]	0.07 [0.02,0.22]	0.27 [0.03,0.49]	65.38 [7.95,117]
partDSA	0.5	0.58 [0.54,0.61]	0.41 [0.32,0.49]	0.98 [1,1]	0.10 [0.08,0.11]	0.18 [0.15,0.2]	0.04 [0.03,0.05]	10.31 [9,12]
MFP	0.5	0.50 [0.46,0.54]	0.49 [0.42,0.57]	0.93 [0,1]	0.87 [0,1]	0.89 [0,1]	0 [0,0.01]	2.14 [2,3]
MFP1df	0.5	0.36 [0.32,0.4]	0.36 [0.28,0.45]	0.83 [0,1]	0.82 [0,1]	0.82 [0,1]	0 [0,0.01]	1.18 [1,2]
GAM	0.5	0.50 [0.46,0.54]	0.49 [0.41,0.56]	0.88 [0,1]	0.38 [0,1]	0.47 [0,1]	0.01 [0,0.03]	4.05 [1,9]
GAMboost	0.5	0.52 [0.48,0.56]	0.49 [0.41,0.57]	0.98 [1,1]	0.05 [0.04,0.07]	0.10 [0.08,0.13]	0.07 [0.06,0.09]	18.46 [14,23]
RF	0.5	0.44 [0.39,0.48]	0.44 [0.36,0.52]	1 [1,1]	0.27 [0.07,0.5]	0.41 [0.13,0.67]	0.02 [0,0.06]	5.76 [2,14]
NNet	0.5	0.56 [0.48,0.64]	0.28 [0.14,0.46]	0.99 [1,1]	0.02 [0.01,0.05]	0.04 [0.01,0.1]	0.35 [0.08,0.7]	82.42 [20,166]
Model	Cor.var	cor^2	cor2Test	Recall	Precision	F-measure	FPR	Nvar
ExWASsp	high	0.48 [0.21,0.7]	0.17 [0.01,0.4]	1 [1,1]	0.02 [0.01,0.03]	0.03 [0.02,0.06]	0.33 [0.14,0.48]	78.19 [34,115]
partDSA	high	0.42 [0.25,0.6]	0.21 [0.01,0.46]	0.94 [0,1]	0.08 [0,0.11]	0.15 [0,0.19]	0.05 [0.04,0.06]	11.84 [9.45,14]
MFP	high	0.30 [0.08,0.52]	0.29 [0.06,0.54]	0.82 [0,1]	0.77 [0,1]	0.78 [0,1]	0 [0,0.01]	1.86 [1,3]
MFP1df	high	0.22 [0.06,0.39]	0.22 [0.04,0.41]	0.66 [0,1]	0.65 [0,1]	0.65 [0,1]	0 [0,0.01]	1.22 [1,2]
GAM	high	0.31 [0.08,0.53]	0.29 [0.05,0.54]	0.77 [0,1]	0.28 [0,1]	0.37 [0,1]	0.01 [0,0.03]	4.15 [1,8]
GAMboost	high	0.34 [0.14,0.54]	0.29 [0.06,0.54]	0.96 [1,1]	0.05 [0.03,0.06]	0.09 [0.07,0.12]	0.09 [0.06,0.11]	21.08 [16,26]
RF	high	0.26 [0.05,0.47]	0.25 [0.03,0.49]	1 [1,1]	0.13 [0.02,0.33]	0.20 [0.03,0.5]	0.08 [0.01,0.24]	20.77 [3,57]
NNet	high	0.44 [0.26,0.61]	0.14 [0.01,0.39]	0.93 [0,1]	0.01 [0,0.04]	0.02 [0,0.07]	0.46 [0.11,1]	109.31 [28,237]
ExWASsp	low	0.35 [0.1,0.61]	0.26 [0.05,0.51]	1 [1,1]	0.20 [0.02,1]	0.26 [0.03,1]	0.08 [0,0.25]	20.97 [1,60]
partDSA	low	0.42 [0.25,0.59]	0.21 [0.01,0.46]	0.95 [1,1]	0.08 [0.07,0.11]	0.16 [0.13,0.2]	0.04 [0.03,0.06]	11.47 [9,14]
MFP	low	0.30 [0.08,0.52]	0.30 [0.07,0.55]	0.93 [0,1]	0.91 [0,1]	0.92 [0,1]	0 [0,0]	1.78 [1,2]
MFP1df	low	0.22 [0.06,0.39]	0.22 [0.04,0.41]	0.92 [0,1]	0.90 [0,1]	0.90 [0,1]	0 [0,0]	1.04 [1,1]
GAM	low	0.31 [0.08,0.53]	0.29 [0.06,0.54]	0.86 [0,1]	0.56 [0,1]	0.64 [0,1]	0.01 [0,0.02]	2.13 [1,5]
GAMboost	low	0.34 [0.14,0.54]	0.29 [0.06,0.54]	0.98 [1,1]	0.05 [0.04,0.07]	0.09 [0.07,0.12]	0.08 [0.06,0.11]	20.4 [15,26]
RF	low	0.25 [0.05,0.46]	0.25 [0.03,0.49]	1 [1,1]	0.19 [0.02,0.5]	0.29 [0.05,0.67]	0.05 [0,0.17]	12.05 [2,40]
NNet	low	0.45 [0.29,0.61]	0.14 [0,0.41]	0.99 [1,1]	0.01 [0,0.04]	0.03 [0.01,0.07]	0.46 [0.11,1]	108.43 [28,237]

Supplementary table 5.6: Performance measures of each model for different scenarios with J-shape associations in terms of the evaluation measures: cor^2, cor2Test, recall, precision, FPR, F-measure and Nvar (number of selected variables). Brackets indicate confidence interval based on 5% and 95% percentiles from simulation results for each measure. The first block contains averaged results across all scenarios and the subsequent blocks contain averaged results across the different assumptions for R^2 and Cor.var (correlation levels between true predictors). ExWASsp: exposome-wide association study with natural cubic splines regressions, partDSA: regression trees model using the algorithm deletion substitution addition, MFP: multivariable fractional polynomial model using stepwise, MFP1df: MFP with one degree of freedom, GAM: generalized additive splines model using backfitting, GAMboost: generalized additive model using boosting, RF: random forest using an implemented variable selection step and NNet: neural network with an implemented variable selection step.

Data Set	Method	ExWASsp	partDSA	MFP	GAM	GAMboost	RF	NNet
INMA_transcriptomics	ExWASsp	2.367	2.366	1.789	479	2.364	2.293	651
	partDSA		7.624	3.758	1.717	7.303	6.974	1.554
	MFP			3.759	942	3.732	3.576	866
	GAM				1.719	1.677	1.614	337
	GAMboost					7.306	6.752	1.504
	RF						6.977	1.463
	NNet							1.554
INMA_methylomics	ExWASsp	1.493	1.396	1.279	543	1.490	1.391	629
	partDSA		6.785	3.725	2.057	6.556	6.254	2.182
	MFP			4.033	1.402	3.963	3.706	1.508
	GAM				2.216	2.180	2.062	770
	GAMboost					6.978	6.334	2.348
	RF						6.544	2.164
	NNet							2.457

Supplementary table 5.7: Number of overlapping significative features (genes or CpGs) at an FDR adjusted p-value < 0.05 and $cor^2 > 0.33$ for the real data sets, INMA_transcriptomics and INMA_methylomics.

Chapter 6

ARE METHYLATION BETA-VALUES SIMPLEX DISTRIBUTED?

Nonell L, González J. Are methylation beta-values simplex distributed? bioRxiv. 2019 Sep 5;753459. DOI: 10.1101/753459

Chapter 7

HOMICS: BAYESIAN HIERARCHICAL MODELS TO ANALYZE *OMICS* DATA WITH PRIOR BIOLOGICAL KNOWLEDGE

HOmics: Bayesian hierarchical models to analyze *omics* data with prior biological knowledge
Lara Nonell and Juan R González
Submitted to Bioinformatics the 1st September 2019

7.1 Abstract

Motivation Incorporating biological information in *omics* association analyses may improve statistical power and reduce false positive results. Additionally, features of different *omics* data such as SNP genotypes, CpG methylation levels or gene expression can be considered as interconnected biological layers and their dependencies can affect the association with disease. Bayesian hierarchical regression models can easily accommodate biological knowledge and handle dependencies among different *omics* data.
Results: HOmics is a new R package that uses Bayesian hierarchical modelling to assess association between *omics* features and traits including prior biological knowledge. Several examples describing how to incorporate information about SNP, CpG or gene expression data are illustrated. In particular, our proposed model is used to assess association between gene expression and ovarian cancer, where the predicted miRNAs of each gene are incorporated as prior information in the association with the disease.
Availability: R package is available at
https://github.com/isglobal-brge/HOmics

7.2 Introduction

In the commitment of disentangling disease, *omics* data generated at cellular level such as genomics, transcriptomics or epigenomics combined with clinical data help us understand complex biological processes. There are several databases providing biological information about the relevance of *omics* features with regard to phenotypes. Additionally, these features are not independent, as the presence of a single SNP or CpG can affect the expression of a gene and have in turn an effect on phenotype. In this sense, *omics* can be regarded as interconnected hierarchical biological layers.

In single *omics* association analyses, each feature is associated independently with the phenotype of interest and features are then ranked according to some statistical criteria. Multiple *omics* data can be inte-

grated using 'parallel methods', in which each *omic* is analyzed separately and then somehow integrated; and 'hierarchical methods', where prior knowledge can be incorporated ([Wu et al., 2019]).

Hierarchical regression models adjust model parameter estimates using prior knowledge but they have been barely used in health sciences. There are some examples in the context of *omics* association studies ([Hung et al., 2004, Conti and Witte, 2003], [Denis and Tadesse, 2015]). However, there are no flexible tools to incorporate biological knowledge using hierarchical regression models. To fill this gap and based on the previous research performed by [Thomas et al., 2009], we have developed **HOmics**, an R package that uses hierarchical models to incorporate prior biological knowledge in *omics* association analyses. The models are fitted using Bayesian inference.

The package allows, for instance, to analyze SNP genotypes or CpG methylation levels incorporating information about their gene position. It can also be used in gene expression analyses where information about gene correlation with CpG beta-values is also available. The model can account for biological information where a set of features are analyzed in a single multivariate regression model. This empowers the user to analyze, e.g., functional pathways that contain several genes of interest. **HOmics** can deal with standard *omics* R/Bioconductor structures including *ExpressionSets* or *GenomicRatioSets*, making the package interoperable with other R/Bioconductor packages used, for instance, to create matrices with a priori biological knowledge. The package can deal with the analysis of both quantitative and qualitative traits and incorporates basic methods to easily visualize the results.

7.3 Method

Given k *omics* features measured in n samples and given some prior knowledge about these features; the relationship between the outcome (Y) and the set of features, affected by the prior knowledge, can be described as a hierarchical model with two levels (see Figure 7.1):

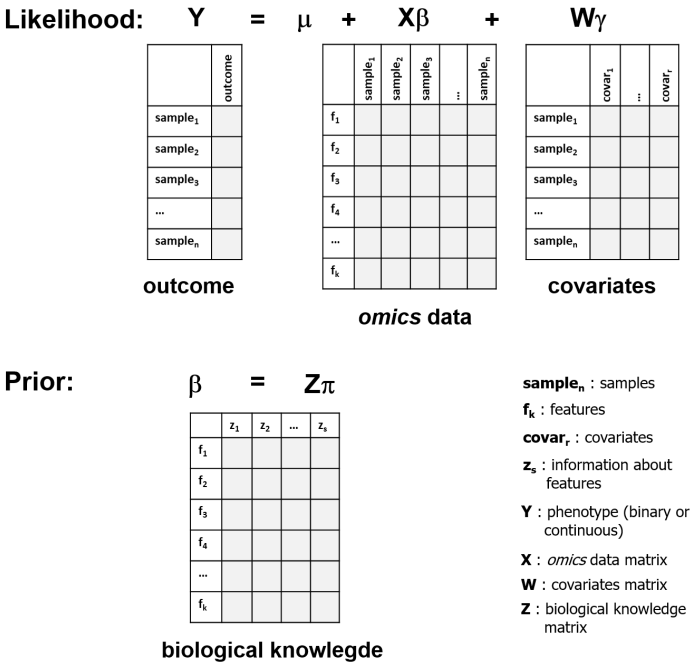


Figure 7.1: Bayesian hierarchical model and required matrices

First level (Likelihood)

$$Y = \mu + X\beta + W\gamma \tag{7.1}$$

where X encodes the *omics* data matrix of the k features measured in n samples and β are the model coefficients. W denotes the matrix of r covariates measured in the n samples with γ coefficients.

Second level (Prior - biological knowledge)

$$\beta = Z\pi + \epsilon, \epsilon \sim N(0, \sigma_\epsilon^2) \tag{7.2}$$

where Z is the matrix containing prior information for the features and π are the prior coefficients (Figure 7.1). Note that when Y is binary, the logit transformation is applied and generalized linear models used.

Models defined in equation (7.1) and equation (7.2) are approached through Bayesian inference formulated using JAGS.

7.4 Functions

The package contains two main functions, the generic **HOmics()** function and **HOmics.meth()**, which is specifically implemented to analyze methylation of genes affecting the outcome (see Supplementary File).

HOmics() function needs the following arguments: 1) **data.matrix** with the measurements of k features in n samples. 2) **cond**: phenotype vector of the n samples and 3) **z.matrix**: prior matrix of the k features. In addition, covariates can be included in the model using the argument **covar.matrix** and multivariate models can be indicated through the argument **agg.matrix**, where each group of features is fitted in a single model.

Results are presented as *S3* objects of class **HOmics** with several attributes including the results of each feature or group of features assessed. Standard *S3* methods **print** and **plot**, depicting the credible intervals, are also available. Besides, the package uses *S3* methods **signif()**, to get significant results, and **filter()**, rendering filtered results by coefficient direction (e.g. down- or up-regulated features). Internally, **HOmics** works with Bioconductor and tidyverse classes. `doParallel` and `foreach` packages deal with parallelization.

7.5 Examples

A wide range of hypothesis can be tested using **HOmics**. The Supplementary File includes several examples and the R code describing how to get prior information for different *omics* features. These examples are: incorporation of genic positions to SNP association analysis in univariate (example 1) or multivariate models (example 2); inclusion of relative positions of CpGs to the closest gene in methylation association to phenotype (example 3); and modulation of gene expression by the predicted miRNAs in the analysis of association with a trait of interest (example 4). In this last example, expression of twenty-four genes related to cancer

were measured in 39 women diagnosed with ovarian cancer (stage 4) and 8 healthy controls. Predicted miRNAs for analyzed genes were obtained from TargetScan and the **z.matrix** was created using different Bioconductor packages (see Supplementary File). Note that each gene is usually binded by several miRNAs and a miRNA may be predicted by several genes. The hierarchical model was fitted individually for each gene using **HOmics()** that accounts for prior information. Nine genes were found associated to the phenotype (see Figure 7 in Supplementary File). Standard analysis using `limma` returned only seven of the nine genes deregulated having the same effect direction (see Figure 8 in Supplementary File). This illustrates that hierarchical modelling increase the power of detecting important genes since we analyzed genes relevant to our disease of interest.

Example 5 in Supplementary File shows how to perform transcriptomic and methylomic integration by using the correlation coefficients between CpGs beta-values and the closest gene expression.

7.6 Conclusion

HOmics is an R package that easily incorporates prior biological knowledge in *omics* association analyses via hierarchical regression models, helping to improve statistical power and reduce false discovery rate. It uses a parallel Bayesian inference approach through JAGS to fit univariate or multivariate models. The developmental version of the package is available at https://github.com/isglobal-brge/HOmics. Future versions will include other examples illustrating how to incorporate information from other databases such as Roadmap Epigenomics, GTEx or TCGA.

7.7 Back matter

Acknowledgements

We acknowledge Xavier Basagaña for his statistical comments and thank Duncan C. Thomas for helpful conversations on hierarchical models. This

work has been supported by the Ministerio de Ciencia, Innovación y Universidades y Fondo Europeo de Desarrollo (RTI2018-100789-B-I00).

Supplementary material

The supplementary file is based on the **HOmics** package vignette and is appended in the following pages:

HOmics: Bayesian hierarchical models to analyze omics data with prior biological knowledge

Supplementary material

Lara Nonell and Juan R. González

Contents

1 Introduction

HOmics is an R package that allows to incorporate previous biological knowledge and enables omics integration.

HOmics uses a Bayesian approach that needs the JAGS (Just Another Gibbs Sampler) environment to be installed. This can be easily done through this link: http://mcmc-jags.sourceforge.net/.

To install **HOmics** use the following commands:

- Windows

```
library(devtools)
install_github("isglobal-brge/HOmics", INSTALL_opts=c("--no-multiarch"))
```

- Linux

```
library(devtools)
install_github("isglobal-brge/HOmics")
```

We illustrate the **HOmics** method with five examples:

1. SNP association studies where information about genic annotations are incorporated in the analyses in a univariate manner,
2. SNP association studies where information about genic annotations are incorporated in the analyses in a multivariate manner,
3. epigenomic gene studies where relative position to the closest gene is incorporated for each CpG using bioconductor's standard classes,
4. gene expression association to ovarian cancer including information about the predicted miRNAs and
5. integration of gene expression with methylation data using correlations in the association to phenotype

To get started let us load **HOmics** and `dplyr`, that will help in data manipulation:

```
library(HOmics)
library(dplyr)
```

2 Incorporating prior knowledge

2.1 Example 1: SNP association analyses

In the GWAS conext, each SNP is associated independently with the phenotype of interest. However, this conventional approach ignores existing information about the analyzed SNPs and assumes that they are all equally likely to impact the phenotype. Instead, one can incorporate information about the SNPs into a hierarchical model, in an attempt to improve the ranking of the p-values for association. There are several ways of incorporating a priori information. For instance, one can weight each SNP association p-value by how well it tags to SNPs that have been previously associated with our phenotype of interest as described in the GWAS catalog (https://www.ebi.ac.uk/gwas/). The genetic distance to these SNPs can also be used. Another option is to incorporate existing information about the SNPs into a second-stage design matrix Z. This information could be: conservation, functional category, gene location, tagging or even linkage. Here we illustrate how genic location depicted in Figure 1 can be used to improve single SNP assocation analyses.

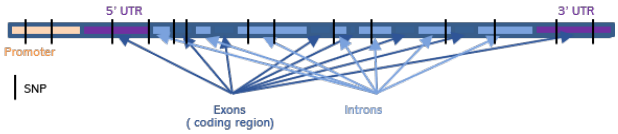


Figure 1: Genic annotation. Components of the gene with the SNP positions relative to the exonic and intronic gene composition.

Let us illustrate how to do the analyses with the first example, using a real data set on obesity.

We have genotypes for a total of 73 SNPs measured in 300 idividuals. Data can be loaded directly from the package by

```
data("obesity", package="HOmics")
```

We have two different objects, one for the genotypes and another for the phenotypic variables. Notice that the rownames of both objects perfectly match:

```
snps[1:6, 1:5]
```

```
     rs12921005 rs1420537 rs4784212 rs9925256 rs1420546
4180          1         2         2         1         1
4938          2         0         2         2         1
2405          1         2         0         2         2
323           2         2         2         0         2
4193          2         1         2         1         1
920           0         1         2         1         1
```

```
head(ob)
```

```
     gender obese age   smoke country
4180   Male     1  41 Current      50
4938   Male     0  44 Current      53
2405   Male     1  46 Current      55
323  Female     0  43 Current      53
```

```
4193 Female     0  49      Ex      53
920    Male     1  54 Current      50
```

```
identical(rownames(ob), rownames(snps))
```

```
[1] TRUE
```

Now the idea is to associate each SNP with the obesity status (0: normal, 1: obese) by incorporating information about the gene position of each SNP as described in Figure 1.

```
rsids <- colnames(snps)
head(rsids)
```

```
[1] "rs12921005" "rs1420537"  "rs4784212"  "rs9925256"  "rs1420546"
[6] "rs9302555"
```

```
length(rsids)
```

```
[1] 73
```

One can locate SNPs in and around genes by using some Bioconductor's packages as follows:

```
library(GenomicRanges)
library(VariantAnnotation)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(SNPlocs.Hsapiens.dbSNP144.GRCh37)

txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
snps.annot <- SNPlocs.Hsapiens.dbSNP144.GRCh37

snpPos <- snpsById(snps.annot, rsids)
snps.loc <- GRanges(seqnames = seqnames(snpPos),
                    IRanges(start=start(snpPos),
                            end=end(snpPos)),
                    rs=snpPos$RefSNP_id)

seqlevelsStyle(snps.loc) <- seqlevelsStyle(txdb)
genome(snps.loc) <- genome(txdb)

loc <- locateVariants(snps.loc, txdb,  AllVariants())
loc <- merge(loc, snps.loc, all.x=TRUE)
m <- findOverlaps(snps.loc, loc)
mcols(loc)[subjectHits(m), "rs"] <-
  mcols(snps.loc)[queryHits(m), "rs"]
loc.unique <- unique(loc)
```

This is the information we have after being processed the previous chunk, a GRanges object:

```
loc.unique[,c("LOCATION","rs")]
```

```
GRanges object with 74 ranges and 2 metadata columns:
       seqnames    ranges strand |   LOCATION          rs
          <Rle> <IRanges>  <Rle> |   <factor> <character>
   [1]    chr16  53125819      + |     intron   rs1108574
   [2]    chr16  53191470      + |     intron  rs12597487
   [3]    chr16  53197934      + |     intron   rs1421069
   [4]    chr16  53238253      + |     intron  rs10521279
   [5]    chr16  53283668      + |     intron   rs8058067
   ...      ...       ...    ... .        ...         ...
```

```
  [70]    chr16  54367420      * | intergenic   rs4783835
  [71]    chr16  54396362      * | intergenic   rs4784375
  [72]    chr16  54407804      * | intergenic    rs748815
  [73]    chr16  54414443      * | intergenic  rs12931564
  [74]    chr16  54415702      * | intergenic  rs11639567
  -------
  seqinfo: 25 sequences from hg19 genome; no seqlengths
```

Then the Z matrix providing a priori biological knowledge can be created using the following code. Notice that: 1) some SNPs may not be annotated; and 2) a given SNP may have 1 or more than one relative positions. This is addressed by adding Z matrix by the rownames. Notice that SNPs rs2908786 and rs3743772 are in two gene relative locations:

```
location <- droplevels(loc.unique$LOCATION)
Z <- model.matrix(~ 0 + location)
colnames(Z) <- levels(location)
rownames(Z) <- loc.unique$rs

Z <- t(sapply(by(Z,rownames(Z),colSums),identity))
dim(Z)
```

```
[1] 72  5
```

```
head(Z)
```

```
           intron threeUTR coding intergenic promoter
rs1004299       0        0      0          1        0
rs1013170       0        0      0          1        0
rs10521279      1        0      0          0        0
rs10521294      0        0      0          1        0
rs1108574       1        0      0          0        0
rs1112899       0        0      0          1        0
```

```
head(sort(apply(Z,1,sum), decreasing = TRUE))
```

```
 rs2908786  rs3743772  rs1004299  rs1013170 rs10521279 rs10521294
         2          2          1          1          1          1
```

Next code illustrates how to fit the hierachical model. We start by selecting from the omic matrix those features with annotated information.

```
rsids.ok <- rownames(Z)
omic.matrix <- as.matrix(t(snps[, rsids.ok]))
```

Then we create the W matrix which includes the covariates:

```
covar.matrix <- model.matrix( ~ gender + age, data=ob)
```

Finally the model is fitted by:

```
mod <- HOmics(data.matrix = omic.matrix,
              cond = as.factor(ob$obese),
              z.matrix = Z,
              covar.matrix = covar.matrix)
```

```
univariate analysis will be performed for each feature
cond is a factor, it has been converted to a numerical vector of 0s and 1s with 0 as the reference leve
```

The results are provided as an object of class *HOmics* with several attributes.

```
mod
```

```
Object of class HOmics
A hierarchical univariate model was fitted for each feature

Call:
HOmics(data.matrix = omic.matrix, cond = as.factor(ob$obese),
    z.matrix = Z, covar.matrix = covar.matrix)
```

```
attributes(mod)
```

```
$names
[1] "results" "call"    "univ"    "cont"

$class
[1] "HOmics"
```

Results can be extracted using the attribute results.

```
mod$results
```

```
[[1]]
# A tibble: 72 x 9
   feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos p.neg
   <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl>
 1 rs1004299  -0.04 -0.42  -0.04    0.34  1      827 0.412  0.588
 2 rs1013170  -0.02 -0.61  -0.01    0.6   1      196 0.486  0.514
 3 rs10521279 -0.08 -0.51  -0.08    0.34  1.01   523 0.356  0.644
 4 rs10521294  0.27 -0.41   0.25    1.06  1.04   110 0.767  0.233
 5 rs1108574  -0.03 -0.45  -0.03    0.38  1.01   647 0.446  0.554
 6 rs1112899  -0.1  -0.68  -0.1     0.47  1      184 0.370  0.630
 7 rs1125338  -0.22 -0.62  -0.21    0.19  1      879 0.146  0.854
 8 rs11639567 -0.44 -0.85  -0.43   -0.03  1      975 0.0168 0.983
 9 rs11642776  0.13 -0.64   0.12    1.08  1.04    83 0.619  0.381
10 rs11859998  0.25 -0.570  0.22    1.26  1.1     73 0.701  0.299
# ... with 62 more rows
```

which is a list of tibbles.

To visualize credible intervals we can use the plot function (Figure 2).

```
plot(mod)
```

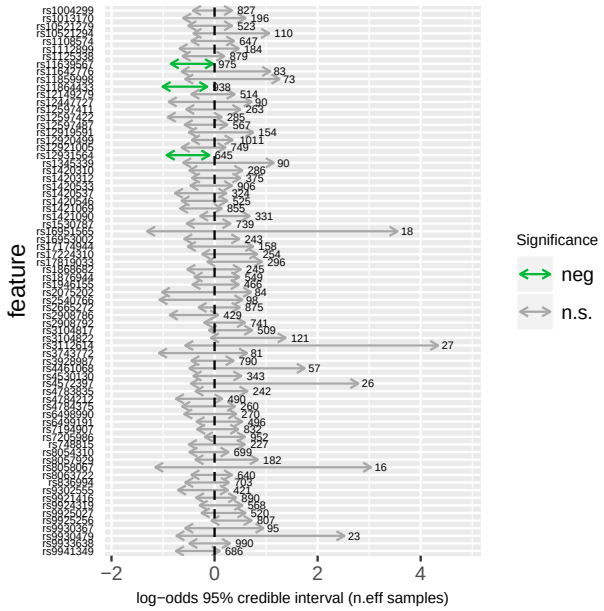


Figure 2: HOmics coefficients. Credible intervals for the assessed SNPs obtained with HOmics.

We then can compare these results with the results obtained from standard association analyses using `SNPassoc` package.

```
library(SNPassoc)
dd <- cbind(snps, ob)
ii <- grep("^rs", colnames(dd))
dd.s <- setupSNP(dd, colSNPs = ii,
                 name.genotypes = c(0,1,2))
ans <- WGassociation(obese, dd.s, model="log")
head(ans)
```

```
           comments log-additive
rs12921005        -      0.44593
rs1420537         -      0.51014
rs4784212         -      0.11534
rs9925256         -      0.11757
rs1420546         -      0.66599
rs9302555         -      0.34211
```

```
ans.odds <- odds(ans)
```

We can now plot `SNPassoc` log-odds confidence interval (Figure 3)

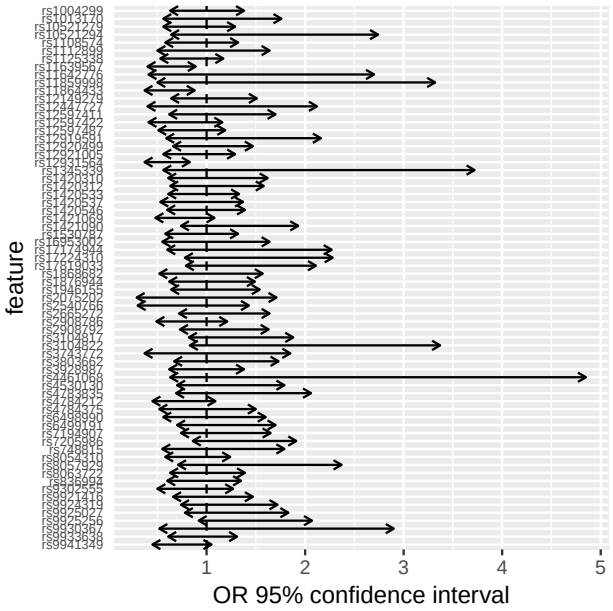


Figure 3: SNPassoc coefficients. Confidence intervals for the assessed SNPs obtained with SNPassoc.

And finally we can compare results by selecting those significative features by each method and transforming `SNPassoc` results to log-odds. Notice that objects are manipulated differently, as they are of different classes:

```
res.homics <- mod$results[[1]]
res.homics.sig <- res.homics %>%
                  filter(sign(`97.5%`) ==  sign(`2.5%`), n.eff>200) %>%
                  mutate(int.l =`97.5%`- `2.5%`)
res.homics.sig
```

```
# A tibble: 3 x 10
  feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos p.neg int.l
  <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
1 rs11639567 -0.44  -0.85 -0.43   -0.03  1      975 0.0168 0.983  0.82
2 rs11864433 -0.56  -1.01 -0.56   -0.14  1      938 0.003  0.997  0.87
3 rs12931564 -0.51  -0.93 -0.51   -0.1   1.01   645 0.0088 0.991  0.83
```

```
res.snpassoc <- ans.odds
res.snpassoc.sig <- res.snpassoc[res.snpassoc$`p-value.log-additive`<0.05 &
                                 !is.na(res.snpassoc$`p-value.log-additive`),]

res.snpassoc.sig$log.odds <- log(res.snpassoc.sig$OR)
res.snpassoc.sig$log.lower <- log(res.snpassoc.sig$lower)
res.snpassoc.sig$log.upper <- log(res.snpassoc.sig$upper)

res.snpassoc.sig$int.l <- res.snpassoc.sig$log.upper - res.snpassoc.sig$log.lower

res.snpassoc.sig[,c("log.odds","log.lower","log.upper")]
```

```
            log.odds  log.lower  log.upper
rs11864433 -0.5621189 -0.9942523 -0.1278334
rs12931564 -0.5798185 -0.9942523 -0.1863296
rs11639567 -0.5108256 -0.9162907 -0.1165338
```

Observe that significative SNPs associated to obesity are the same but the credible intervals (equivalent to confidence intervals) for **HOmics** approach have changed after including prior information.

2.2 Example 2: Multivariate SNP association analysis

Instead of performing a univariate analysis for each feature, we can include all SNPs in a multivariate model to study the association with phenotype. For that, the parameter agg.matrix must be specified with groups of features as depicted in Figure 4.

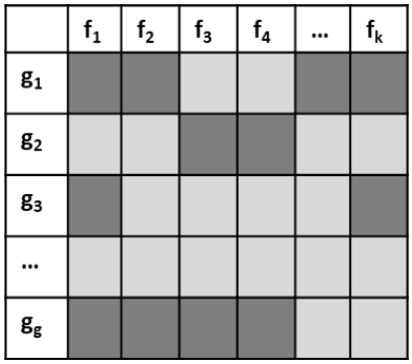


Figure 4: aggregation matrix. aggregation.matrix, where gi represents the group i and fi the feature i. For each feature fi in group gi, dark grey indicates belonging (1 in the matrix) and pale grey indicates non belonging (0 in the matrix).

In this case we will create a 1's matrix containing just one row representing a group and 73 columns, one for each SNP. We therefore consider all features included in group "g1".

```
agg.matrix <- matrix(data=1,nrow=1,ncol=nrow(omic.matrix))

rownames(agg.matrix) <-"g1"
colnames(agg.matrix) <- rownames(omic.matrix)

mod.multiv <- HOmics(data.matrix = omic.matrix,
               cond = as.factor(ob$obese),
               z.matrix = Z,
               covar.matrix = covar.matrix,
               agg.matrix = agg.matrix)
```

```
cond is a factor, it has been converted to a numerical vector of 0s and 1s with 0 as the reference leve
```

```
mod.multiv
```

```
Object of class HOmics
A hierarchical multivariate model was fitted for each group

Call:
HOmics(data.matrix = omic.matrix, cond = as.factor(ob$obese),
```

```
    z.matrix = Z, covar.matrix = covar.matrix, agg.matrix = agg.matrix)
```

If we want to assess more groups, it should be specified in the aggregation matrix, with each group as a row.

The results are again provided as an *HOmics* object, and results show this time the coefficients for the multivariate model

```
mod.multiv$results
```

```
$g1
# A tibble: 72 x 9
   feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos p.neg
   <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl>
 1 rs1004299   0.1  -0.37   0.1    0.56   1.01   672 0.654  0.346
 2 rs1013170  -0.04 -0.570 -0.04   0.49   1.01   280 0.446  0.554
 3 rs10521279  0.13 -0.46   0.12   0.76   1.02   281 0.652  0.348
 4 rs10521294  0.04 -0.53   0.04   0.62   1.03   221 0.560  0.440
 5 rs1108574   0.11 -0.31   0.11   0.55   1      649 0.687  0.313
 6 rs1112899  -0.15 -0.81  -0.14   0.5    1      168 0.325  0.675
 7 rs1125338  -0.28 -0.73  -0.28   0.14   1.01   765 0.0915 0.908
 8 rs11639567 -0.24 -0.75  -0.24   0.25   1      539 0.171  0.829
 9 rs11642776 -0.04 -0.68  -0.03   0.580  1      164 0.461  0.539
10 rs11859998 -0.06 -0.69  -0.06   0.62   1      165 0.425  0.575
# ... with 62 more rows
```

And we plot the coefficients of the multivariate model to visualize credible intervals (Figure 5).

```
plot(mod.multiv)
```

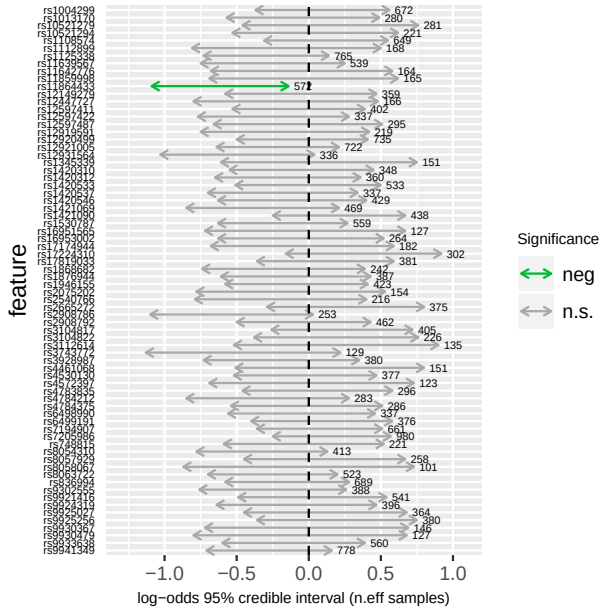


Figure 5: HOmics coefficients in the multivariate model. Credible intervals for the assessed SNPs obtained with HOmics in the multivariate model.

2.3 Example 3: CpG methylation analyses using bioconductor's classes

In methylomics, CpGs are studied individually in their association to phenotype. However, there is some information that can be added to the model such as their relative position to the gene. We could here perform a similar analysis to the one displayed in previous section but **HOmics** contains a specific function, **HOmics.meth()** that takes advantage of standard Bioconductor classes. The function is prepared to process ExpressionSet (`Biobase`) or GenomicRatioSet (`minfi`), standard classes when downloading GEO data using the `GEOquery` package. These objects have the following components:

- pheno data: information about the variables to analyze including phenotype and covariates
- annotation data: information about features (featureData for an ExpressionSet)

The relative position of a CpG to the closest gene is in this case the prior information and it is directly extracted from the annotation data of the object.

We will in this example assess an ExpressionSet, extracted from the GEO series GSE117929. Data was previously downloaded using package `GEOquery` and is accessible as the data object *GSE117929* in **HOmics**. `Biobase` package is needed to manipulate ExpressionSet class objects.

GSE117929 contains a methylome-wide analysis of 37 samples of peripheral blood mononuclear cells of systemic sclerosis patients (SSc, N=18) and normal controls (NC, N=19).

```
library(Biobase)

data("GSE117929", package="HOmics")

GSE117929
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 312028 features, 37 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM3315436 GSM3315437 ... GSM3315472 (37 total)
  varLabels: title geo_accession ... gender:ch1 (37 total)
  varMetadata: labelDescription
featureData
  featureNames: cg00000029 cg00000108 ... cg17014186 (312028
    total)
  fvarLabels: ID Name ... SPOT_ID (37 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL13534
```

```
table(pData(GSE117929)$"diagnosis:ch1")
```

```
 NC SSc
 19  18
```

The list of genes to model are obtained from PMC5988798, and we just call the function

```
genes <- c("CCR5","CXCR4")

mod.meth <- HOmics.meth(meth.data = GSE117929,
                    pheno.cond.col = "diagnosis:ch1",
                    annot.gene.col = "UCSC_RefGene_Name",
                    annot.z.col = "UCSC_RefGene_Group",
                    annot.mult.sep = ";",
```

```
                    gene.list = genes,
                    cores = 1)
```

```
some missing values were detected, only complete features will be selected
diagnosis:ch1 is a factor, it has been converted to a numerical vector of 0s and 1s with NC as the refe
```

```
mod.meth
```

```
Object of class HOmics
A hierarchical multivariate model was fitted for each group

Call:
HOmics.meth(meth.data = GSE117929, pheno.cond.col = "diagnosis:ch1",
    annot.gene.col = "UCSC_RefGene_Name", annot.z.col = "UCSC_RefGene_Group",
    annot.mult.sep = ";", gene.list = genes, cores = 1)
```

Let us see the results:

```
mod.meth$results[[1]]
```

```
# A tibble: 4 x 9
  feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff p.pos  p.neg
  <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl>
1 cg00803692 12.8   -0.09 12.1     30.6  1.43    21 0.973 0.027
2 cg04131610  6.6   -2.93  5.21    21.2  1.04    24 0.876 0.124
3 cg07616471  8.7   -3.05  8.03    21.7  1.12    23 0.920 0.0798
4 cg15239694  4.46  -6.29  4.09    17.3  1.17    35 0.843 0.157
```

We filter the results of those CpGs in genes with high probability of positive coefficients (betas) and also of negative coefficients in the adjusted Bayesian hierarchical model. In this example we filter at a significance level of 0.8 for demo purposes.

```
res.f.pos <- filter(mod.meth, param = "p.pos", threshold = 0.8, as.data.frame = T)
```

```
notice that this is a probability, so values above threshold will be selected
```

```
res.f.pos
```

```
# A tibble: 4 x 10
  group feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff p.pos  p.neg
  <chr> <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl>
1 CCR5  cg00803692 12.8   -0.09 12.1     30.6  1.43    21 0.973 0.027
2 CCR5  cg04131610  6.6   -2.93  5.21    21.2  1.04    24 0.876 0.124
3 CCR5  cg07616471  8.7   -3.05  8.03    21.7  1.12    23 0.920 0.0798
4 CCR5  cg15239694  4.46  -6.29  4.09    17.3  1.17    35 0.843 0.157
```

```
res.f.neg <- filter(mod.meth, param = "p.neg", threshold = 0.8, as.data.frame = T)
```

```
notice that this is a probability, so values above threshold will be selected
```

```
res.f.neg
```

```
# A tibble: 1 x 10
  group feature     mean `2.5%` `50%` `97.5%`  Rhat n.eff p.pos p.neg
  <chr> <chr>      <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
1 CXCR4 cg12595667 -6.66  -27.9 -3.61    0.99   1.2    29 0.102 0.898
```

We finally plot 95% credible interval and use method signif to obtain significant results, ie those features with credible intervals not containing 0 (Figure 6)

```
plot(mod.meth)
```

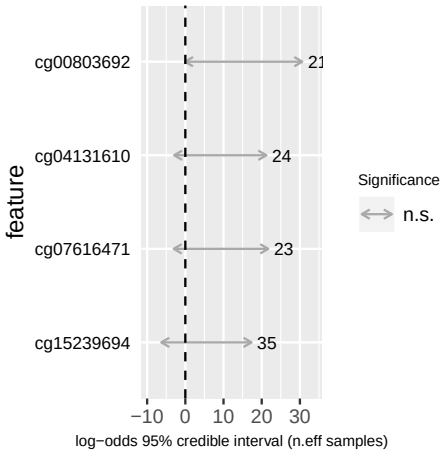


Figure 6: HOmics coefficients. Credible intervals for the assessed CpGs obtained with HOmics.

```
signif(mod.meth)
```

```
# A tibble: 0 x 0
```

We will now adjust the model by sex variable, which is specified in phenoData as 'gender:ch1'

```
genes <- c("CCR5","CXCR4")

mod.meth.sex <- HOmics.meth(meth.data = GSE117929,
                  pheno.cond.col = "diagnosis:ch1",
                  annot.gene.col = "UCSC_RefGene_Name",
                  annot.z.col = "UCSC_RefGene_Group",
                  annot.mult.sep = ";",
                  pheno.covar.col = "gender:ch1",
                  gene.list = genes,
                  cores = 1)
```

```
some missing values were detected, only complete features will be selected
diagnosis:ch1 is a factor, it has been converted to a numerical vector of 0s and 1s with NC as the refe
covariate gender:ch1 has been converted to a numerical vector
```

```
class(mod.meth.sex)
```

```
[1] "HOmics"
```

And we filter adjusted results

```
res.meth.f.sex.pos <- filter(mod.meth.sex)
```

```
notice that this is a probability, so values above threshold will be selected
```

```
res.meth.f.sex.pos
```

```
# A tibble: 0 x 0
```

```
res.meth.f.sex.neg <- filter(mod.meth.sex, param="p.neg")
```

```
notice that this is a probability, so values above threshold will be selected
```

```
res.meth.f.sex.neg
```

```
# A tibble: 0 x 0
```

2.4 Example 4: Ovarian cancer gene expression with targeted miRNAs

Gene expression can be modulated by the miRNAs, that bind to genes during transcription. We will see how this binding can affect the association with phenotype. For this example we will work with data from three different sources:

- ExpressionSet obtained from bicoconductor's package `curatedovarianData`
- Genes related to ovarian cancer obtained from The human protein atlas, ovarian cancer and available as a data object in **HOmics**
- Prior information of target miRNA downloaded from TargetScan website (v 7.2 http://www.targetscan.org/vert_72/) and available as a data object in **HOmics**

We will assess the gene expression association with ovarian cancer by fitting a hierarchical model where prior information known about genes are their predicted miRNAs.

```
# BiocManager::install("curatedOvarianData")

library(curatedOvarianData)
data(TCGA_eset)
TCGA_eset
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 13104 features, 578 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: TCGA.20.0987 TCGA.23.1031 ... TCGA.13.1819 (578
    total)
  varLabels: alt_sample_name unique_patient_ID ...
    uncurated_author_metadata (31 total)
  varMetadata: labelDescription
featureData
  featureNames: A1CF A2M ... ZZZ3 (13104 total)
  fvarLabels: probeset gene
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 21720365
Annotation: hthgu133a
```

Notice that TCGA_eset is an object of class ExpressionSet. We will compare ovarian cancer tumor samples in stage 4 with recurrence versus healthy samples. From the original TCGA_eset object, we will extract the expression data and the phenotype variables as follows

```
pheno <- pData(TCGA_eset)
table(pheno$sample_type)
```

```
healthy   tumor
      8     570
```

```
TCGA_subset <- TCGA_eset[,pheno$sample_type=="healthy" |
                           (pheno$sample_type=="tumor" &
                              pheno$tumorstage==4 &
                              pheno$recurrence_status=="recurrence")]

expr.m <- exprs(TCGA_subset)
dim(expr.m)
```

```
[1] 13104    47
```

```
pheno <- pData(TCGA_subset)
```

Let us obtain the condition vector

```
cond <- pheno$sample_type
table(cond)
```

```
cond
healthy   tumor
      8      39
```

The prior matrix with miRNAs that target our set of genes is obtained from the downloaded information obtained at TargetScan

```
data("targets.hs.7.2", package="HOmics")

targets
```

```
# A tibble: 718,233 x 11
   `miR Family` `Gene ID` `Gene Symbol` `Transcript ID` `Species ID`
   <chr>        <chr>     <chr>         <chr>                  <dbl>
 1 miR-23-3p    ENSG0000~ A1BG          ENST0000026310~         9606
 2 miR-23       ENSG0000~ A1BG          ENST0000026310~         9598
 3 miR-23-3p    ENSG0000~ A1BG          ENST0000026310~         9544
 4 miR-23       ENSG0000~ A1BG          ENST0000026310~         9615
 5 miR-302-3p/~ ENSG0000~ A1CF          ENST0000037400~         9544
 6 miR-302c-3p~ ENSG0000~ A1CF          ENST0000037400~         9606
 7 miR-520      ENSG0000~ A1CF          ENST0000037400~         9598
 8 miR-15/16/1~ ENSG0000~ A1CF          ENST0000037400~         9913
 9 miR-15/16/1~ ENSG0000~ A1CF          ENST0000037400~         9615
10 miR-153      ENSG0000~ A1CF          ENST0000037400~         9598
# ... with 718,223 more rows, and 6 more variables: `UTR start` <dbl>,
#   `UTR end` <dbl>, `MSA start` <dbl>, `MSA end` <dbl>, `Seed
#   match` <chr>, PCT <chr>
```

Let us filter the targets to get only those with probability of conserved targeting (PCT) > 0.5 and generate the prior information matrix

```
targets.f <- targets %>% mutate(PCT= as.numeric(PCT)) %>% filter(PCT >0.5)

z.table <- table(targets.f$`Gene Symbol`, targets.f$`miR Family`)

z.matrix <- as.matrix(as.data.frame.matrix(z.table))

z.matrix[z.matrix>1]<-1

table(z.matrix) # 0s and 1s only
```

```
z.matrix
      0       1
1714767  118666
```

Ovarian specific related genes are stored in a data object

```
data("ov.genes", package="HOmics")

head(ov.genes$Gene)
```

```
[1] "PDCL2"   "DEFB126" "RBPJL"   "OVGP1"   "COX8C"   "IMPG2"
```

To make sure that features (in this case genes) in the three data sets are the same, we do:

```
common.genes <- intersect(intersect(rownames(expr.m), ov.genes$Gene),
                          rownames(z.matrix))
length(common.genes)
```

```
[1] 24
```

```
expr.m <- expr.m[common.genes,]
z.matrix <- z.matrix[common.genes,]

mod.ov <- HOmics(data.matrix = expr.m,
              cond = cond,
              z.matrix = z.matrix)
```

```
univariate analysis will be performed for each feature
cond is a factor, it has been converted to a numerical vector of 0s and 1s with healthy as the referenc
```

```
mod.ov$results[[1]]
```

```
# A tibble: 24 x 9
   feature  mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos  p.neg
   <chr>   <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl>  <dbl>
 1 ALK      3.36   0.84  3.33    6.21  1.03    20 1      0
 2 CDH6     0.7    0.14  0.66    1.37  1.03    68 0.993  0.007
 3 CLIC5    1.86   0.91  1.82    2.99  1.02    43 1      0
 4 DOK5     0.44  -0.08  0.42    1.05  1.13    64 0.943  0.0568
 5 EMX2    -1.29  -2.47 -1.3    -0.31  1.12    17 0.0013 0.999
 6 ESR1    -2.53  -3.73 -2.81   -0.87  1.63    21 0      1
 7 FGF18    1.17   0.45  1.13    2.03  1.02    87 1      0
 8 GPR12   -0.22  -1.47 -0.26    1.32  1.05    74 0.342  0.658
 9 HOXD1   -0.14  -0.64 -0.14    0.3   1.02   105 0.285  0.715
10 HOXD3   -0.13  -0.63 -0.12    0.35  1.02   117 0.300  0.700
# ... with 14 more rows
```

```
signif(mod.ov)
```

```
# A tibble: 10 x 10
   group feature  mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos  p.neg
   <chr> <chr>   <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl>  <dbl>
 1 1     ALK      3.36   0.84  3.33    6.21  1.03    20 1      0
 2 1     CDH6     0.7    0.14  0.66    1.37  1.03    68 0.993  0.007
 3 1     CLIC5    1.86   0.91  1.82    2.99  1.02    43 1      0
 4 1     EMX2    -1.29  -2.47 -1.3    -0.31  1.12    17 0.0013 0.999
 5 1     ESR1    -2.53  -3.73 -2.81   -0.87  1.63    21 0      1
 6 1     FGF18    1.17   0.45  1.13    2.03  1.02    87 1      0
 7 1     PAX2    -0.77  -1.36 -0.76   -0.33  1.03   150 0      1
```

```
 8 1     PRSS21   0.85   0.3   0.84    1.46  1.01   180 0.999  0.001
 9 1     SOX11    1.33   0.03  1.27    3.15  1.12    59 0.980  0.0205
10 1     SOX17    1.18   0.08  1.13    2.6   1.02    18 0.982  0.018
```

We compare these results with those obtained by standard approaches.

limma is used to assess differential expression in the selected genes.

```
library(limma)
library(Biobase)

cond <-as.factor(cond)
design<-model.matrix(~0+cond)
colnames(design) <- gsub("cond","",colnames(design))

fit<-lmFit(expr.m,design)
contrast.matrix<-makeContrasts(tumor-healthy ,levels=design)
fit2<-contrasts.fit(fit,contrast.matrix)
fite<-eBayes(fit2)
tt<-topTable(fite,coef=1,number=Inf,adjust="BH",confint = TRUE)
res.limma <- tt[tt$adj.P.Val<0.05,]
res.limma
```

```
          logFC       CI.L       CI.R  AveExpr         t      P.Value
CLIC5   1.756459  1.0046932  2.5082247 7.794603  4.698622 2.258957e-05
PAX2   -3.014797 -4.4275924 -1.6020024 4.821611 -4.291351 8.641504e-05
ESR1   -1.295692 -1.9399870 -0.6513966 9.627492 -4.044193 1.909784e-04
PRSS21  2.083699  0.8537279  3.3136707 6.269091  3.406866 1.343950e-03
FGF18   1.613308  0.5824364  2.6441803 6.057279  3.147224 2.841231e-03
CDH6    1.378219  0.3162892  2.4401486 8.295964  2.609982 1.206470e-02
         adj.P.Val          B
CLIC5  0.0005421498  2.4981621
PAX2   0.0010369805  1.2249433
ESR1   0.0015278270  0.4761951
PRSS21 0.0080636971 -1.3484875
FGF18  0.0136379075 -2.0391419
CDH6   0.0482588085 -3.3498186
```

```
res.limma.unadj <- tt[tt$P.Value<0.05,]
res.limma.unadj
```

```
          logFC       CI.L       CI.R  AveExpr         t      P.Value
CLIC5   1.756459  1.0046932  2.5082247 7.794603  4.698622 2.258957e-05
PAX2   -3.014797 -4.4275924 -1.6020024 4.821611 -4.291351 8.641504e-05
ESR1   -1.295692 -1.9399870 -0.6513966 9.627492 -4.044193 1.909784e-04
PRSS21  2.083699  0.8537279  3.3136707 6.269091  3.406866 1.343950e-03
FGF18   1.613308  0.5824364  2.6441803 6.057279  3.147224 2.841231e-03
CDH6    1.378219  0.3162892  2.4401486 8.295964  2.609982 1.206470e-02
ALK     1.054318  0.2138109  1.8948254 4.303671  2.522582 1.504445e-02
         adj.P.Val          B
CLIC5  0.0005421498  2.4981621
PAX2   0.0010369805  1.2249433
ESR1   0.0015278270  0.4761951
PRSS21 0.0080636971 -1.3484875
FGF18  0.0136379075 -2.0391419
CDH6   0.0482588085 -3.3498186
ALK    0.0515809778 -3.5462168
```

Which are less results than those obtained by **HOmics**, even if non adjusted for multiple comparisons. Finally, let us plot the results to see how intervals get adjusted (Figure 7):

```
plot(mod.ov)
```

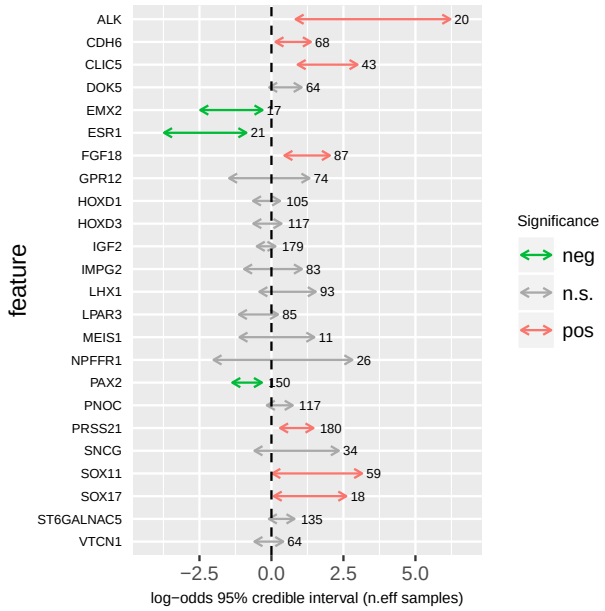


Figure 7: HOmics gene coefficients. Credible intervals for the assessed genes obtained with HOmics.

and `limma` (Figure 8).

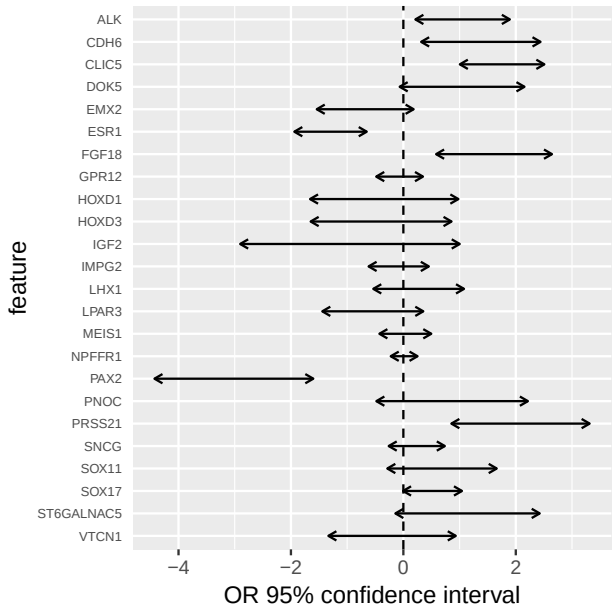


Figure 8: Limma gene coefficients. Confidence intervals for the assessed genes obtained with limma.

3 Integration of *omics* data

3.1 Example 5: Methylation beta-values integrated to gene expression

In this example, we will use GEO public data from a data superseries (GSE117931) containing methylation data (GSE117929), used in previous example, but also expression data (GSE117928) from the same set of samples. GSE117928 expression data was downloaded as an ExpressionSet using package `GEOquery` and is accessible as the data object *GSE117928* in **HOmics**.

We will integrate methylation of CpGs to the expression of the closest gene in the association with the phenotype variable, composed of 18 systemic sclerosis patients and 19 normal controls. Integration between gene expression and methylation is usually performed by correlations, as for instance, a hypermethylated site in the promoter of the gene can reduce its expression. Therefore, the prior information will be in this case the correlation between the gene and the CpGs.

We will take the same genes as in previous example and will construct a model for each of them.

First we load the data

```
library(Biobase)

data("GSE117928", package="HOmics")

GSE117928
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 29373 features, 37 samples
  element names: exprs
protocolData: none
```

```
phenoData
  sampleNames: GSM3315399 GSM3315400 ... GSM3315435 (37 total)
  varLabels: title geo_accession ... race:ch1 (38 total)
  varMetadata: labelDescription
featureData
  featureNames: ILMN_1343291 ILMN_1651209 ... ILMN_3311190 (29373
    total)
  fvarLabels: ID Transcript ... GB_ACC (28 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL14951
```

We need to extract and prepare the objects

```
genes <- c("CCR5","CXCR4")

expression <- log2(exprs(GSE117928)+24)
beta.values <- exprs(GSE117929)

fdata.exp <- fData(GSE117928)
fdata.meth <- fData(GSE117929)

cond <- pData(GSE117928)$"diagnosis:ch1"
table(cond)
```

```
cond
 NC SSc
 19  18
```

The expression matrix is given with the microarray IDs. To transform this into a matrix with Symbols, it is easy to use `dplyr` functions

```
expression.t<-as_tibble(expression)
expression.t <- left_join(fdata.exp %>% dplyr::select(ID,Symbol),
                      expression.t %>% mutate(ID = rownames(expression)),
                      by = 'ID')

expression.t <- expression.t %>%
                dplyr::select(-ID) %>%
                group_by(Symbol) %>%
                summarize_all(funs(mean))
```

Now, the construction of the hierarchical model for each gene will be done with a simple loop, where the z.matrix is constructed as a vector of correlations between the gene expression and beta-values of the CpGs in the region

```
n <- length(genes)
n
```

```
[1] 2
```

```
res.gene <- list()

for(i in 1:n){

  gene <- genes[i]

  gene.val <- unlist(expression.t %>% filter(Symbol==gene) %>% dplyr::select(-Symbol))
```

```
  cpgs.gene<- fdata.meth[grep(gene, fdata.meth$"UCSC_RefGene_Name"),
                         c("ID","UCSC_RefGene_Name" )]
  cpgs.val <- beta.values[cpgs.gene$ID,]

  data.matrix <- as.matrix(t(gene.val))
  rownames(data.matrix) <- gene

  z.matrix <- cor(gene.val,t(cpgs.val))
  rownames(z.matrix) <-gene

  res.gene[[i]] <- HOmics(data.matrix = data.matrix,
                          z.matrix = abs(z.matrix),
                          cond = cond)$results[[1]]
  names(res.gene)[i] <- gene
}
```

```
univariate analysis will be performed for each feature
cond is a factor, it has been converted to a numerical vector of 0s and 1s with NC as the reference lev
univariate analysis will be performed for each feature
cond is a factor, it has been converted to a numerical vector of 0s and 1s with NC as the reference lev
```

```
res.gene.t <- bind_rows(res.gene)

res.gene.t
```

```
# A tibble: 2 x 9
  feature  mean `2.5%` `50%` `97.5%`  Rhat n.eff  p.pos p.neg
  <chr>   <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl>
1 CCR5     0.28  -0.72  0.23    1.4   1.14    37 0.666  0.334
2 CXCR4   -0.8   -1.74 -0.79   -0.07  1.06    20 0.0133 0.987
```

If we compare these results to `limma` standard results

```
expression.sym <- data.frame(expression.t[,-1])
rownames(expression.sym) <- expression.t$Symbol

cond <-as.factor(cond)
design<-model.matrix(~0+cond)
colnames(design) <- gsub("cond","",colnames(design))

fit<-lmFit(expression.sym,design)
contrast.matrix<-makeContrasts(SSc-NC ,levels=design)
fit2<-contrasts.fit(fit,contrast.matrix)
fite<-eBayes(fit2)
tt<-topTable(fite,coef=1,number=Inf,adjust="BH",confint = TRUE)

tt[genes,c("logFC","CI.L","CI.R","P.Value","adj.P.Val")]
```

```
            logFC       CI.L       CI.R    P.Value adj.P.Val
CCR5   0.06567442 -0.2157401  0.3470890 0.63989864 0.8556032
CXCR4 -0.81055386 -1.4351193 -0.1859884 0.01224881 0.1299548
```

We see that, in this case, although significance is the same, credible intervals are wider in the **HOmics** model in comparison to `limma` confidence intervals.

Part III

Discussion

Chapter 8

GENERAL DISCUSSION

I will discuss the results obtained in this thesis, according to the different objectives. Some of these points have already been included in the discussions of the respective articles but I will here readdress them to provide a discussion in the global context of this thesis.

In science, we are ultimately interested in understanding nature, its biological processes and more precisely those related to human and disease to try improve them. The more knowledge we acquire the better management of diagnosis, treatment and prognosis of diseases. *Omics* world splits biology in layers that can be studied separately or as a whole. Each of the different *omics* details just a part of a biological process. In this thesis we have addressed several issues to gain insight into *omics* data through specific analyses. These topics are mainly related to data distribution and their association with phenotype.

Going deep into the *omics* data analysis, it is important to start by exploring and checking how data are distributed and their oddities. This is analogous to the analysis performed by statisticians in epidemiology but with some particularities. In epidemiology, each variable is explored and analysed with the most suitable statistical test. In contrast, *omics* analyses are performed at a high-throughput level, where there are thousands of features (genes, CpGs, SNPs, etc.), although each one is assessed independently. In this context, global exploration analyses are conducted

and usually the same assumptions are considered for all features in the subsequent analyses, inferring this way biased results. One of the simplifications that is generally done in high-throughput (whole genome) analyses, when studying association between variables, is the assumption of a linear relationship. Most of the analyses performed in this area assume a linear association and perform Pearson's correlation, or apply a linear regression model between two continuous variables. In addition, as demonstrated through the analysis of the first objective, these relationships are ideal but not real and several methods can be used in this context such as the multivariable fractional polynomials (MFP). We have demonstrated that MFP can be used in many contexts, despite real data heterogeneity. The results obtained in the article *Non-linear models to link exposome with omic data* apply to the association between continuous exposures (from the exposome) and any *omic* with continuous features such as transcriptomics or methylomics. However, this can be extended to any association performed between two continuous *omics* data features. Table 1.2 in the introduction summarizes the types of data that can benefit from these findings and the package that we have developed, *nlOmicAssoc*, will help the scientific community to explore and assess their data. The article was submitted some months ago to BMC Bioinformatics but is still under revision. The four reviewers selected by the editor have pointed out some interesting improvements, such as a new function in *nlOmicAssoc* to select the best method for each feature or testing the algorithms with other parameters. We are currently working on these topics.

Related to the *omics* data distribution, it has been a while since the publication of the first microarray quality control article with the objective of developing guidelines for transcriptomic microarray data analysis (MAQC-I, [Patterson et al., 2006]). Since then, normalization and `limma` analysis are some of the standard procedures to analyse transcriptomic data. Transcriptomic data originally obtained by microarrays were log_2 transformed and then linear models were applied. Later, with the emergence of NGS, which produces count data, transformations were developed, such as the voom transformation, to bring the count data to the continuous scale and apply similar methods. In methylomics, some assumptions are

actually taken on the data distribution that affect the analysis and more precisely their association with phenotype. Beta-values, the standard measurements of CpG methylation, represent the proportion of methylation in a CpG. Their distribution across samples is occasionally normal-like, sometimes bimodal and can even have marked peaks in 0 or 1 extremes. CpG measurements are usually assumed to be beta distributed -hence their name- but we have seen through the analyses written in the paper *Are methylation beta-values simplex distributed?* that they match in fact with a simplex distribution pattern most of the times. This conditions posterior analyses with regression models. We have demonstrated in this ground that simplex regression models but also linear models can be applied in many scenarios and that RRBS data are very scarce and variable. Quantile regression models, that *a priori* seemed suitable to identify differentially methylated sites, did not fulfill the expectations and no optimal results were obtained. This work, submitted currently to bioRxiv, will be submitted in the near future to BMC Epigenetics & Chromatin. We plan to create an R package with the developmental functions that are available at Github so that investigators can test the distribution of their data and apply the most appropriate model. In this regard, we will also add an improvement to all the assessed models so that covariates can be included in the analysis.

Finally, I did not want to conclude my research work without humbly contributing to the integromics field. One of the concerns about *omics* analyses and their integration is that there are interdependencies among *omics* data. Although many methods have been developed in this field, most of them are parallel or are focused in clustering [Wu et al., 2019]. Bayesian hierarchical regression models cope with *omics* dependencies by incorporating prior knowledge into the model. Given the lack of R tools to apply this kind of models, we have in this settings developed the R package *HOmics*. We present the package in the application note *HOmics: Bayesian hierarchical models to analyze omics data with prior biological knowledge*. This was initially developed to integrate methylation in the context of their gene positions but has been generalized to incorporate many kinds of previous knowledge into the hierarchical model. The application is available at Github and the paper has been submitted to

Bioinformatics as an application note. We will study the feasibility of including the tool in Bioconductor. The issue is that *HOmics* needs JAGS, so we have to carefully study this software dependency before submitting it to Bioconductor.

In this research, we have performed simulations but have also worked with real data. Simulations assessments are usually depicting consistent results, as shown in the non-linear analysis of association and also in the methylation data study. This is partly due to the fact that tests are performed in a controlled environment but real data sets tend to be more heterogeneous and it is not easy to control every situation in the simulations.

As general limitations I must mention the selection of methods, packages and functions that we have used; which were frequently chosen from a vast list of options. Although they have been often compared we can not claim that results would have been different if other options would have been used. The same applies to the function parameters, mostly applied with default options. We have also commented shortly about preprocessing procedures, which can really affect the distribution of data and consequently subsequent analyses and results. Because we were very conscious about this issue we tried to perform as many tests as possible, some of them not even included in the articles.

To conclude, despite there are some future tasks for each of the three objectives that will be developed in the coming months, the work reported in this thesis provides comprehensive discernments in the *omics* world. These new findings are my modest contribution to the scientific community (with the help of many people) and I am convinced that will inspire future works.

Chapter 9

CONCLUSIONS

There are many topics in the *omics* data distribution and analysis that can be studied to fully understand their data and find best methods to assess them. Here we have addressed some, related to non-linear associations, methylation data distribution and integration of *omics* information using Bayesian hierarchical models. These are the final conclusions of this thesis, arranged by objective:

- Related to the first objective:
  - Multivariable fractional polynomials show good performance when modelling non-linear associations between *omics* data and multiple exposures.
  - Multivariable fractional polynomials can also be applied in other types of studies having multiple continuous predictors.
  - *nlOmicAssoc* is an R package designed to analyse the association between continuous data such as the exposome measurements and other *omics* data. The package contains different functions to fit several methods and is programmed to work with the standard R objects such as matrix or data.frame but also with ExpressionSet or SummarizedExperiment, enabling the interaction with other Bioconductor packages.

- Related to the second objective:
  - Beta-values obtained from methylation microarrays are simplex distributed.
  - Distribution of methylation ratios obtained from NGS data depend on the type of sequencing (WGBS or RRBS) and on the data set. RRBS data are not easily addressed through regression models.
  - In the analysis of methylation data there are some recommendations that should be followed:
    1. Use data sets of at least 10 samples per studied condition for microarrays or 30 in NGS
    2. Apply a simplex or beta model in microarray data
    3. Apply a linear model in any other case
  - Quantile regression models do not fit beta-values as well as other regression models.
- Related to the third objective:
  - Hierarchical regression models can be applied to incorporate previous biological knowledge. They can be extended to account for dependencies among different *omics*.
  - The integration of prior data to Bayesian hierarchical regression models contribute to modulate model coefficients and reduce type I error.
  - *HOmics* is an easy to use R package that implements Bayesian estimations to capture *omics* data dependencies and can be applied to integrate prior knowledge and *omics* data. It works with standard R and Bioconductor classes.

List of abbreviations

A: Adenine

AIC: Akaike's information criterion

C: Cytosine

CN: Copy number

CDS: Coding sequencing

CTD: Comparative toxigogenomics database

DMS: Differentially methylated site

DNA: Deoxyribonucleic acid

ExWASsp: Exposome-wide association study with natural cubic splines regressions

FDR: False discovery rate

FN: False negatives

FP: False positives

G: Guanine

GAM: Generalized additive splines model using backfitting

GAMboost: Generalized additive model using boosting

GLM: Generalized linear model

GO: Gene ontology

GWAS: Genome wide association studies

KEGG: Kyoto encyclopedia of genes and genomes

KS: Kolmogorov-Smirnov test

LOH: Loss of heterozygosity

MFP: Multivariable fractional polynomial model using stepwise

MFP1df: Exposome-wide association study

MLE: Maximum likelihood estimation

MOM: Method of moments

MSigDB: Molecular signature database

NNet: Neural network with an implemented variable selection step

NGS: Next generation sequencing

partDSA: Regression trees model using the algorithm deletion substitution addition

RF: Random forest using an implemented variable selection step

RNA: Ribonucleic acid (mRNA: messenger RNA)

RNA-seq: RNA sequencing

RRBS: Reduced representation bisulfite sequencing

SNP: Single nucleotide polymorphism

T: Thymine

TCGA: The cancer genome atlas

TN: True negatives

TP: True positives

TSS: Transcription starting site

U: Uracil

WES: Whole exoome sequencing

WGBS: Whole genome bisulfite sequencing

WGS: Whole genome sequencing

Bibliography

[Agier et al., 2019] Agier, L., Basagaña, X., Maitre, L., Granum, B., Bird, P. K., Casas, M., Oftedal, B., Wright, J., Andrusaityte, S., de Castro, M., et al. (2019). Early-life exposome and lung function in children in europe: an analysis of data from the longitudinal, population-based helix cohort. *The Lancet Planetary Health*.

[Agier et al., 2016] Agier, L., Portengen, L., Chadeau-Hyam, M., Basagana, X., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., Gonzalez, J. R., Nieuwenhuijsen, M. J., Vineis, P., Vrijheid, M., Slama, R., and Vermeulen, R. (2016). A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. *Environ. Health Perspect.*, 124(12):1848–1856.

[Akalin et al., 2012] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology*, 13(10):R87.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular biology of the cell 4th edn (new york: Garland science). *Ann Bot*, 91:401.

[Allenby et al., 2005] Allenby, G. M., Rossi, P. E., and McCulloch, R. E. (2005). Hierarchical bayes models: A practitioners guide. ssrn scholarly paper id 655541. *Social Science Research Network, Rochester, NY*.

[Ambler and modified by Benner, 2015] Ambler, G. and modified by Benner, A. (2015). mfp: Multivariable fractional polynomials.

[Arivaradarajan and Misra, 2018] Arivaradarajan, P. and Misra, G. (2018). *Omics Approaches, Technologies And Applications*. Springer.

[Aryee et al., 2014] Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics*, 30(10):1363–1369.

[Assenov et al., 2014] Assenov, Y., Mueller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Compehensive analysis of dna methylation data with rnbeads. *Nature Methods*, 11(11):1138–1140.

[Barndorff-Nielsen and Jørgensen, 1991] Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of multivariate analysis*, 39(1):106–116.

[Barrera-Gómez et al., 2017] Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., Robinson, O., Vlaanderen, J., González, J. R., Nieuwenhuijsen, M., et al. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health*, 16(1):74.

[Barrett et al., 2012] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995.

[Bauer et al., 2016] Bauer, M., Fink, B., Thürmann, L., Eszlinger, M., Herberth, G., and Lehmann, I. (2016). Tobacco smoking differently influences cell types of the innate and adaptive immune system—indications from cpg site methylation. *Clinical epigenetics*, 8(1):83.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(1):289–300.

[Beyerlein, 2014] Beyerlein, A. (2014). Quantile regression—opportunities and challenges from a user's perspective. *American journal of epidemiology*, 180(3):330–331.

[Bonora et al., 2019] Bonora, G., Rubbi, L., Morselli, M., Ma, F., Chronis, C., Plath, K., and Pellegrini, M. (2019). Dna methylation estimation using methylation-sensitive restriction enzyme bisulfite sequencing (mrebs). *PloS one*, 14(4):e0214368.

[Braun et al., 2019] Braun, J. M., Kalloo, G., Kingsley, S. L., and Li, N. (2019). Using phenome-wide association studies to examine the effect of environmental exposures on human health. *Environment international*, 130:104877.

[Brazma et al., 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., et al. (2001). Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall, first edition.

[Buehlmann and Hothorn, 2007] Buehlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, 22(4):477–505.

[Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

[catalog, 2019] catalog, G. (2019). Gwas catalog. Accessed the 19th July 2019.

[Cavalcante and Sartor, 2016] Cavalcante, R. G. and Sartor, M. A. (2016). annotatr: Associating genomic regions with genomic annotations. *bioRxiv*, page 039685.

[Chen et al., 2010] Chen, Y. A., Almeida, J. S., Richards, A. J., Müller, P., Carroll, R. J., and Rohrer, B. (2010). A Nonparametric Approach to Detect Nonlinear Correlation in Gene Expression. *Journal of Computational and Graphical Statistics*, 19(3):552–568.

[Cobb, 2017] Cobb, M. (2017). 60 years ago, francis crick changed the logic of biology. *PLoS biology*, 15(9):e2003243.

[consortium, 2019] consortium, B. (2019). Blueprint epigenome. Accessed the 29th July 2019.

[Consortium et al., 2012] Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57.

[Conti and Witte, 2003] Conti, D. V. and Witte, J. S. (2003). Hierarchical modeling of linkage disequilibrum: genetic structure and spatial relations. *The American Journal of Human Genetics*, 72(2):351–363.

[Coro, 2017] Coro, G. (2017). Gibbs sampling with jags: Behind the scenes.

[CTD, 2019] CTD (2019). http://ctdbase.org/. Accessed the 29th July 2019.

[Davis et al., 2017a] Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. (2017a). The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.*, 45(D1):D972–D978.

[Davis et al., 2017b] Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., et al. (2017b). The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801.

[Delignette-Muller and Dutang, 2015] Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34.

[Denis and Tadesse, 2015] Denis, M. and Tadesse, M. G. (2015). Evaluation of hierarchical models for integrative genomic analyses. *Bioinformatics*, 32(5):738–746.

[Diaz-Uriarte and Alvarez de Andres, 2006] Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

[Diaz Zapata, 2018] Diaz Zapata, J. C. (2018). *ZOIP: ZOIP Distribution, ZOIP Regression, ZOIP Mixed Regression*. R package version 0.1.

[Dolzhenko and Smith, 2014] Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15(1):215.

[Draghici, 2003] Draghici, S. (2003). *Data analysis tools for DNA microarrays*. Chapman and Hall/CRC.

[ensembl, 2019] ensembl (2019). Ensembl. Accessed the 29th July 2019.

[Esteller, 2011] Esteller, M. (2011). Cancer epigenetics for the 21st century: what's next? *Genes & cancer*, 2(6):604–606.

[Ezkurdia et al., 2014] Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22):5866–5878.

[Faraway, 2016] Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

[Fave et al., 2018] Fave, M. J., Lamaze, F. C., Soave, D., Hodgkinson, A., Gauvin, H., Bruat, V., Grenier, J. C., Gbeha, E., Skead, K., Smargiassi, A., Johnson, M., Idaghdour, Y., and Awadalla, P. (2018). Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nat Commun*, 9(1):827.

[Feng et al., 2014] Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69.

[Ferrari and Cribari-Neto, 2004] Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

[Gasparrini et al., 2015] Gasparrini, A., Guo, Y., Hashizume, M., Kinney, P. L., Petkova, E. P., Lavigne, E., Zanobetti, A., Schwartz, J. D., Tobias, A., Leone, M., et al. (2015). Temporal variation in heat–mortality associations: a multicountry study. *Environmental health perspectives*, 123(11):1200.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

[Gevrey et al., 2003] Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables

in artificial neural network models. *Ecological Modelling*, 160:249–264.

[GO, 2019] GO (2019). Gene ontology. Accessed the 29th July 2019.

[Grosskopf, 2019] Grosskopf, M. (2019). Wikipedia. Accessed the 9th July 2019.

[Grün et al., 2012] Grün, B., Kosmidis, I., and Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, 48(11):1–25.

[Guxens et al., 2012] Guxens, M., Ballester, F., Espada, M., Fernández, M. F., Grimalt, J. O., Ibarluzea, J., Olea, N., Rebagliato, M., Tardón, A., Torrent, M., et al. (2012). Cohort profile: the inma—infancia y medio ambiente—(environment and childhood) project. *International journal of epidemiology*, 41(4):930–940.

[Hahne et al., 2010] Hahne, F., Huber, W., Gentleman, R., and Falcon, S. (2010). *Bioconductor case studies*. Springer Science & Business Media.

[Han et al., 2015] Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of rna sequencing and challenges. *Bioinformatics and biology insights*, 9:BBI–S28991.

[Hastie, 2018] Hastie, T. (2018). *gam: Generalized Additive Models*. R package version 1.15.

[Hastie and Tibshirani, 1986] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

[Hebestreit et al., 2013] Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.

[Hernandez-Ferrer and Gonzalez, 2018] Hernandez-Ferrer, C. and Gonzalez, J. R. (2018). CTDquerier: A Bioconductor R package for Comparative Toxicogenomics DatabaseTM data extraction, visualization and enrichment of environmental and toxicological studies. *Bioinformatics*.

[Heyn and Esteller, 2012] Heyn, H. and Esteller, M. (2012). Dna methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13(10):679.

[Hung et al., 2004] Hung, R. J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., and Witte, J. S. (2004). Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiology and Prevention Biomarkers*, 13(6):1013–1021.

[Ji et al., 2014] Ji, L., Sasaki, T., Sun, X., Ma, P., Lewis, Z. A., and Schmitz, R. J. (2014). Methylated dna is over-represented in whole-genome bisulfite sequencing data. *Frontiers in genetics*, 5:341.

[Jung and Pfeifer, 2015] Jung, M. and Pfeifer, G. P. (2015). Aging and dna methylation. *BMC biology*, 13(1):7.

[Kchouk et al., 2017] Kchouk, M., Gibrat, J.-F., and Elloumi, M. (2017). Generations of sequencing technologies: From first to next generation. *Biology and Medicine*, 9(3).

[KEGG, 2019] KEGG (2019). Kegg. Accessed the 29th July 2019.

[Kieschnick and McCullough, 2003] Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, 3(3):193–213.

[Koenker, 2018] Koenker, R. (2018). *quantreg: Quantile Regression*. R package version 5.36.

[Koenker and Bassett Jr, 1978] Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

[Law et al., 2014] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29.

[Leisch et al., 2016] Leisch, F., Hornik, K., and Ripley, B. D. (2016). mda: Mixture and flexible discriminant analysis.

[Lesnoff et al., 2012] Lesnoff, M., Lancelot, and R. (2012). *aod: Analysis of Overdispersed Data*. R package version 1.3.1.

[Li-Ping et al., 2014] Li-Ping, T., Li-Zhi, L., and Fang-Xiang, W. (2014). Nonlinear-model-based analysis methods for time-course gene expression data. *The Scientific World Journal*, page Article ID 313747.

[Liaw and Wiener, 2002] Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

[Lumley et al., 2002] Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, 23(1):151–169.

[Lunn et al., 2009] Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Stat Med*, 28(25):3049–3067.

[Maitre et al., 2018] Maitre, L., De Bont, J., Casas, M., Robinson, O., Aasvang, G. M., Agier, L., Andrušaitytė, S., Ballester, F., Basagaña, X., Borràs, E., et al. (2018). Human early life exposome (helix) study: a european population-based exposome cohort. *BMJ open*, 8(9):e021311.

[Matlin et al., 2005] Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5):386–398.

[May and Bigelow, 2005] May, S. and Bigelow, C. (2005). Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. *Dose-Response*, 3(4):dose–response.

[McGinnis et al., 2016] McGinnis, D. P., Brownstein, J. S., and Patel, C. J. (2016). Environment-Wide Association Study of Blood Pressure in the National Health and Nutrition Examination Survey (1999-2012). *Sci Rep*, 6:30373.

[Medvedev et al., 2010] Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome research*.

[Meng et al., 2016] Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics*, 17(4):628–641.

[Molaro et al., 2011] Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W. R., Hannon, G. J., and Smith, A. D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–1041.

[Molinaro and Lostritto, 2010] Molinaro, A. and Lostritto, K. (2010). Statistical bioinformatics: a guide for life and biomedical science researchers. *Statistical resampling for large screening data analysis such as classical resampling, Bootstrapping, Markov chain Monte Carlo, and statistical simulation and validation strategies. John Wiley & Sons, Inc., Hoboken, New Jersey*, pages 219–248.

[Molinaro et al., 2010] Molinaro, A. M., Lostritto, K., and van der Laan, M. (2010). partdsa: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. *Bioinformatics*, 26(10):1357–1363.

[mSigDB, 2019] mSigDB (2019). http://software.broadinstitute.org/gsea/msigdb. Accessed the 29th July 2019.

[Müller et al., 2019] Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2019). Rnbeads 2.0: comprehensive analysis of dna methylation data. *Genome biology*, 20(1):55.

[Nair et al., 2018] Nair, S. S., Luu, P.-L., Qu, W., Maddugoda, M., Huschtscha, L., Reddel, R., Chenevix-Trench, G., Toso, M., Kench, J. G., Horvath, L. G., et al. (2018). Guidelines for whole genome bisulphite sequencing of intact and ffpet dna on the illumina hiseq x ten. *Epigenetics & chromatin*, 11(1):24.

[Niemi, 2016] Niemi, J. (2016). Jarad niemi youtube channel. Accessed in November 2018.

[Olden et al., 2004] Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178:389–397.

[Park and Wu, 2016] Park, Y. and Wu, H. (2016). Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453.

[Patel et al., 2010] Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). An environment-wide association study (ewas) on type 2 diabetes mellitus. *PLoS ONE*, 5(5):e10746.

[Patterson et al., 2006] Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., Collins, P. J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E. S., Hager, J., Tikhonova, I. R., et al. (2006). Performance comparison of one-color and two-color platforms within the microarray quality control (maqc) project. *Nature biotechnology*, 24(9):1140.

[Pidsley et al., 2013] Pidsley, R., Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C. (2013). A data-driven approach to

preprocessing illumina 450k methylation array data. *BMC genomics*, 14(1):293.

[Plummer et al., 2003] Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 10. Vienna, Austria.

[Qu and Xu, 2006] Qu, Y. and Xu, S. (2006). Quantitative Trait Associated Microarray Gene Expression Data Analysis. *Mol Biol Evol*, 23(8):1558–1573.

[R Core Team, 2016] R Core Team, u. (2016). R: A language and environment for statistical computing.

[Raineri et al., 2014] Raineri, E., Dabad, M., and Heath, S. (2014). A note on exact differences between beta distributions in genomic (methylation) studies. *PLoS One*, 9(5):e97349.

[research network, 2019] research network, T. (2019). The cancer genome atlas. Accessed the 29th July 2019.

[Rigby and Stasinopoulos, 2005] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

[Ritchie et al., 2015] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

[rnbeads.org, 2018] rnbeads.org (2018). Rnbeads. Accessed in November 2018.

[Robinson et al., 2014] Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, 5:324.

[Royston and Altman, 1994] Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society*, 43(3):429–467.

[Royston et al., 1999] Royston, P., Ambler, G., and Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol*, 28(5):964–974.

[Saxonov et al., 2006] Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417.

[Shafi et al., 2017] Shafi, A., Mitrea, C., Nguyen, T., and Draghici, S. (2017). A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in bioinformatics*, 19(5):737–753.

[Sheffield et al., 2017] Sheffield, N. C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., et al. (2017). Dna methylation heterogeneity defines a disease spectrum in ewing sarcoma. *Nature medicine*, 23(3):386.

[Singmann et al., 2015] Singmann, P., Shem-Tov, D., Wahl, S., Grallert, H., Fiorito, G., Shin, S.-Y., Schramm, K., Wolf, P., Kunze, S., Baran, Y., et al. (2015). Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & chromatin*, 8(1):43.

[Smyth, 2004] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.

[Solvang et al., 2011] Solvang, H. K., Lingjærde, O. C., Frigessi, A., Børresen-Dale, A.-L., and Kristensen, V. N. (2011). Linear and non-linear dependencies between copy number aberrations and mrna expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics*, 12(1):1–12.

[Song, 2007] Song, P. X.-K. (2007). *Correlated data analysis: modeling, analytics, and applications*. Springer Science & Business Media.

[Song, 2009] Song, P. X.-K. (2009). Dispersion models in regression analysis. *Pakistan Journal of Statistics*, 25(4).

[Storey et al., 2005] Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842.

[Sun et al., 2014] Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). Moabs: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38.

[Sun et al., 2015] Sun, Z., Cunningham, J., Slager, S., and Kocher, J.-P. (2015). Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*, 7(5):813–828.

[Teschendorff and Relton, 2018] Teschendorff, A. E. and Relton, C. L. (2018). Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, 19(3):129.

[Thomas et al., 2009] Thomas, D. C., Conti, D. V., Baurley, J., Nijhout, F., Reed, M., and Ulrich, C. M. (2009). Use of pathway information in molecular epidemiology. *Human genomics*, 4(1):21.

[Tsaprouni et al., 2014] Tsaprouni, L. G., Yang, T.-P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., Viñuela, A., Grundberg, E., Nelson, C. P., Meduri, E., et al. (2014). Cigarette smoking reduces dna methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics*, 9(10):1382–1396.

[ucsc, 2019] ucsc (2019). Ucsc genome browser. Accessed the 29th July 2019.

[Van Steen and Malats, 2015] Van Steen, K. and Malats, N. (2015). Perspectives on data integration in human complex disease analysis. In *Big Data Analytics in Bioinformatics and Healthcare*, pages 284–322. IGI Global.

[Vandenberg et al., 2012] Vandenberg, L. N., Colborn, T., Hayes, T. B., Heindel, J. J., Jacobs Jr, D. R., Lee, D.-H., Shioda, T., Soto, A. M., vom Saal, F. S., Welshons, W. V., et al. (2012). Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. *Endocrine reviews*, 33(3):378–455.

[Venables and Ripley, 2002] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, fourth edition.

[Vineis et al., 2017] Vineis, P., Chadeau-Hyam, M., Gmuender, H., Gulliver, J., Herceg, Z., Kleinjans, J., Kogevinas, M., Kyrtopoulos, S., Nieuwenhuijsen, M., Phillips, D., et al. (2017). The exposome in practice: design of the exposomics project. *International journal of hygiene and environmental health*, 220(2):142–151.

[Vrijheid et al., 2014] Vrijheid, M., Robinson, O., Basagaña Flores, X., Bustamante Pineda, M., Casas, M., Estivill, X., van Gent, D., González, J. R., Júlvez Calvo, J., Kogevinas, M., et al. (2014). The human early-life exposome (helix): project rationale and design. *Environmental Health Perspectives. 2014; 122 (6): 535-544*.

[Wang et al., 2015] Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J. C., and Chen, W. (2015). A systematic study of normalization methods for infinium 450k methylation data using whole-genome bisulfite sequencing data. *Epigenetics*, 10(7):662–669.

[Wild, 2005] Wild, C. P. (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure mea-

surement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, 14(8):1847–1850.

[Wild, 2012] Wild, C. P. (2012). The exposome: from concept to utility. *International journal of epidemiology*, 41(1):24–32.

[Willyard, 2018] Willyard, C. (2018). New human gene tally reignites debate. *Nature*, 558(7710):354–356.

[Wu et al., 2019] Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019). A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4.

[www.nature.com, 2019] www.nature.com (2019). Nature. Accessed the 19th July 2019.

[Xiao et al., 2017] Xiao, X., Gasparrini, A., Huang, J., Liao, Q., Liu, F., Yin, F., Yu, H., and Li, X. (2017). The exposure-response relationship between temperature and childhood hand, foot and mouth disease: A multicity study from mainland China. *Environ Int*, 100:102–109.

[Yadav, 2007] Yadav, S. P. (2007). The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT*, 18(5):277.

[Yee, 2019] Yee, T. W. (2019). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-1.

[Yong et al., 2016] Yong, W.-S., Hsu, F.-M., and Chen, P.-Y. (2016). Profiling genome-wide dna methylation. *Epigenetics & chromatin*, 9(1):26.

[Yousefi et al., 2015] Yousefi, P., Huen, K., Davé, V., Barcellos, L., Eskenazi, B., and Holland, N. (2015). Sex differences in dna methylation assessed by 450 k beadchip in newborns. *BMC genomics*, 16(1):911.

[Zhan et al., 2011] Zhan, H., Chen, X., and Xu, S. (2011). A stochastic expectation and maximization algorithm for detecting quantitative trait-associated genes. *Bioinformatics*, 27(1):63–69.

[Zhang et al., 2016] Zhang, P., Qiu, Z., and Shi, C. (2016). simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software*, 71(11):1–21.

[Zhou et al., 2011] Zhou, Y.-H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of rna sequence count data. *Bioinformatics*, 27(19):2672–2678.