Chapter 13
Perspectives on Data Integration in Human Complex Disease Analysis

Kristel Van Steen
University of Liége, Belgium & University of Liege, Belgium

Nuria Malats
Spanish National Cancer Research Centre (CNIO), Spain

ABSTRACT

The identification of causal or predictive variants/genes/mechanisms for disease-associated traits is characterized by "complex" networks of molecular phenotypes. Present technology and computer power allow building and processing large collections of these data types. However, the super-rapid data generation is counterweighted by a slow-pace for data integration methods development. Most currently available integrative analytic tools pertain to pairing omics data and focus on between-data source relationships, making strong assumptions about within-data source architectures. A limited number of initiatives exist aiming to find the most optimal ways to analyze multiple, possibly related, omics databases, and fully acknowledge the specific characteristics of each data type. A thorough understanding of the underlying assumptions of integrative methods is needed to draw sound conclusions afterwards. In this chapter, the authors discuss how the field of "integromics" has evolved and give pointers towards essential research developments in this context.

INTRODUCTION

DNA and RNA microarray technologies have made it possible to relate genome structure with gene expression patterns and physiological cell states. This paved the way towards a better understanding of tumor development, diseases progression, and drug response (Trevino, Falciani, & Barrera-Saldana, 2007). Since their appearance, these technologies have been used to detect single nucleotide polymorphisms (SNPs) and other structural variations in the genome, such as copy number variations (CNVs) (Feuk, Carson, & Scherer, 2006; Macdonald, Ziman, Yuen, Feuk, & Scherer, 2014; Pang et al., 2010; Pang, Migita, Macdonald, Feuk, & Scherer, 2013), as

DOI: 10.4018/978-1-4666-6611-5.ch013

well as to examine changes involving all aspects of epigenetic interactions (Colyer, Armstrong, & Mills, 2012). Next Generation Sequencing (NGS) can also be used to identify novel mutations. In addition, NGS allows the identification of protein binding to chromatin, RNA quantification, and the investigation of spatial interactions, amongst others. Compared to micro-array experiments, sequencing-based experiments are more widely applicable, since they exhibit a potentially richer information content, but at the expense of higher analytical costs and the need for more sophisticated analytic tools and well-equipped IT-infrastructures to deal with the vast amounts of data they generate (Nekrutenko & Taylor, 2012).

Genome-wide association studies (GWAS) typically assay hundreds of thousands of SNPs in thousands of individuals (Johnson & O'Donnell, 2009). Such studies have reproducibly identified numerous SNP-trait associations, as are catalogued in the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies) (Hindorff et al., 2009). The catalog includes over 1,500 curated publications of over 10,000 SNPs. With the bloom of analytic tools for gene-gene interaction analysis using SNPs (Van Steen, 2012), gene interaction studies are gradually being incorporated in the catalog as well (Welter et al., 2014). However, apart from gene-gene interactions, several other factors exist that makes GWAS less efficient, including compound or multiple phenotypes, genomic imprinting, gene-environment interactions. The latter type of interactions can be taken very broadly, realizing that intermediate phenotypes, such as gene or protein expression, DNA methylation or histone modification, also respond to variations in DNA, cascading into changes for the trait of interest. Clearly, independently carried out omics data analyses are unlikely to be sufficient to obtain a full comprehension of all the underlying principles that govern the functions of biological systems (Joyce & Palsson, 2006).

TAXONOMY

What "Integration" Is and Is Not

Data integration may mean different things in different contexts. In order to share scientific data and to analyze them as such, the available data sources need to be integrated via a uniform interface that accommodates different data types and/or data accessing from different remote locations. In this context, the term *data fusion* pops up. It refers to fusing records on the same entity into a single file, and involves putting measures in place to detect and remove erroneous or conflicting data (Wang et al., 2014). Several definitions of data fusion or data merging exist. Some of these definitions use "data integration" in their definition. However, although some data integration efforts will rely on data fusion processes, data fusion and data integration are not equivalent. In fact, new fusion techniques may appear as more complicated omics data integration tasks need to be accomplished. In line with our own viewpoints, and starting from mathematical formalisms, Oxley and Thorsen (Oxley & Thorsen, 2004) concluded that fusion can be defined as the process of optimally mapping several objects into a single object. In contrast, *integration* is the process of connecting systems (which may have fusion in them) into a larger system (Oxley & Thorsen, 2004).

A single data type omics analysis can be regarded as a comprehensive assessment of biochemical processes or interactions between molecules that belong to one specific layer of a cellular system. Examples of such layers are genomics, transcriptomic, proteomics and metabolomics. The data are characterized by their high-throughput. The analyses are often data-driven, but depending on the focus of the study (for instance, studying interactions between genetic variants or molecular interactions in cells) and/or the availability of sufficient IT infrastructure, these analyses may need to be hypothesis-driven rather than hypothesis-free. Obviously, a single omics study only provides lim-