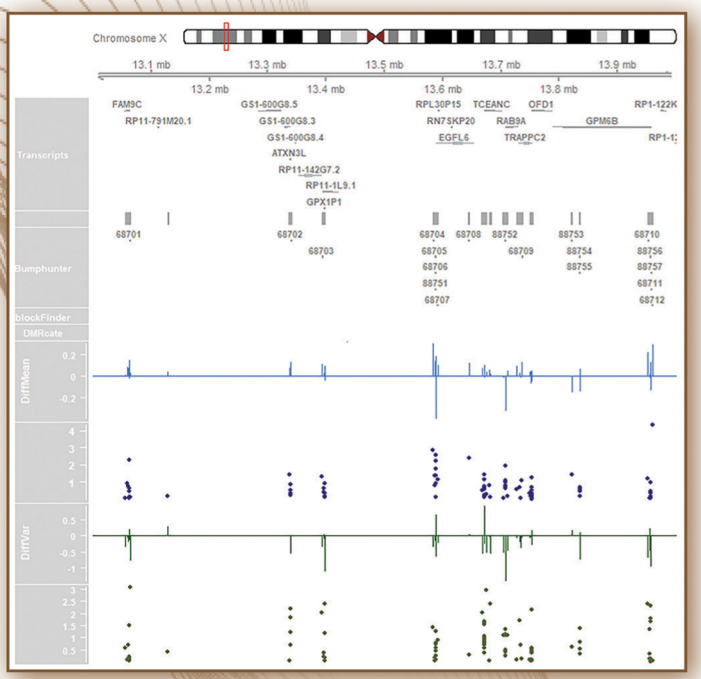
Omic Association Studies with R and Bioconductor



Juan R. González
Alejandro Cáceres

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Omic Association Studies with R and Bioconductor

Omic Association Studies with R and Bioconductor

Juan R. González
Alejandro Cáceres

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
International Standard Book Number-13: 978-1-138-34056-5 (Hardback)

To our families.

Contents

Preface

The aim of the book is to offer a practical guide to researchers, graduate students and those interested in the analysis of *omic* data. While our emphasis is on the use of data in publicly available repositories, the reader interested in analyzing novel data will find settled methods for inquiring into high-dimensional biological data. We have conceived the book as a first reference to tackle specific types of data, as well as a textbook for a bioinformatics course at the MSc level. Our objective is to demonstrate how to analyze genomic, transcriptomic, epigenomic and exposomic data to explain phenotypic differences among individuals. We describe the first analyses and methods of inquiry that should be used to identify the patterns in the data that associate with a trait of interest. During the past decade numerous methods have been developed and, due to the complexity of the data, we expect many more to be devised. Nonetheless, we describe some of the most established methods that are available in the Bioconductor and `R` repositories, which should constitute the first line of inquiry and to which future developments should be compared against.

The methods and applications described here are all publicly available and are accessible to anyone comfortable with fitting a linear regression model in `R`. While we direct the reader to numerous introductory books in `R` and basic statistical methods, the present book is directed to users. From a basic user level, we aim to guide the readers to expand their toolkit in order to deal with *omic* data with confidence.

All the methods discussed here are part of our daily toolkit. We are regular users of all the methods and are also developers of many of them. The book is the result of compiling workshop and class material, of software package development and of years of research carried out in Juan R. González's Bioinformatics Group in Genetic Epidemiology, within ISGlobal. We have thus developed expertise in the use of the methods and in their communication, and have realized the need to offer a guide to new researchers in the field. There is a wealth of publicly available software and data, yet the landscape is overwhelming to newcomers. We offer them starting points from which to begin inquiring into the *omic* data of interest. We do not offer a complete or global view but indicate safe up-to-date entry points. As developers of some of the packages discussed, we are committed, as part of the Bioconductor community, to offer clear and reproducible documentation, clarify doubts and update new versions. We insist that packages and pipelines to assist users are also implemented so they are further improved by other developers.

The material discussed in the book is largely based on cheap high-throughput methods. They include microarrays and some sequencing methods such as RNA-sequencing. We are also aware of the developments in the collection of new high-dimensional biological data, such as Next-Generation Sequencing or those aimed at single cells. There are, however, important advantages in the use and analysis of microarrays which will keep them relevant for many years. First, association studies require cohorts and technologies to be scalable to hundreds of thousands of individuals to properly power epidemiological inferences. Microarrays clearly meet the target. While we may conceive such scalability for future sequencing, the preprocessing of data may change but the basic methods of inference would likely remain the same. In addition, microarray data is widely available and it has been an important source of continuous reanalysis to test novel focused hypotheses, confirm new results or reproduce previous findings. Finally, SNPs arrays can be additionally used to explore other genomic variants, for which specific high-throughput technology is not yet available. Therefore, association analyses in large cohorts can be performed on inversion polymorphisms and mosaicism, including the loss of chromosome Y.

This book is the result of the joint effort with other colleagues whom we have collaborated throughout the years. We would like to explicitly acknowledge and thank Carles Hernández-Ferrer, Marcos López and Carlos Ruiz who have contributed with their ideas, work and coding hours to the `R` packages that we have developed at the BRGE and that are discussed within the book. We are thankful to them for starting their research careers with us and for the valuable input that they have given us through their PhD projects. Roger Pique-Regi is also acknowledged for his fruitful collaboration with the R-GADA package. We would also like to thank our colleagues and collaborators from whom we continuously learn, get encouragement and intellectual stimulation. We particularly would like to mention Luis Perez-Jurado, Mariona Bustamante, Xavier Basagaña, Manolis Kogevinas, Jordi Sunyer and Martine Vrijheid, and all our colleagues from ISGlobal. We also would like to thank Tonu Esko from the Estonia Biobank for providing access to large amounts of data to test our methods, when data sharing was not a standard procedure. Finally, we would like to acknowledge support from Ministerio de Economía y Competitividad y Fondo Europeo de Desarrollo (grant number MTM2015-68140-R), Ministerio de Ciencia e innovación (grant numbers MTM2011-26515 and MTM2010-09526-E) and Ministerio de Educación y Ciencia (grant number MTM2008-02457/MTM).

The material presented in the book has been conceived as complete analysis sessions, in which initial data is available and the reader can follow, step by step, the `R` commands that will lead to a concrete result. Concepts and theory are introduced and explained as we go along with the analysis demonstrations. As such, all data, software and code are freely available and can be accessed and reproduced in any platform. Most data can be downloaded from the Internet. Data from the main repositories can be accessed directly within an `R`

session, otherwise, we indicate functioning URLs at the time of publishing. Some functions have been implemented to add functionality to existing software. We have deposited them in the our GitHub repository which is publicy available at `https://github.com/isglobal-brge/book_omic_association`. Our GitHub repository (`https://github.com/isglobal-brge/`) also contains vignettes for most of the packages used in this book describing more detailed analyses. Also, the repository contains the most updated versions of the packages which include new features and bugs fixed. Specific instructions for data access and software needed are explained within each analysis demonstration.

1

Introduction

CONTENTS

1.1 Book overview

This book is concerned with the analysis of high dimensional data that is acquired at specific biological domains. The aim of the analyses is the explanation of phenotypic differences among individuals. We, therefore, search for endogenous and exogenous factors that may influence such differences. The endogenous domains on which we turn our attention are those at the molecular level involving basic DNA structure and function, which have been labeled with the *omic* suffix. In particular, we will describe current methods to analyze genomic data, which is high-dimensional at the gene (DNA sequence) level, transcriptomic data involving transcription of DNA into mRNA and epigenomic/methylomic data that relate to the epigenetic modifications of DNA. Many of the methods used at each domain overlap due to the biological nature and high dimensionality of data. However, important specificities remain, some derived from the acquisition of data and others from differences in the underlying biological processes. Within the exogenous domain, we study the high dimensional acquisition of exposure factors that are believed to influence the development or progression of individual traits.

1.2 Overview of *omic* data

Omic data refers to data collected massively in a specific *omic* domain. The notion of an unbiased scan of numerous biological entities of similar nature comes to mind. The definition is clearly operational, given the differences in understanding biological similarity and, perhaps more challenging, the varying characteristics of different biological levels. Genomic data, for instance, is ultimately concerned with the full characterization of the DNA sequence of an individual. As such, it is highly stable across tissues and the individual's lifespan. Some variations may arise in terms of somatic mutations that give rise to mosaicisms or to specific mutations, as found in tumorous cells. By contrast, transcriptomic or epigenomic data are highly variable across tissues, each of which changes on different time scales. Transcription data is highly dynamic and responds to physiological activity while epigenetic changes are expected to occur at developmental and aging rates.

An additional consideration is the differences in the expected coverage of each *omic* data or the data's dimensionality. Nowadays, for instance, one can expect from current technology that the complete DNA sequence of an individual may be determined, or estimated to high accuracy; and therefore, genomic data is close to full coverage. However, transcriptomic data is currently far from giving us the full picture: the complete set of transcripts of an individual in a given time across all cell types. While transcriptomic data is

clearly not complete, it is, however, a highly-dimensional unbiased-scan of a possible state of the transcriptome; that is, the complete set of transcripts in a biological sample of the individual.

Current extensions of *omic* data include metabolomics and proteomics, and other domains not strictly associated with specific molecular levels. These include, for instance, phenomic and exposomic data, which record multiple phenotypes and exposures at any level: molecular, organic or population. Studies including such data, therefore, allow high dimensionality on the response, traits or environmental conditions of individuals. Here, we will be concerned with studies of single phenotypes and conditions that are controlled or can be adjusted for covariates. We are primarily interested in describing subject variability on single phenotypes at a molecular level. We, therefore, study high dimensional data of DNA structure and function of groups of individuals, whose analysis methods show wide consensus. Some attention will also be given to exogenous factors given by exposomic data, which is a massive collection of environmental conditions in an unbiased manner.

1.2.1 Genomic data

The genome of an individual is the entire DNA content of all the individual's chromosomes. Genomic data comprises extensive and unbiased measurements of all the chromosomes' nucleotide sequences. Therefore, the highest possible dimensionality of genomic data is the number of nucleotides in the genome. However, it is the comparison between genomes what informs about their biological and meaningful substructures. As such, a collection of genomic data across individuals is based on the sequence variability of given structures.

1.2.1.1 Genomic SNP data

The simplest and most common structural variants in the genome are single nucleotide polymorphisms (SNPs). They are changes in only one nucleotide within a short DNA sequence that is otherwise conserved across individuals. The changes considered as SNPs are those given by only one substitution of a nucleotide for another, they are bi-allelic mutations and not rare in the population. Their allele frequencies are considered to be higher than 1%. SNPs can be detected with microarrays or sequencing techniques.

1.2.1.2 SNP arrays

Short DNA sequences, with their variant nucleotides at their ends, constitute probes that can be interrogated by its hybridization with the DNA of a given subject, which has been amplified, cut and marked with fluorescent dyes, one for each variant nucleotide or allele. Microarrays are scilico chips of millions of immobilized probes that capture the luminous DNA fragments of the subject, creating an optical pattern that is given by the individual's allele pairs, or genotypes, at each probe.

Different microarray technologies are used to genotype individuals with this approach, which is currently the most efficient and economical method to measure a substantial part of the genomic variability between individuals. The end result is an extensive coverage of SNP variants across the genomes of thousands/hundred-of-thousands of individuals. For large studies, the dimensionality of this data can achieve 10^5 (individuals) times 10^7 (SNPs), where the SNP variables are typically encoded as 0, 1 and 2 for annotated homozygous, heterozygous and variant homozygous, respectively. Annotations are complementary data on the genomic variables containing the two possible alleles at a given SNP; among adenine (A), thymine (T), cytosine (C) or guanine (C); the DNA strand, 5' to 3' (+) or 3' to 5' (-); and the alleles that should be considered as reference. Other specific considerations, that influence posterior analysis, include quality measurements of technical and biological conditions affecting SNPs and individuals.

A typical human SNP array assay includes a couple of millions of reference SNPs, from about 85 million SNPs existent in humans [23]. Neighboring SNPs are, however, highly correlated. Due to recombination, the correlation between SNPs diminishes with their distance but it is still substantial ($R^2 \sim 0.2$) for SNPs as far as 200,000 base pairs. Blocks of correlated SNPs, namely haplotypes, in reference populations have been used to impute the value of unmeasured SNPs and thus help to increase the number of SNPs of a particular study or facilitate the merging of genomic data from multiple studies [59]. The scalability of microarray-based studies is, therefore, their biggest asset to identify the likely small independent effects of numerous SNPs on complex traits [89].

SNP microarrays collect the genetic variability of individuals in known sequence variants. The known variants have been determined from reference population samples which have been fully sequenced. It remains to be determined the extent to which the selected references can offer a complete and unbiased coverage of different population samples. Despite the benefits of microarray genotyping, genome sequencing is still the ultimate source of information to fully define the genomic variability of individuals.

1.2.1.3 Sequencing methods

High-throughput sequencing methods aim to sequence all the DNA content of individuals. Broadly, in these methods, DNA is cut at small sizes ($\sim$ 100 base-pairs) or reads. Hundreds of millions of reads are then produced, which can cover the genome a number of times ($\sim$ 5/8), and need to be assembled to reconstruct an individual genome. Specific sequence variants of individuals can be estimated with high accuracy. The mapping of the reads of different individual genomes to a reference genome recovers genomic SNP data with the greatest coverage, unconditioned to ancestry. The scalability of genomic data, obtained from sequencing, is, however, limited. Current technology is expensive and computationally demanding and a suitable increase in the number of

individuals, required to detect the likely small effects of common variants, is at the moment unattainable.

Sequencing call of structural variants, therefore, remains an important tool to investigate rare variations and specific genomic architectures, while SNP arrays are most powerful in large studies of common genomic variation.

1.2.2 Genomic data for other structural variants

Genomic variation is rich, even between individuals with common recent ancestry. In a specific population, several DNA segments, of various lengths and up to the order of mega bases-pairs, can be found inserted, duplicated, deleted, translocated or inverted. While DNA sequencing is the best way to detect genomic variation, its price and analysis demands in large cohorts limit its use. SNP microarrays can, however, be exploited to detect many of these variants. For instance, luminous intensities used to genotype SNPs, can also be utilized to either detect regions with copy number alterations or cell populations with different genotypes (mosaicism)[167, 45]. In addition, specific haplotype patterns, which are produced by suppression of recombination, are indicative of mispairing between homologous chromosomes due to likely structural differences between them. Large and divergent haplotype groups have been associated with the suppression of recombination due to inversion polymorphisms. From genomic SNP data, inversion genotyping can be performed and their variability and functional impact can be studied in large cohorts [16].

Microarray SNP data opens the possibility to study more complex structural DNA variation in population samples across the genome. We can, therefore, exploit SNP data to have a more complete knowledge of genomic variability and to study the potential role of large structural variation in the phenotypic differences between individuals.

1.2.3 Transcriptomic data

Complex biochemical reactions are involved in the de-codification, or transcription, of DNA sequences. A direct product of these reactions is the production of RNA molecules some of which is further processed to produce proteins, the basic tools of the cells' physiology. Transcriptomic data is, therefore, a large-scale survey of the transcribed RNA repertoire of a biological sample.

The dimensionality of transcriptomic data is much smaller than that of genomic data. While in the production of a single RNA molecule, extensive and disjoint DNA may be involved, the structure of the molecule can be mapped to one gene. Genes constitute specific genomic regions of high variability in extent (up to 10^6 base-pairs) but cover in average 10,000 base-pairs of DNA sequence. In humans, the number of coding genes is estimated to be around 20,000. As such, the coding region of the genome, composed by all genes,

may represent only $2 - 8\%$ of the genome. Consequently, the dimensionality of transcriptomic, or gene expression, data can widely cover a transcriptome.

Unlike the genome, there are numerous transcriptomes per individual. Given that the RNA repertoire of a cell underlies its specific functions there is at least as many transcriptomes as cell types. In addition, transcriptomes are dynamic and therefore full coverage of the conditions that alter the transcriptome is currently limited.

1.2.3.1 Microarrays

Microarrays have been extensively used to study gene expression. Messenger RNA (mRNA), RNA destined to be translated into protein, is collected from a biological sample and used to synthesize complementary DNA (cDNA). cDNA is then amplified, cut and marked with fluorescent dye and hybridized on a chip containing probes, which are DNA segments of the genes' encoding known mRNAs. Luminous patterns are analyzed to measure the content of specific parts of mRNAs, as markers for their abundance.

Each mRNA maps to a gene but one gene can be mapped by numerous mRNAs. A gene encodes different mRNA transcripts, or isoforms, that are produced by the alternative splicing of the primary RNA, the molecule directly transcribed from the DNA sequence. Therefore, transcriptomic data, obtained from hybridization of gene probes is typically a mixture of the mRNA transcripts that map to a common gene. While specific junction probes can be designed to test the abundance of a particular mRNA transcript, the complete transcriptome, given by the abundance of all genes' isoforms, has to be derived from high dimensional data of probes that cover the entire structure of all the possible mRNA transcripts. Because the coding region of a gene is given by a set of disjoint sequences, exons, it is sufficient with one probe for one exon. Therefore, a transcriptome can be inferred from exon microarray data of hundreds of thousands of probes, as many as exons in the genome. While microarray data are giving way to other sequencing-based technologies, there is a large amount of available transcriptomic data that researchers can access for re-analysis. In addition, it is economically viable for many studies.

1.2.3.2 RNA-seq

Microarrays query specific points of mRNAs requiring *a priori* knowledge. An unbiased scan of the RNA repertoire in a biological sample is clearly a sequencing of its entire RNA content. RNA-seq is the application of high throughput sequencing to RNA. Reads are mapped to the gene exons of a reference genome. As there are numerous transcripts in the sample, the production of reads in a given region is a measure of the content of the mRNAs that is encoded by the region. Therefore, the read count per exon is the main output of this type of transcriptomic data. Again, the data is not the direct observation on the transcriptome but a mixture of the gene isoforms at each

exon. However, RNA-seq also provides junction reads that can inform on a particular mRNA transcript.

Compared to microarrays, RNA-seq allows detection of low expressed genes and genes with higher fold change between conditions. Furthermore, RNA-seq does not need *a priori* knowledge of probes, allows detection of genomic variants and it does not present problems like cross-hybridization. While RNA-seq is expensive and its analysis complex, transcriptomic data collection is transitioning from microarrays to RNA-seq. This is aided by the fact that expression differences between conditions (tissue or disease) are already detectable in hundreds of individuals and not in hundreds of thousands, as required by genomic studies.

1.2.4 Epigenomic data

The accessibility to DNA material is essential for the expression of genes. The way in which DNA is packed or its structure modified at a specific location can alter the fate of gene translation. For instance, the addition of a methyl group to the cytosine of a cytosine-guanine sequence 5'-3' (CpG), reduces the accessibility of DNA at this point by affecting the binding of proteins that promote transcription. DNA methylation has attracted much attention because it contributes to epigenetics (non-sequence modifications of DNA that are heritable), cell differentiation and cellular response to the environment at the genomic level [65].

Methylomic data is, therefore, the survey of the methyl modifications in the genome. Methylomic data is tissue-specific and dynamic through an organisms development and aging. Therefore, while the dimensionality of the data is bounded by the number of CpG sequences in the genome (1% in humans), a total covering of the methylomes of an individual is constrained by the number of cell types and the individual's age. CpG content is uneven across the genome and tends to concentrate in islands on the promoter regions of genes, which in humans are around 45,000 islands. Concentration on CpG islands and regions near genes can reduce the dimensionality of the data.

DNA methylation can be measured by means of DNA hybridization or sequencing. Similar to genomic and transcriptomic data, methylomic data can be obtained from microarray and high-throughput sequencing methods. In microarray-based methods, DNA material of a biological sample is firstly treated with bisulfite, which converts the unmethylated cytosines to uracils, and then amplified, which converts uracils to thymines. Therefore, methylation levels can be observed from SNP type probes of variant alleles C/T that mark the methylation status methylated/unmethylated at their genomic locations. As the collection of methylomic data is reduced to the sequencing of treated DNA, where nucleotide replacements of unmethylated cytocines by thymines are induced, high-throughput sequencing can also be applied.

Methylomic data is highly sensitive to cell type and is highly variable between individuals. Therefore, large association studies are required to observe

reliable effects of methylation on phenotypes. As such, methylomic microarrays are currently favored due to their scalability to multicentric studies.

1.2.5 Exposomic data

Environmental factors are important contributors to the etiology of most complex diseases. Therefore, individual differences in multifactorial traits will not be completely explained as long as environmental conditions are not taken into account [35]. In 2005, Christopher Wild defined the *exposome* as every exposure to which an individual is subjected from conception to death [170]. While the definition parallels that of molecular *omic* data, as the total content of similar entities of a given biological domain, the totality of an individual's exposure is clearly immesurable. Efforts to bound the exposome to an operational definition need to consider both the nature of exposures and their changes over time [170]. As for the nature of exposures, one can consider those at the internal environment, specific external environment, and general external environment [171]. Whereas characterizing the time scale of the exposome is more challenging as the exposure dynamics and kinetics can change in orders of magnitude depending on the exposure. While a state of the exposome is difficult to define, following the parallels from the genome or transcriptome, a high dimensional collection of exposures can be performed, each of which is under their own spatiotemporal characteristics and methods of acquisition. Unbiased assessment of the exposome is harder to achieve as targeted exposures need to be previously defined. Higher degrees of unbiased measures can be achieved with methods that intend to scan the compounds from the exogenous origin within an organism, using mass spectrometry. However, it is still early days for such developments.

1.3 Association studies

The most immediate interest in *omic* data is to underpin trait differences between individuals at lower biological domains. Many causal relations do exist in cases were specific differences at the molecular level are amplified at a population level. For instance, Mendelian mutations in the gene *F8* can lead to hemophilia A, or de-regulation of biochemical signaling of insulin can result in diabetes. However, complex multifactorial traits typically emerge from the interactions between many units at different biological levels. Tracing back the differences between individuals at lower levels is greatly challenged by the complexity within and between biological domains.

Omic data offers an unbiased high-dimensional scan of a biological domain; and therefore, consistent patterns that associate with given individual differences in the population can be searched. In this approach, the patterns

do not arise from a specific scientific hypothesis, but rather from the more general question of whether we can observe a consistent, reproducible set or arrangement of variables in a given *omic* domain that associates with subject differences. The analysis of *omic* data is therefore not based on the testing of mechanistic hypotheses. It is based on the discovery of plausible biological patterns that can guide researchers into the mechanisms. For instance, genome-wide association studies have identified numerous genomic variants associated with late-onset Alzheimer's disease [73]. Some of the variants had been associated with the disease before genomic data were available, such as the variants in *APOE*[28]. Whereas most variants are within genes, not previously associated with the disease, but offer new insights into its etiology, like the probable role of endocytosis by variants in *PICALM*[159].

1.3.1 Genome-wide association studies

Genome-wide association studies (GWASs) are based on the analysis of genomic data that try to identify SNP variants that are independently associated with differences between population samples. GWASs are any type of observational studies where specific subject differences are of interest. Specifically, the question that these studies address is if there exists *any* SNP that independently associates with subject differences. The patterns searched in this type of study are, therefore, at the univariate level. As such, massive univariate tests are performed, one for each SNP in the dataset, for which suitable inferences are drawn and tested for statistical consistency, scientific reproducibility, and biological plausibility.

A large number of GWASs have been performed in the last decade. From 2008, the GWAS catalog, sponsored by the National Human Genome Research (NHGRI) and the European Bioinformatics Institute (EMBL-EBI), has systematically collected the GWAS results of over 2,500 publications on humans [168]. These include the significant associations of SNPs for several common traits, including cardiovascular disease, cancer, type-2 diabetes and human morphology, amongst many others.

GWASs have demonstrated the ability to discover genomic variants associated with complex traits. However, they have also shown that numerous variants are needed to explain sizable amounts of common phenotypic variability [176]. In particular, for many heritable traits, there is still an amount of heritability not explained by GWASs, as the associations are typical of small in effect size [90]. In addition, the hypothesis of detecting *any* significant association amongst all the genomic variants requires strong adjustment of significance thresholds to account for random findings. The adjustment for multiple comparisons, as it is known, is in the order of the dimensionality of the genomic data, reducing statistical thresholds to orders of 10^{-8}. Therefore, studies of large sample sizes have been required to achieve enough statistical power to detect and validate findings [9]. Reasons for genomic data not accounting for the expected heritability of given traits, as measured in

twin studies, are the complexity between SNP interactions or the contribution of other genetic variants, such as translocations, CNVs or inversions. Gene-environment interactions can also contribute to explain phenotype variability and lack of validation between studies. All these types of associations are also considered in specific *omic*-wide association studies.

1.3.2 Whole transcriptome profiling

While genomic data can be regarded as structural data, inasmuch as it defines the DNA content of an individual, transcriptomic data is functional. That is, it is defined by the biological states of specific tissues of an individual, at given moments. Given that some biological states are accessible to some tissues and not others, such as the neurotransmission release in the brain, or $T3$ production in thyroid, transcriptome profiling informs on the physiological functions of the tissues. As genes are tightly correlated in biochemical pathways, transcriptomic data reflects the co-regulation of genes by the correlation between their transcripts levels. Therefore, the correlation structure of transcritomic data is in itself of biological interest. Intense research is dedicated to revealing the network structure of the transcriptomic data under different conditions, such as disease state or different tissues. Numerous biochemical pathways are known and have been carefully reconstructed from well-detailed experiments [68]. Therefore, specific transcriptomic data can be methodologically compared to this pathway knowledge and inferences can be drawn on whether such-and-such pathway was active or differentiated in the data collected in a given population sample [149].

Transcriptome-wide association studies have also been extensively performed to study transcriptional differences between individuals or transcriptional signatures of phenotypes. For instance, significant and consistent differences in estrogen receptor signaling have been identified in breast cancer subtypes [29]. Given the large variability in transcriptomic data, that arises from technical (i.e. batch) and biological differences, large meta-analyses have been required to validate findings. In addition, there is an increased effort to reduce biological variability by profiling the transcriptome of single cell types.

Integration between genomic and transcriptomic data is also highly informative. Association studies in which transcription levels can be explained by genomic variation aim to determine quantitative trait loci (eQTL). These are variants that can modulate the transcription of genes by altering their coding or regulatory sequences. It has been observed that significant SNPs in GWAS are likely to be eQTLs, offering further information into the biological mechanisms underlying the associations with phenotypes [103].

1.3.3 Epigenome-wide association studies

Epigenome-wide association studies (EWAS) use epigenomic data to determine which methylated site can explain more phenotypic variability. Given

that DNA methylation is affected by genomic variation and environmental exposures, the interest is to find regions in the genome that are responsive to the environment and can contribute either to adaptation or disease. Phenotypes with known heritable burden but also explained by differences in environmental exposures are of high interest. For instance, as methylation is important during development for tissue differentiation, there is interest in studying methylation patterns on head circumference, body-mass index and other developmental measures in children [131, 151]. In addition, phenotypes from mental diseases which require a gene-environment framework, such as schizophrenia or post-traumatic stress disorder [174, 160], have been approached by EWAS.

Similar to transcriptomic data, epigenomic data is strongly dependent on tissue and technical collection of data. As with GWAS, the aim of EWAS is to identify the methylation probes that are independently associated with phenotype differences between subjects. Because there are correlations between neighboring methylation probes, univariate analysis has been extended to multivariate analysis that comprises probes in extended genomic regions, which can alter the expression of a given gene or cluster of genes. Methylomic data is highly variable between individuals, therefore large meta-analyses are also considered to account for this variability and to validate findings.

1.3.4 Exposome-wide association studies

Exposome-wide association studies (ExWAS) turn their attention to the environmental risk factors as a source of phenotypic differences between individuals. Rapid developments in technology and declining costs have led to a massive increase in the amount of exposure data that can be collected for individuals over time. Current epidemiological studies are able to simultaneously measure hundreds of exposures using a combination of questionnaires, arrays of sensors and biochemical assays (see Table 1.1). Commonly assessed exposures include chemicals in the air, water, food, or household products, as well as information about individual behaviors, activities, and surrounding physical environments. Exposomic data is therefore highly heterogeneous, as it is the conjunction of different modalities that are derived from different experimental methods. The correlational structure of exposomic data Exposomic data is also dependent on the sub-types of data that are included in the exposure matrix. Therefore, given its complexity, association analyses are also performed mainly at the univariate level, in which the objective is to detect any exposure factor that is significantly associated with trait differences. Scalability and power analyses at the exposome level are difficult to establish given the heterogeneity of the exposures and the underlying mechanisms of action. Therefore, as in other *omic* studies replication is necessary to validate a finding.

TABLE 1.1
Most relevant research projects studying the *exposome* in human health.

Project	Web Site
The HELIX Project	http://www.projecthelix.eu/
The EXPOsOMICS project	http://www.exposomicsproject.eu/
HEALS	http://www.heals-eu.eu/
The Human Exposome Project	http://humanexposomeproject.com/

1.4 Publicly available resources

The ability to survey biological domains with high-throughput technology has been matched with the ability to share the data through the Internet. Collection and analysis of *omic* data are complex and can yield to not reproducible observations, particularly, if the effects are small, the biology complex and the between-study variability large.

A first objective of making the raw data of *omic* studies publicly available is to promote their independent reanalysis to increment reproducibility. Many meta-analyses have been made possible through the access of the data. However, it is becoming increasingly the important use of data repositories to confirm specific hypothesis, to find supporting evidence of initial findings at different biological domains or to test new analysis methods that adapt better to biological complexity. There are numerous data repositories, here we cite some of the most widely used with a particular emphasis on association studies. All data available through these repositories are susceptible to be analyzed using the methods described in this book.

1.4.1 dbGaP

The database of genotypes and phenotypes (dbGaP) is a public repository of genomic, epigenomic, somatic mutations, transcriptomic and microbiomic data, with associated phenotypes [88]. The repository is provided by the National Center for Biotechnology Information (NCBI). At the time of the publication of the book, the repository contained assay data for over 1.6 million of SNP arrays (2.3 hundred imputed), 10 thousand expression arrays and 10 thousand methylation arrays. It also contains high-throughput sequencing assays for 150 thousand whole exome sequencing, 50 thousand whole genome sequencing, and 25 thousand RNA-seq.

dbGaP offers open-access data and controlled-access data. Open data can be accessed without permission and pertains to general data about the study, including some phenotypic variables and summary results. dbGaP controls access to de-identified genotypes and phenotypes. Formal requests to use the

data are required to ensure its use for scientific purposes, to comply with the ethical standards of the studies and to warrant proper use of sensitive data.

dbGaP is, therefore, a primary source of *omic* data and its influence is only expected to grow, as specific studies will be required to contrast their results with published raw data, and new methodologies will be able to access large sets of observations to assess their viability.

1.4.2 EGA

The European Genome-phenome Archive (EGA) is a permanent archive that promotes the distribution and sharing of genetic and phenotypic data consented for specific approved uses but not fully open, public distribution. It enables collaboration and data sharing of individual patient-level genomic and phenotype data through a controlled-access system. The EGA includes data collections for human genetics research [74].

The repository contains raw data from DNA sequencing and array-based genotyping applications, e.g. gene expression experiments, transcriptomics, epigenomics, sequencing or proteomics assays. It has processed datatypes such as genotypes, structural variations or whole genome sequence. Phenotypic data is also available, all consented for research purposes.

The archive is used as the repository of large genomic studies that include the International Cancer Genome Consortium (ICGC), the International Human Epigenome Consortium (IHEC), the The International Human Microbiome Consortium (IHMC), the UK10K project for Rare Genetic Variants in Health and Disease or the Deciphering Developmental Disorders (DDD) project among others.

1.4.3 GEO

The Genome Expression Omnibus (GEO) is a data repository specialized in functional genomic studies, including transcriptomic data from microarray and RNA-seq[8]. GEO is hosted by NCBI and has a number of on-line analysis results for specific datasets. An important advantage of GEO is that its data can be directly retrieved with Bioconductor packages in `R`. It hosts more than 90,000 accession entries. Most of the entries are for expression microarrays (50,000) but fastest growth of submissions are for high-throughput sequencing data (15,000). Large meta-analyses, including numerous studies of common phenotypes, can be routinely performed, such as those for breast cancer or Alzheimer's disease. Similar to dbGaP, GEO constitutes an archive of raw data for studies to be continuously consulted to advance understanding of trait differences at lower biological domains.

1.4.4 1000 Genomes

Other public data resources correspond to large multi-centric studies that have ambitiously collected data to characterize *omic* data across conditions of great interest [23]. The 1000 Genomes study is, for instance, based on the characterization of human genomes across numerous ancestries. The idea was to create a detailed map of human genomic variability, offering a platform to further support research in genetics, medicine, bioinformatics, pharmacology, and biochemistry.

The 1000 Genomes comprises genomic data from the sequencing of 2,504 individuals from 26 different ancestries with 4 × genome coverage that allows detection of variants with more than 1% frequency. This genomic data is, therefore, a reference panel of populations to impute SNPs that have not been genotyped in specific studies and to help merge genomic data across multiple studies. High dimensional meta-analysis of GWAS can be thus performed on 85 million SNPs and over 5,000 haplotypes. A subset of 423 subjects from 4 European and one African ancestry was selected for transcriptomic data collection using RNA-seq. The data was produced by the GUEVADIS project and is also freely available. The aim of GUEVADIS was, in particular, to study the transcriptome variability, in the lymphoblastoid cell line, across human populations. Focused on European ancestry, the study has shown that significant SNPs in GWASs are likely to be eQTLs, demonstrating that the integration of transcriptomic and genomic data can reveal causal variants and biological mechanisms of diseases.

1.4.5 GTEx

The genotype tissue-expression (GTEx) project aimed to study transcriptome variability across 53 different human tissues [85]. The project collected transcriptomic data for 714 donors, 635 of which were also genotyped, which allows the study of specific changes in the relationship between genomic and transcriptomic data across tissues. Genotype data were collected with SNP microarrays covering 5 million SNPs, while transcriptomic data was obtained from RNA-seq, with 50 million aligned reads per sample. eQTL analyses have been performed and its results are available through a web-browser. Specific queries can be performed that inform on the SNPs that modulate gene expression and splicing across tissues. The integration of genomic and transcriptomic data through eQLT analysis aims to guide GWAS results into the mechanisms underlying the associations between SNPs and phenotypes.

Data is freely available, as the GTEx project intended to offer a resource to find further support of novel findings, develop new methods for integration of genomic and transcriptomic data and investigate the variability of transcription across tissues.

1.4.6 TCGA

The cancer genome atlas (TCGA) is an initiative to collect multiomic data to support cancer research [156]. It is sponsored by the national cancer institute (NCI) and the NHGRI. The project has collected data on 33 different types of tumors in 11 thousand patients. *Omic* data includes high-throughput DNA and RNA sequencing , SNP, DNA methylation and reverse-protein arrays. It has generated 2.5 petabytes of information by its closure in 2017. However, further initiatives plan to build on this initial effort.

The objective of the project was to support research aimed at assessing the extent to which multiomic variability can explain differences between individuals in cancer susceptibility, cancer types, progression and treatment. As such, the data has supported, at the time of publication of this book, more than a thousand studies, including those by independent researchers from the TCGA network. Clearly, the analysis of such massive data to account for biological complexity across different biological domains will take decades to complete.

1.4.7 Others

There are several Biobanks that also provide free access to different *omic* data. The UK Biobank (UKB) is a prospective cohort study with deep genetic, physical and health data collected on 500,000 individuals across the United Kingdom from 2006-2010 (`https://www.ukbiobank.ac.uk/`). The Estonian Biobank contains genetic information about 50,000 individuals from the Estonian population as well as data from different resources including medical records (`https://www.geenivaramu.ee/en`). The BioBank Japan project has a registry of patients diagnosed with any of 47 common diseases and genomic data of 200,000 patients.

ReCount is a specialized repository for RNA-seq data. It stores processed and summarized expression data for nearly 70,000 human RNA-seq samples `http://bowtie-bio.sourceforge.net/recount`. The data, accessible through a web-application (`https://jhubiostatistics.shinyapps.io/recount/` and the Bioconductor's package *recount* [22], is made available with the aim of reproducing the expression profiles of reported findings. For instance, the RNA-seq data from the GTEx project is stored in ReCount. There are several Bioconductor data packages including *omic* data. Among others, the package *curatedTCGA* contains different objects corresponding to TCGA tumors that integrates RNA-seq, copy number, mutation, microRNA, protein with clinical/pathological data.

A source of publicly available exposomic data is offered by the National Health and Nutrition Examination Survey (NHANES) `https://www.cdc.gov/nchs/nhanes/index.htm` [107]. NHANES is a US national survey that covers demographics, health, nutrition, and environmental chemical exposures. NHANES started surveying in 1999, repeating every two-year cycles. The data is also available through `R` by the package *RNHANES*.

1.5 Bioconductor

Collection of high dimensional data at different biological domains demands the development of new analysis methods and generalizations of old ones. There are multiple ways in which data can be stored, preprocessed and analyzed. Diversity arises from technical capabilities, experimental conditions, and scientific hypotheses to be tested. Therefore, as volume and complexity of data increases so analytical methodology does. A proliferation of "in-house" software to address the specific needs of studies greatly challenged reproducibility, absorption and further development of the methods by a wide community of users. Bioconductor is an open source software project aimed to address these issues, by orchestrating software development in the programing language `R` to analyze high-throughput biological data [40].

1.5.1 `R`

Parallel and independently from the development of *omic* data acquisition, the data analyst community made important advancements into the integration and sharing of methodologies through the extensions of `R`. `R` is a high-level programing language that initially focused on the implementations of functions for statistical analysis. From the developer's perspective, `R` offers flexible syntax for object-oriented programming, which is easily packaged into software units with clear functionality. As free software, package contents can be modified by other developers or incorporated into other packages. The great flexibility of `R` has promoted software development in diverse research fields, in particular, those that need to quickly integrate new statistical analyses and visualization methods. `R` packages are shared through various public repositories such as The Comprehensive R Archive Network (CRAN), GitHub and Bioconductor.

In the production of `R` packages, great effort is put into documentation that includes detailed information on how to use its functions, example data, and demonstrations. In addition, manuals in form of vignettes are distributed to guide users into the specific tasks supported by the packages. The code in the examples and in the vignettes must be reproducible in any platform and by any `R` user. Initiatives to increase the reproducibility of code and reporting have been also incorporated into the production of packages. Comprehensive information of software, clearly delimited to achieve concrete tasks, have greatly incremented the users base of `R` to non-developer analysts. The `R` user's community is highly active on the Internet and packages are typically found by common Internet queries on specific topics of interest. In its website, CRAN offers the list of the packages available, all of which can be installed in `R` using the command

```
> install.package("nameOfPackage")
```

once the package is installed, it is accessible in each R session by

```
> library("nameOfPackage")
```

1.5.2 *Omic* data in Bioconductor

Bioconductor offers methods to analyze a wide range of *omic* data and to access publicly available resources [63]. Important additional tools include annotation resources and visualization capabilities. Bioconductor's project integrates all these facilities under common data structures that enable easy integration of new data and methods into existing workflows. Users can find sufficient information at different levels, such as functions, packages, and workflows, which allows them to combine and develop analysis strategies with specific needs. Bioconductor's packages are continuously growing, as technologies evolve and produce new data, and developers create packages with new capabilities. However, numerous packages have been settled into standard procedures, through an intense use and feedback, from user to developer, that continuously put to test the underlying methods. In particular, Bioconductor's methods to preprocess and perform association studies of genomic, epigenomic and transcriptomic data have achieved great consensus. They include microarray and high-throughput sequencing data.

Integration between analysis methods and retrieval of data from repositories is highly coordinated. Specific packages have been developed to use R packages to query specific databases. For instance, GEO, can be queried with the package *GEOquerry*. The package has implemented the function `getGEO` to download data of a study with a given accession number from the GEO website. Data is retrieved and made available in R as a variable of class `ExpressionSet`, a recognizable data structure of Bioconductor. The data can thus be analyzed following established workflows, to reproduce the study's reported results, or to further exploit the data to test alternative hypotheses or methods. Specific projects like TCGA and GTEx have also R packages, supported by Bioconductor, to retrieve their data.

Previously, Bioconductor's installations used the commands

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("nameOfPackage")
```

Currenlty, packages in Bioconductor, like *GEOquery*, are installed with the library *BiocManager*

```
> library(BiocManager)
> install("GEOquery")
```

Once installed, the package can be loaded as an usual R package

```
> library("GEOquery")
```

GitHub `https://github.com/` is another `R` repository, mainly used to deposit development version of new packages. Packages from GitHub can be installed into `R` using the *devtools* package

```
> library(devtools)
> install_github("nameOfRepository/nameOfPackage")
```

For instance, the author's GitHub repository is `isglobal-brge`, from which multiple packages discussed in the book can be installed

```
> library(devtools)
> install_github("isglobal-brge/nameOfPackage")
```

1.6 Book's outline

The book is designed to give an introduction on how to use established tools to analyze association studies of genomic, transcriptomic, methylomic and exposomic data. We focus our discussion on publicly available datasets. The aim for this is double: to help readers who are interested in acquiring the analytical tools to start analyzing real datasets and to show those working on statistical methodologies how to access a large amount of biological data. In addition, readers who are interested in learning how to exploit available data can start proving specific hypotheses, or find further support for specific results. We, therefore, start in Chapter 2 with case studies, whose main objective is to illustrate how to access particular data repositories and how to obtain the main results that can be expected from a standard analysis. Chapter 3 describes how to deal with *omic* data in Bioconductor. Chapters 4 to 9 are dedicated to explaining in detail the preprocessing and analysis methods, functions and visualization tools of genomic, transcriptomic, epigenomic and exposomic association studies. Chapter 10 gives a first approach to the integration of *omic* data with biochemical pathways, through enrichment analysis methods. Chapter11 describes integration between different *omic* data-sets including how to gather results into functional, disease and pathway annotations and how to perform multi-*omic* data analysis using advanced multivariate methods.

All data in the book is freely available. Most of it can be directly downloaded from the public repositories however some data has been compiled in an R package to explain specific analyses. Those data are available at `https://www.github.com/isglobal-brge` and can be installed with

```
> library(devtools)
> install_github("isglobal-brge/brgedata")
```

The data is loaded in `R` with the command

```
> library(brgedata)
```

The *brgedata* package contains several files in specific formats (binary or text files) that will be used throughout the book. Data stored in file format are found in the folder `extdata` that is created when installing the package and can be accessed by

```
> path <- system.file("extdata", package="brgedata")
```

Binary data are accessed by using the `data` function. For instance, data of a SNP association study can be retrieved into `R` by:

```
> data(asthma, package = "brgedata")
> asthma[1:5, 1:10]
  country  gender      age      bmi smoke casecontrol rs4490198 rs4849332
1 Germany   Males 42.80630 20.14797     1           0        GG        TT
2 Germany   Males 50.22861 24.69136     0           0        GG        GT
3 Germany   Males 46.68857 27.73230     0           0        GG        TT
4 Germany Females 47.86311 33.33187     0           0        AG        GT
5 Germany Females 48.44079 25.23634     0           1        AG        GG
  rs1367179 rs11123242
1        GC         CT
2        GC         CT
3        GC         CT
4        GG         CC
5        GG         CC
```

The package contains the following datasets:

```
> library(brgedata)
> ls("package:brgedata")
 [1] "asthma"           "breastMulti"      "breastMulti_list"
 [4] "brge_expo"        "brge_gexp"        "brge_methy"
 [7] "brge_prot"        "genesAD"          "gwascatalog"
[10] "lusc"
```

2

Case examples

CONTENTS

2.1 Chapter overview

In this chapter, we will show, with five case examples, how to retrieve data from five public repositories and some basic analysis that can be performed on the data. We introduce the functions and packages in `R`/Bioconductor that can be used to perform specific queries, retrieval, and analysis, all within a single `R` session. Further chapters will treat in detail the packages and the functions used to produce the results.

2.2 Reproducibility: The case for public data repositories

Accessibility to primary data has been strongly motivated by the research community to encourage reproducibility of results. Studies based on *omic* data collection are particularly sensitive to reproducibility issues due to the variety of methods and strategies of analyses, and the numerous small effects and complex interactions that may underlie a particular pattern in the data.

For a given dataset and a given analysis strategy, it is at least expected that the results obtained by two different analysts will be the same. This level of reproducibility is analytical and can be tested with independent analyses of one reference study. This is easily achieved when primary research data

is freely available for reanalysis and the methods together with their implementations are clearly explained and documented [147]. The second level of reproducibility refers to the validity of a scientific observation. In this case, we expect that under one analysis strategy the pattern observed in one study is reproduced in another independent study. Having the data of independent studies freely available clearly motivates validation of results.

In addition to reproducibility, the access to primary data has motivated the testing of novel methodologies. In this case, two different methodologies can be tested on the same dataset and study their differences and commonalities.

The use of freely available data has greatly contributed to advance reproducible research in studies with high dimensional data. As such, the initiatives of sharing data together with strengthening public repositories are only going to increase and become a common practice in future research programs.

2.3 Case 1: dbGaP

dbGaP is a data repository of primary research on genome-wide association studies. The detailed description of the studies available can be queried in `https://www.ncbi.nlm.nih.gov/gap`. Studies can be queried by keywords, i.e. "Alzheimer" or directly by its accession number. Our first case example is based on the summarized data of a study on late-onset Alzheimer's disease (LOAD). The NIA-LOAD study was carried out by the National Institute of Aging (NIA) on families with at least two affected siblings and unrelated controls [77]. The database contains 5,273 individuals with genomic data (SNP microarray), and 5,220 individuals with phenotypic information. The dbGap's accession number is phs000168.v2.p2, a full description of the study can be found in the dbGap's website, corresponding to the accession number.

Data is available under controlled-access. Authorization for use of genotype data needs to be granted by the NIH Data Access Committee (DAC), who evaluate the purpose of use and handling of data by researchers and institutions. Research and dissemination of results are encouraged with proper acknowledging of the study. Interested readers should apply for data access. Note that this is the only example where data access is controlled, all other data in the book is unrestricted. Here, we illustrate how to display *summarized* results of reported GWAS of the LOAD-NIA study. The phenotypic variables are distributed in the dataset `LOAD610K_Subject_Phenotypes`. In particular, they report the first four principal components (PC) of SNP array data, comprising 599,011 SNPs and 3,007 subjects. A PC analysis of genomic data is used to determine the components that capture most genetic variability between the subjects. Therefore, individuals represented in the first PC components will be clustered in groups according to similar genetic background. Clearly, ancestry is the strongest predictor of common genetic background

and therefore PC analysis is used to infer ancestral similarities and differences between individuals, based on the genomic data. Therefore, one can expect that there is a strong correlation between the first PC of the genomic data and the self-reported ancestry. A scatter plot between the first two PCs of the LOAD-NIA genomic data (encoded in variables `AllEthnicity_PC1` and `AllEthnicity_PC2`), colored by the self-reported ancestry (encoded in `Race`) clearly illustrates this point.

Data is loaded in `R` as it is distributed in dbGap, in the appropriate phenotype subdirectory of the complete `LOAD610K_Subject_Phenotypes` dataset.

```
> data <- read.delim(
+ file = "phs000168.v2.pht000707.v2.p2.c1.LOAD610K_Subject_Phenotypes.GRU.txt",
+ comment.char = "#")
>
> names(data)[1:18]
 [1] "dbGaP_Subject_ID"   "SUBJ_NO"            "SEX"
 [4] "Dx_Level"           "Case_Control"       "RecruitedAsControl"
 [7] "ConType"            "BirthYr"            "AgeAtLastEval"
[10] "VitalSt"            "AgeDeath"           "Autopsy"
[13] "Race"               "Hispanic"           "AllEthnicity_PC1"
[16] "AllEthnicity_PC2"   "AllEthnicity_PC3"   "AllEthnicity_PC4"
```

We obtain the self-reported `race` variable as described in the variable report documentation

```
> race <- as.factor(data$Race)
> table(race)
race
   1    2    3    4   50   99
2692  123    3    3   12    1
```

We can then plot the first two PCs in the dataset and color them according to `race`.

```
> mycols <- c("gray90", "black", "gray70", "gray50", "gray20", "white")
> cols <- as.character(factor(race, labels=mycols))
> plot(data$AllEthnicity_PC1, data$AllEthnicity_PC2,
+      type="n", main="PCA LOAD-NIA",
+      xlab="PC1", ylab="PC2")
> points(data$AllEthnicity_PC1, data$AllEthnicity_PC2,
+        col = cols, pch=0:5)
>
> legend("bottomright", c("white", "black", "american indian",
+                         "asian", "other", "missing"),
+        col=mycols, pch=0:5)
```

PC components are strong predictors of ancestry and relevant measures of population stratification. Therefore, they can be used to identify individuals with strong genomic differences from a given population sample. In addition, they are important covariates to account for in association studies of genetic variants, as we will discuss in Chapter 4.

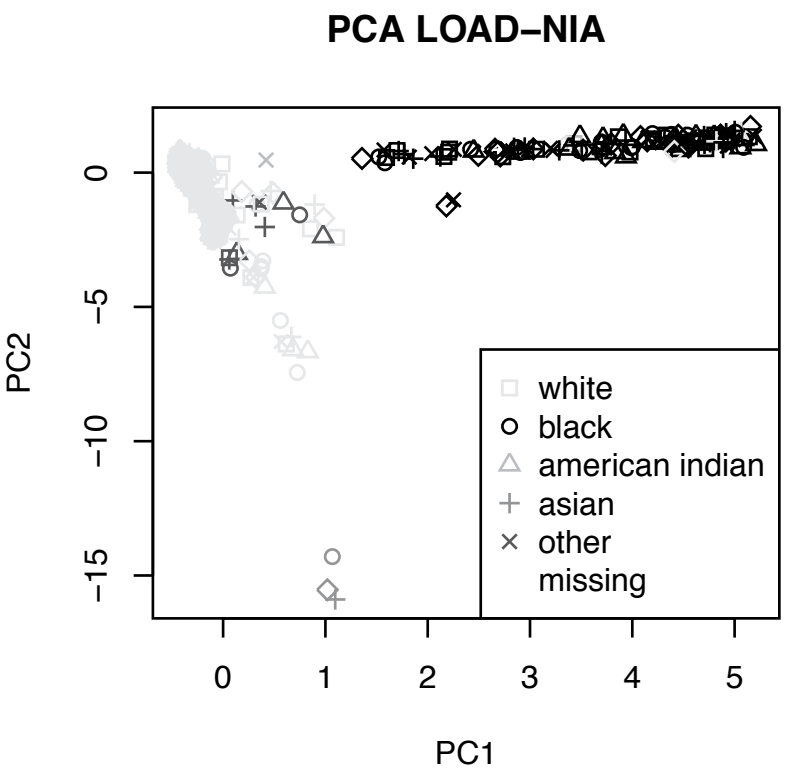


FIGURE 2.1
Genome-wide PCA of LOAD-NIA dataset.