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.