# ABSTRACT

## A FRAMEWORK FOR BIOLOGICAL DATA INTEGRATION AND FEATURE SELECTION IN LARGE DATA SETS

By

Agustin Gonzalez-Reymundez

The increasing volume of high-dimensional biological data (*omics*) has intensified the discovery of thousands of biomarkers across the different fundamental components of the cell (e.g., genome, transcriptome, proteome, epigenome) and allowed the characterization of complex phenotypes (e.g., metabolome, imaginome, phenome). However, the ability to integrate omics into informative results is constantly challenged by a seemingly ever-increasing volume of data. Furthermore, huge data sizes impose a tradeoff between how complex an omic integration algorithm can be and how much data it can handle (e.g., how fast can the algorithm be scaled to integrate large data sizes). In this dissertation, we explore statistical frameworks to face the challenges of modern omic data, including the integration of high-dimensional data of large sample sizes. We have developed a novel framework of competitive analytical performance compared with existing methods but suitable for omic data reaching biobank scales (i.e., hundreds of thousands of samples and variables). We implemented this method as an R package and showed its application on two traits of a complex molecular basis: cancer and regulation of energy intake and expenditure. In chapter one, we review the technologies and methods used to generate and integrate omic data. Chapter two describes our novel method and software of omic integration, shows examples in synthetic data, and evaluates its computational and analytical performance. Chapter three presents an application of our method to reveal a novel pan-cancer classification of tumors beyond

the tissue of origin, regulated by distinct sets of molecular signatures. In chapter four, we present an application of our method to integrate phenomics data and identify patterns of energy balance regulated by genomic variation. Finally, in chapter five, we offer general conclusions to the entire thesis.