CAMBRIDGE

Cambridge University Press
978-1-107-06911-4 - Integrating Omics Data
Edited by George Tseng, Debashis Ghosh and Xianghong Jasmine Zhou
Excerpt
More information

Introduction

In the past two decades, high-throughput experimental techniques such as mass spectrometry, microarrays, and next-generation sequencing have revolutionized biomedical research with abundant genome-scale data. The fruits of this research are paving the way toward precision medicine. “Ultra-big” data sets are routinely generated now that the cost of these experiments has greatly decreased. As a result, more and more of these data sets are made available in the public domain. This abundance of data permits us to study biological processes and disease mechanisms in a multifaceted manner, drawing insights from DNA variations (e.g., genotyping and mutation), RNA transcription (e.g., gene or isoform expression and fusion transcripts), gene regulation by epigenetic changes (e.g., methylation, protein–DNA interaction, and miRNA expression), and protein expression/modification.

The enormous scope of high-throughput results creates many statistical and computational obstacles to storing, analyzing, integrating, and interpreting the data. Generally speaking, the research community is pursuing two kinds of integrative studies: horizontal meta-analysis (data from different cohorts, often from different labs) and vertical multi-omics analysis (multiple experiments performed on the same cohort). Either of these may also integrate results from the growing pathway and pharmacogenetics databases. The vast range of available data and new biomedical questions that can be answered calls for research teams with multidisciplinary quantitative expertise, including in computer science, statistics, applied math, and machine learning. This edited book collects state-of-the-art computational and statistical methods recently developed in the booming field of omics data integration. Its purpose is to showcase a wide range of cutting-edge methods and tools for our readers, in hopes of inspiring new biological and methodological research techniques to advance the field.

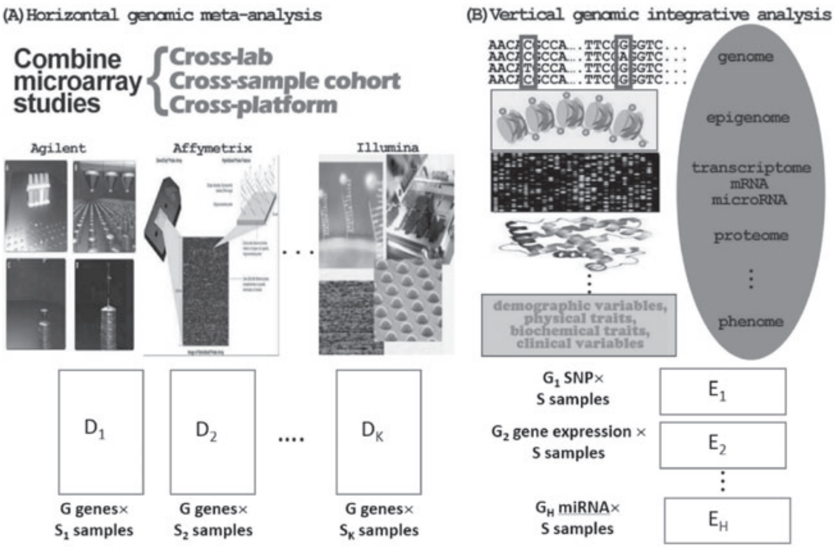

Figure I.1 (A) Horizontal omics meta-analysis. (B) Vertical multi-omics integrative analysis.

The microarray boom of the late 1990s introduced a now common convention for raw omics data: samples are arranged on the columns of the matrix, while gene features are on the rows (the main reason being that Microsoft Excel could only manage 256 columns at the time). This is in contrast to the traditional statistical convention to place samples on the rows, but in this book we keep the popular bioinformatic convention. Therefore, when multiple omics data sets from different labs are combined, the studies are integrated horizontally (Figure I.1A). In this context, many of the data integration problems now being published are analogous to traditional meta-analysis. This is why we name cross-cohort data integration "horizontal omics meta-analysis" in the preceding paragraphs and in Chapters 1–5. Alternatively, when multiple types of omics experiments are performed on the same cohort, the data sets are vertically aligned (Figure I.1B). The next set of chapters describes various types of "vertical multi-omics integrative analysis." Chapters 6–11 cover methods applicable to any type of omics data (e.g., clustering or dimension reduction methods not specific to the biological property or structure of the omics data). Chapters 12–19 cover methods that are specific to certain omics data types.

In the following we give an overview of the book's contents, based on the different biological purposes and quantitative techniques described herein.

Dimension reduction. Multi-omics data sets have naturally drawn attention to many dimension reduction methods. Chapter 2 introduces a variant of principal component analysis (MetaPCA), whereas Chapter 11 describes a method called *joint and individual variation explained* (JIVE). The first is designed for horizontal analysis, and the second is for vertical analysis. Chapter 6 proposes variations of the partial least squares (PLS) and nonnegative matrix factorization (NMF) methods, named *sparse multi-block partial least squares* (sMBPLS) regression and joint NMF. These methods reduce dimensionality and identify coherent modules in vertical multi-omics cancer data.

In addition, many published methods that have applied latent variable models and/or matrix factorization can also be considered dimension reduction techniques. The iCluster method in Chapter 7 and the iFad and iPad methods in Chapter 19 are examples of these.

Unsupervised analysis. An incrasingly popular type of analysis in omics data is to identify novel disease subtypes of clinical importance in a complex disease via unsupervised machine learning (also known as cluster analysis). Chapter 2 introduces the MetaSparseKmeans method for horizontal meta-clustering analysis. Chapter 7 proposes an iCluster method developed for multi-omics analysis. Chapter 11 develops a Bayesian consensus clustering (BCC) method that tracks both consensus and source-specific clustering in multi-omics data.

Integration with biological pathway information. Integration of omics data with public pathway databases sheds light on the key functional pathways associated with an underlying disease mechanism or other experimental perturbations. The methods described in Chapters 2, 15, and 19 include such pathway-based analyses. Chapter 2 uses the MetaPath algorithm to combine multiple transcriptomic studies for pathway analysis. Chapter 15 surveys different approaches to identifying significantly mutated pathways in cancer patients. Chapter 19 integrates transcriptome profiles, drug response profiles, and a pathway database to form drug-pathway associations.

Meta-analysis methods and the homogeneity/heterogeneity issue. Many horizontal omics meta-analysis problems have settings similar to traditional meta-analysis, but the new data structures and biological questions are inspiring novel developments. Chapters 1–5 cover this area. Chapter 1 describes new methods and practical guidelines for meta-analysis of genome-wide association studies (GWAS). An increasingly relevant problem in data integration is managing homogeneity and heterogeneity in the analysis (see Chapters 2, 5, 8, and 11). The adaptively weighted meta-analysis approach in Chapter 2 directly searches a feature-dependent subset of studies with concordant signals for horizontal meta-analysis. Chapter 5 applies the concept of motifs to handle exponentially increasing homogeneity/heterogeneity patterns in ChIP-chip and ChIP-seq meta-analysis. Chapter 8 provides homogeneity and heterogeneity regularization models for outcome association analysis. Chapter 11 develops a joint principal component analysis (PCA) framework that can separate the homogeneous and heterogeneous signals during dimension reduction and a Bayesian consensus clustering (BCC) method that tracks consensus and source-specific information in clustering formation.

Graphical model and network analysis. Graphical and network methods are powerful tools to model and elucidate associations, message flows, and gene regulation in biological systems. Many methods in this book make use of graphical and network models (e.g., Chapters 2, 3, 4, 9, 12, 13, 14, and 15). Chapter 2 describes the MetaDiffNetwork method for identifying recurrent network modules that are highly connected in one condition but altered in another condition across multiple transcriptomic studies. Chapter 3 reviews several novel graph mining algorithms to identify frequent and heavy subgraphs across a series of large weighted graphs and discover frequent coupled subgraphs in a series of two-layered graphs. Chapter 4 describes computational methods for modeling genetic information flow in networks and studying differential connectivity in co-expression networks. Chapter 9 proposes a graphical model using a Bayesian approach to study regulatory relationships of multi-omics data. Chapter 12 extends the expression quantitative trait loci (eQTL) analysis to directed graphical models. Chapter 13 discusses integrative methods for inferring miRNA regulatory networks. Chapter 14 presents a probabilistic graphical model that integrates diverse omics data to infer cancer patient-specific pathway activities, as well as a mathematical model to isolate important subnetworks. Chapter 15 contains network-based approaches to identify recurrent combinations of mutated genes in cancer genomes.

Bayesian modeling and inference. Hierarchical Bayesian models provide a natural solution to interpreting many multi-omics data structures and answering biological questions. The potential downsides of Bayesian analysis include arguable prior distribution specifications and the high computing cost of Monte Carlo simulations. Chapters 5, 9, 10, 11, and 19 contain examples of Bayesian approaches to data integration. Chapter 6 applies an EM algorithm to derive the posterior probabilities in the outcome association analysis. Chapter 9 adopts a Bayesian inference for a Markov random field model that can investigate multi-omics regulatory relationships. Chapter 10 proposes a multilayer Bayesian hierarchical model to integrate miRNA, copy number variation, methylation, mRNA expression, and clinical phenotype. Chapter 11 develops a Bayesian consensus clustering model using conjugate priors and Gibbs sampling for inference. Chapter 19 applies an advanced collapsed Gibbs sampling technique to speed up the posterior probability approximation.

Regularization and penalization methods. The techniques of feature regularization and penalization have gradually gained popularity in genomic research. This trend arises naturally because the high dimensionality of the models often diminishes their stability and obscures interpretation. Regularization methods "shrink" the effect sizes of the majority of features so that they provide zero contribution to the model, thereby achieving a model with limited dimensionality and good theoretical properties. The penalization and regularization methods are seen in Chapters 2, 6, 7, 8, 10, and 19. Chapter 2 applies regularization in the meta-analysis framework of sparse K-means when combining multiple transcriptomic studies to identify disease subtypes (the MetaSparseKmeans method). Chapter 6 applies network regularization in the joint NMF method. The iCluster method in Chapter 7 uses regularization in the latent variable model before it performs clustering analysis. Chapter 8 performs penalization and feature selection in the high-dimensional association models. The Bayesian hierarchical model in Chapter 10 incorporates ideas from the statistical regularization literature for combining multiple levels of omics data. Chapter 19 adopts regularization in the drug-pathway association analysis.

Data integration to study gene regulation. Gene regulation is a complex process, subject to multilevel controls. Much of the data integration effort has been devoted to deciphering the mechanisms and implications of gene regulation. Chapters 3, 4, 5, 9, 12, 13, 16, 17, and 18 contain computational and statistical methods to study various aspects of gene regulation. Chapter 3 presents methods that integrate many microarray or RNA-seq data sets to reconstruct transcriptional regulatory networks and splicing regulatory networks and explore how transcription and splicing simultaneously take place. Chapter 4 reviews several computational approaches that model the flow of genetic information to gene expression in biological networks. Chapter 5 develops a novel statistical framework for integrative analyses of ChIP-X data to improve peak calling and study allele-specific binding. Chapter 9 proposes a Bayesian graphical model to study regulatory relationships involving copy number variation, DNA methylation, and mRNA expression. Chapter 12 reviews methods to estimate directed graphical models with eQTL data. Chapter 13 focuses on predicting microRNA targets and microRNA regulatory networks. Chapter 16 discusses a model-based approach to quantitatively dissect the contributions

Cambridge University Press
978-1-107-06911-4 - Integrating Omics Data
Edited by George Tseng, Debashis Ghosh and Xianghong Jasmine Zhou
Excerpt
More information

of RNA-level and protein-level regulation in the variation in gene expression. Chapter 17 discusses statistical models to quantify the relationship between TF binding, histone modification, and gene expression. Chapter 18 presents some integrative analysis approaches to identify lncRNAs that are specific to cancer subtypes and predict those that are potential drivers of cancer progression.

Cambridge University Press
978-1-107-06911-4 - Integrating Omics Data
Edited by George Tseng, Debashis Ghosh and Xianghong Jasmine Zhou
Excerpt
More information

PART A

HORIZONTAL META-ANALYSIS

1

Meta-Analysis of Genome-Wide Association Studies: A Practical Guide

WEI CHEN

Abstract

Meta-analysis is an effective approach to combining summary statistics across multiple studies. This approach has been widely used in recent genome-wide association studies (GWAS) and next-generation sequencing (NGS) studies. As a result, numerous disease-susceptibility loci, which cannot be found in a singe GWAS, have been identified through the meta-analysis of multiple studies. In this chapter, we give an overview how meta-analysis techniques can be used in consortium projects and provide guidance for future studies. Sections 1.1.1 and 1.1.2 cover background information on GWAS and imputation techniques, which play a key role in the meta-analysis of multiple studies. Section 1.2.1 discusses the methods of meta-analysis for single variant tests and provides a basic workflow of meta-analysis in a typical consortium project. Section 1.2.2 presents an application of Section 1.2.1 from a meta-analysis of age-related macular degeneration (AMD). Next, Section 1.2.3 discusses a method for meta-analysis for a gene-level test. Section 1.2.4 presents an application of Section 1.2.3 from a meta-analysis of plasma lipid levels. Section 1.2.5 provides a discussion of popular software for meta-analysis of genetic studies. Finally, Section 1.3 closes the chapter and discusses future directions.

1.1 Introduction

1.1.1 Meta-Analysis of Genome-Wide Association Studies

In the past decade, new technologies have enabled researchers to examine genetic and genomic data on a whole-genome scale. A genome-wide association study (GWAS) is known as a popular design for assessing thousands to millions of common and rare genetic variants associated with a disease or a trait. Thousands of disease-susceptible variants have been discovered through the GWAS of hundreds or thousands of individuals [1, 2]. To summarize the findings of these studies, the National Human Genome Research Institute has

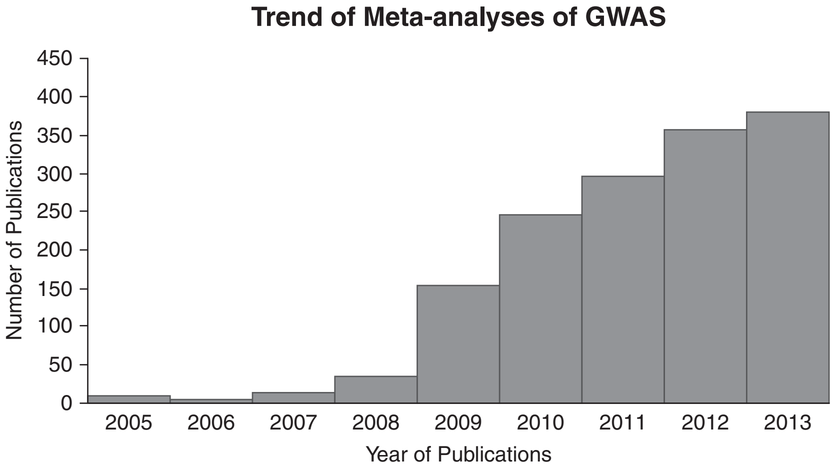

Figure 1.1 Number of publications by year from 2005 to 2013.

organized a catalog of published genome-wide association studies with frequent updates (http://www.genome.gov/gwastudies/). However, single-center GWAS typically has a limited number of samples, thus the power to detect those variants is small, especially for variants with small to modest effect sizes, as observed in many complex diseases. Although it is ideal to combine genetic data for as many individuals as possible, local institutional review board policy makes sharing of individual-level data from each study site difficult or impossible. In such situations, meta-analysis becomes a popular and powerful approach for combining summary statistics from multiple GWAS by increasing the sample size without sharing individual-level data. Multiple consortia have been founded to exchange and merge genetic data sets from multiple sites, with the central goal of identifying more disease-susceptibility loci [3]. Figure 1.1 illustrates the rapidly increase in the number of publications in PubMed using search terms “meta analysis” and “GWAS” from 2005 to 2013. If we focus on one disease or trait, we see a greatly increased number of participating studies and total sample sizes.

1.1.2 Imputation

One technical difficulty in combining different studies comprises the various genotyping platforms, which differ in density, position, and genotyping accuracy. Consequently, complete summary statistics of only a small set of