1
Integrative Analysis of Omics Data

Tobias Österlund, Marija Cvijovic, and Erik Kristiansson

Summary

Data generation and analysis are essential parts of systems biology. Today, large amounts of omics data can be generated fast and cost-efficiently thanks to the development of modern high-throughput measurement techniques. Their interpretation is, however, challenging because of the high dimensionality and the often substantial levels of noise. Integrative analysis provides a framework for analysis of the omics data from a biological perspective, starting from the raw data, via preprocessing and statistical analysis, to the interpretation of the results. By integrating the data into structures created from biological information available in resources, databases, or genome-scale models, the focus moves from the individual transcripts or proteins to the entire pathways and other relevant biochemical functions present in the cell. The result provides a context-based interpretation of the omics data, which can be used to form a holistic and unbiased view of biological systems at a molecular level. The concept of integrative analysis can be used for many forms of omics data, including genome sequencing, transcriptomics, and proteomics, and can be applied to a wide range of fields within the life sciences.

1.1
Introduction

Systems biology is an interdisciplinary approach to biology and medicine that employs both experimentation and mathematical modeling to achieve a better understanding of biological systems by describing their shape, state, behavior, and evolutionary history. An important aim of systems biology is to deliver predictive and informative models that highlight the fundamental and presumably conserved relationships of biomolecular systems and thereby provide an improved insight into the many cellular processes [1]. Systems biology research methodology is a cyclical process fueled by quantitative experiments in combination with mathematical modeling (Figure 1.1) [2, 3]. In its most basic form, the cycle starts with the formulation of a set of hypotheses, which is followed by knowledge generation

Systems Biology, First Edition. Edited by Jens Nielsen and Stefan Hohmann.

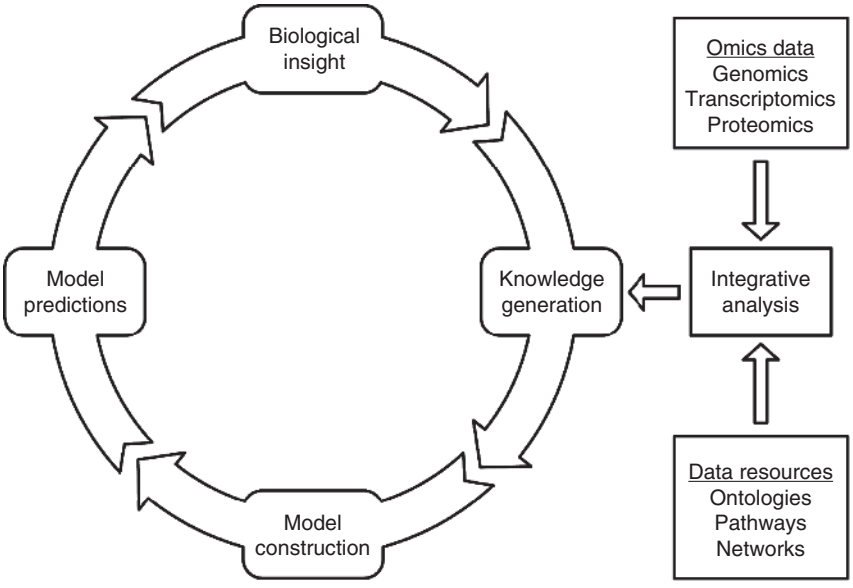

Figure 1.1 Systems biology research methodology. In the systems biology cycle, novel hypotheses are first formulated, which is followed by knowledge generation, model construction, and model predictions, which, in turn, leads to new biological insights. The development of high-throughput techniques have enabled rapid and cost-efficient generation of omics data from, for example, genome sequencing, transcriptomics, and proteomics. Integrative analysis provides a framework where omics data is systematically analyzed in a biological context, by data integration into known biological networks or other data resources, which enables improved interpretation and easier integration into quantitative models.

and model construction where an abstract description of the biological system (a model) is formulated and its parameters are estimated from data taken from the literature. The final step is defined by model predictions, where the constructed model is used to address the original hypotheses by providing a quantitative analysis of the system, which, in turn, generates new biological insight.

The development of high-throughput measurement techniques in the recent years has resulted in an unprecedented ability to rapidly and cost efficiently generate molecular data. Bioassays are today established for large-scale characterization of genes and their expression at the different layers defined by the central dogma: the genome, the transcriptome, and the proteome. The resulting data, which in this chapter will be referred to as *omics data*, is however complex because of its high dimensionality and is therefore hard to interpret and directly integrate into quantitative models. The concept of *integrative analysis* is a framework to systematically analyze the different components of omics data in relation to their corresponding biological functions and properties. The resulting biological interpretation can be used to form a holistic and unbiased view of biological systems at a molecular level. Thanks to the comprehensiveness of the

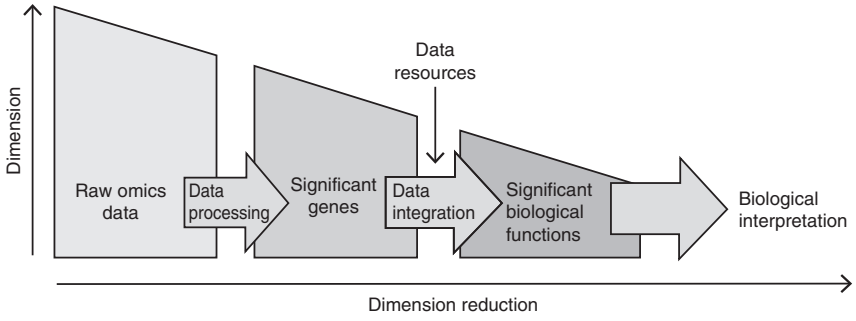

Figure 1.2 Description of the concept of integrative analysis as a tool for reduction of the dimension of omics data. Integrative analysis starts with raw omics data, which is typically affected by high levels of noise and errors. Computational and statistical approaches are first used to process the data to produce a ranked list of genes that are found to be of significant importance in the experiment. The gene list is used as input to the data integration, where known biological information is used as a basis for the interpretation of the data. During integrative analysis, the dimension of the data is significantly reduced, from potentially millions of data points to a limited number of significant biological functions and pathways, which considerably facilitates the interpretation.

omics data, all components (i.e., genes, transcripts, or proteins) can be measured simultaneously, which opens up opportunities for testing of existing hypotheses as well as generation of completely new hypotheses of the studied biological system.

The process of integrative analysis can be divided into two main steps: data processing and data integration (Figure 1.2). Integrative analysis starts from raw omics data and ends with the biological interpretation, and during this process the dimensionality of the data is reduced. The first step, the data processing, takes the high-dimensional omics data, and by applying computational and statistical tools, removes noise and errors while identifying genes and other components that contain information significant for the experiment. The next step, the data integration, uses the list of identified genes to pinpoint relevant functions and pathways by integrating the data on top of a "scaffold" built using established biological information collected from various resources and databases. The result, which is based on the combined analysis of the genes with similar functional properties, has a substantially reduced dimension, which considerably facilitates its interpretation.

Many studies in the life sciences aim to understand biological systems, often in relation to a perturbation caused by, for example, disease, genetic variability, changes in environmental parameters, or other factors introduced through laboratory experiments. A commonly used measurement technique is transcriptomics, where the transcriptional response is analyzed and the genes that are differentially expressed between investigated conditions are identified. In this setting, the data integration shifts the focus from what *genes* are differentially expressed to providing a *biological context* where activated and repressed pathways, functions, or subnetworks can be identified. This provides a more relevant view of the data, which paves the way toward more sound and detailed biological conclusions.

In this chapter, we provide a broad overview of integrative analysis of omics data. We will describe the general concept of integrative analysis and provide an outline of the many associated computational steps. It should, however, be pointed out that this topic has been extensively researched during the recent years and – due to the scope of the topic at hand – we will not be able to cover all aspects and details in a single chapter. We have therefore provided a comprehensive set of references throughout the text, which are the recommended starting points for further reading. Also, our main focus throughput this chapter will be on data generated by techniques from genomics, transcriptomics, and proteomics. This means that other types of data, which are commonly encountered in systems biology, such as metabolomics and lipidomics, will receive little attention, and here we instead refer the reader to the recent reviews by Robinson *et al*. [4] and Kim *et al*. [5].

The chapter is organized as follows. Section 1.2 contains an overview of some the types of omics data that are commonly used in integrative analysis. This is followed by Section 1.3, where we focus on the data processing, starting from the quality assessment of the raw data to statistical analysis. Section 1.4 explains the concepts of data integration and describes the different approaches and data resources that can be used. We end the chapter with an outlook discussing future challenges related to the continuous growth of biological information.

1.2 Omics Data and Their Measurement Platforms

In this section three commonly used types of omics data will be described, namely genome sequencing, transcriptomics (RNA sequencing and microarrays), and mass spectrometry (MS)-based proteomics.

1.2.1 Omics Data Types

Genome sequencing is used for determining the order of the complete set of nucleotides present in an organism. The comparative analysis of the genome of a strain or a multicellular organism in relation to a reference genome is referred to as "resequencing," which enables identification of the complete genotype and its variation between individuals. This includes both small mutations, such as single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels), and larger structural variations such as genome rearrangements and copy number alterations [6]. The resulting information, containing a list of all identified genetic variants, is often subjected to integrative analysis in order to provide a biological context where the genotype can be linked to a phenotype [7]. Whole-genome and exome resequencing are important techniques for the study of human disease [8], and in, for example, cancer, the set of germline and somatic mutations are

often good predictors of the tumor phenotype, including aggressiveness, ability to metastasize, and drug resistance [9].

Transcriptomics is the large-scale analysis of gene expression at the transcript level. Modern transcriptomics is based on RNA-seq, which is the process where RNA is reversed-transcribed into complementary DNA (cDNA) and then sequenced *en masse* [10]. From the resulting data, the relative abundance of expressed mRNA and other functional noncoding RNA can be estimated. RNA-seq can also provide detailed information about alternative splicing and expression of isoforms as well as antisense transcription [11]. Analogous to transcriptomics, proteomics is the study of the gene expression but at the protein level. Large-scale proteomics data is generated by bottom-up tandem MS (shotgun proteomics), where a mixture of proteins extracted from a sample is first enzymatically digested (using, e.g., trypsin) followed by peptide separation using liquid chromatography. The peptides are then subjected to two consecutive mass spectrometry runs where the individual peptides are first separated and then fragmented to generate a set of mass spectra. The resulting data provides information about the peptide sequences and their relative abundance in the sample [12]. Proteomics can also be used to study post-translational modifications, such as phosphorylation and ubiquitination [13]. Integrative analysis of transcriptomic and proteomic data has long been popular to study and interpret differences in gene expression between tissues and individuals, as well as medical, environmental, or experimental conditions [14, 15].

1.2.2 Measurement Platforms

The recently introduced next-generation sequencing (NGS) technology has revolutionized large-scale characterization of DNA [16]. In contrast to the traditional Sanger sequencing, which is inherently a serial process, NGS is massively parallel and can characterize billions of DNA fragments simultaneously. This has enabled rapid and cost-efficient generation of vast volumes of DNA sequence data, and, consequently, genome resequencing and transcriptomics are today almost exclusively based on NGS. There are several NGS platforms available, and they all have differences in their performance and characteristics [17]. The Illumina platform uses a sequencing-by-synthesis approach where fluorescence-tagged nucleotides are consecutively incorporated to form the reverse strand of single-stranded DNA fragments. Each incorporated base is registered using a camera, which provides information about the nucleotide sequence of billions of fragments simultaneously. The Illumina sequencing technique has a high throughput, where one single run can generate more than 1 terabase of sequence data. The length of the generated reads are however relatively short (currently 100–300 bases) [18]. The IonTorrent platform also applies sequencing-by-synthesis scheme, but the incorporated bases are instead registered by semiconductor measurement of fluctuations in pH resulting from the release of hydrogen ions [17]. The IonTorrent platform provides quick sequencing runs and can generate reads

up to 400 bases but has a lower throughput that the Illumina platform. A third commonly used platform is Pacific Bioscience (PacBio), which uses a sequencing technique where fluorescence pulses of the incorporated tagged nucleotides are detected in real time [18]. PacBio can generate sequence reads up to 20 000 bases but has still a limited throughput compared to the Illumina and IonTorrent platforms [19].

Similar to those of DNA sequencing technology, the performance and throughput of MS-based proteomics have increased drastically during the last decade. This is a result of the improvements in and optimization of the many of the steps in the proteomics workflow. In particular, improved protein digestions through the use of multiple proteases, optimized chromotographic peptide separation, and novel instrumentation with higher resolving power and scan speed have significantly increased the performance – both with respect to sequencing depth and proteome coverage [20]. As a consequence, MS-based proteomics can today be used to identify >10 000 unique proteins in a single sample using low volumes of starting material and thus generate a comprehensive snapshot of the proteome [21, 22].

Microarray technology, first introduced 20 years ago, is based on fluorescence-tagged cDNA that is hybridized to unique gene-specific probes distributed over a chip. A laser scanner is used to extract information about the amount of DNA captured by each probe. Microarrays were previously popular, for example, for large-scale transcriptomics and identification of SNPs but have, compared to NGS-based techniques, lower resolution and are plagued by high technical variation and systematic error [23, 24]. Even though the microarray measurement technology has to a large extent been superseded, there is a large accumulated body of microarray data present in the public repositories that can be subjected to integrative analysis [25]. There is a vast literature regarding all steps of the processing of microarray data, and it will therefore be less extensively covered in this chapter [26, 27].

1.3 Data Processing: Quality Assessment, Quantification, Normalization, and Statistical Analysis

All forms of omics data need to be computationally processed before any biological conclusions can be reached. Data processing, which is the first main step of the integrative analysis, can be split into four parts: (i) quality assessment, (ii) quantification, (iii) normalization, and (iv) statistical inference, all of which are necessary to ensure a reliable end result (Figure 1.3). While data processing shares conceptual similarities between data types, there are also important differences related to the measurement platforms and their error patterns. In this section, we will describe the purpose of each of the four parts and provide references for suitable tools and software. The key methods for the different analysis steps have been summarized in Table 1.1.

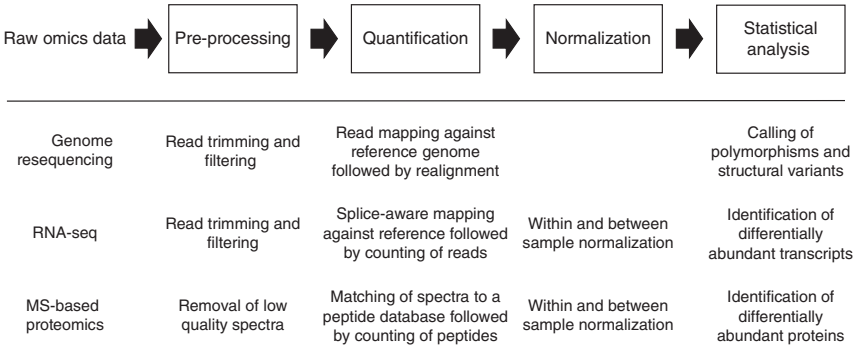

Figure 1.3 Overview of data processing, which start with raw omics data and produces a list of significant genes. Data processing is divided into four main steps: quality assessment, quantification, normalization, and statistical analysis. Omics data types have important differences related to their measurement platforms and their error patterns, and different data processing methods are therefore necessary. For each data type, the figure summarizes the most important parts of the analysis in each of the processing steps. For examples of available methods for each step, see Table 1.1.

1.3.1 Quality Assessment

DNA sequencing and tandem MS are inherently noisy, and the generated data contain errors and irregularities. If not properly removed, erroneous information can propagate through the consecutive analysis steps and into the final results. Quality assessment of high-throughput data is therefore a vital step and should always be performed. The nature of the errors is heavily dependent on the specific bioassay and its biochemical properties, and methods for quality assessment should therefore be selected based on the applied measurement platform.

In high-throughput DNA sequencing, the most common type of error is incorrect base calls introduced during the sequencing process [45]. The characteristics of the errors differ between the sequencing platforms: while the Illumina platform is almost exclusively associated with incorrect substitutions [46], the IonTorrent and PacBio platforms are dominated by insertions and deletions, often within homopolymeric regions [47]. Furthermore, the reliability of the sequencing process typically decreases along the processed DNA fragments, in some cases leading to substantially decreased quality at the end of the sequenced read. The general strategy for quality assessment of sequence data is therefore to exclude bases that are likely to be inaccurate, either by trimming the end of the generated sequence reads or by completely discarding reads from the analysis. The exclusion is based on a base-specific quality score that is provided by all sequencing platforms, which estimates the probability of a sequenced base being incorrect. Quality score thresholds can be used to tune the stringency of the quality assessment in relation to the application at hand.

Table 1.1 Examples of key methods for processing of omics data.

Method	Purpose	Type of data
Quality assurance and filtering		
FASTX toolkit	Quality control and filtering	Genomics, transcriptomics
Trim Galore!	Quality filtering and removing adapters	Genomics, transcriptomics
NGS QC Toolkit [28]	Quality control and filtering	Genomics, transcriptomics
Spectrum quality [29]	Filtering of MS spectra	Proteomics
Quantification		
BWA [30]	Mapping of reads to reference	Genomics, transcriptomics
Bowtie2 [31]	Mapping of reads to reference	Genomics, transcriptomics
TopHat [32]	Splice-aware mapping of reads to reference	Transcriptomics
Star [33]	Splice-aware mapping of reads to reference	Transcriptomics
SEQUEST [34]	Matching MS spectra to a database of full-length peptides	Proteomics
MASCOT [35]	Matching MS spectra to a database of full-length peptides	Proteomics
InsPecT [36]	Matching MS spectra to a database of peptide patterns	Proteomics
Normalization		
RPKM/FPKM [37]	Normalization by transcript length and total abundance	Transcriptomics
Upper quartile normalization [38]	Normalization of transcript abundance	Transcriptomics
Trimmed mean of M-values (TMMs) [39]	Normalization of transcript abundance	Transcriptomics
Linear regression normalization [40]	Normalization of peaks in MS spectra	Proteomics
Statistical analysis		
GATK toolkit [41]	Identification of significant genotype variants	Genomics
MuTect [42]	Identification of somatic point mutations in cancer	Genomics
edgeR [39]	Identification of differentially expressed genes	Transcriptomics
deSeq2 [43]	Identification of differentially expressed genes	Transcriptomics
QPROT [44]	Identification of differentially expressed genes at the protein level	Proteomics

RPKM/FPKM – Reads/fragments per kilobase per millions of mapped reads

Multiple algorithms have been developed for quality assessment of sequence data for the different platforms, for example, the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and NGS QC Toolkit [28]. In addition to quality filtering, several of these methods can also remove other types of inconsistencies in the sequence data, such as adaptor contamination and duplicated reads.

Tandem MS generates a large number of spectra, of which only a small proportion corresponds to identifiable peptides. The major part is, instead, dominated by spectra with lower quality, which provide no, or in the worst case ambiguous, information and may result in false positives in the downstream analysis [48]. The quality of each spectrum can therefore be assessed based on its characteristics, such as peak intensity, peak distance, and signal-to-noise ratio, either using statistical models or unsupervised machine learning algorithms [29, 49]. Spectra that are deemed to not pass a prespecified quality threshold are excluded from further analysis.

1.3.2 Quantification

The quantification step transforms the quality-assessed raw data into quantitative values describing the abundance of the genetic variants, transcripts, or proteins. In genome resequencing, this is a two-step process in which the reads are matched first to a reference genome to identify their correct position, typically using computationally efficient alignment-free mapping algorithms (e.g., BWA or bowtie2) [30, 31]. This is often followed by a more sensitive realignment step, where reads in regions with a high dissimilarity between the sequenced and reference genomes are realigned [50]. From the resulting alignment, differences between the sequenced and reference genomes can be identified and their relative abundance of variants estimated.

Quantification of RNA-seq is done through a process called binning, where the sequenced reads are first mapped to a reference that is annotated with any feature that may be of interest in the study (i.e., the reference containing "bins"). The relative abundance of each bin is then derived based on the number of matching reads [51]. The reference is often the genome from the studied organism, and the mapping needs therefore to be done using algorithms that are splice-aware and can correctly align reads that extend over exon boundaries (e.g., TopHat or STAR) [32, 33]. After mapping, the quantification can be done for genes, isoforms, or single exons based on the number of matching reads [51, 52]. If no suitable reference genome is available, a reference can be assembled *de novo* from the generated sequence data [53].

Proteomics data is quantified by matching the measured spectra against a comprehensive database with theoretical spectra calculated from known peptides. A similarity score is used to measure the similarity between the measured and the theoretical spectra and, based on the score, a best match is identified [12].

The matching can be done either for complete peptides (using methods such as SEQUEST or MASCOT) or tag-wise based on sub-peptide patterns (using, e.g., Inspect), which also enables identification of peptides that are not in the database [54]. Next, the identified peptides are matched to full-length proteins, and the abundance of each protein is calculated based on the number of matching spectra. Alternatively, proteomics data can be quantified based on labeling where the intensity of each spectrum is compared to a spiked internal standard of known quantity [12].

1.3.3 Normalization

Omics data exhibit large variations and biases, of which a substantial part is of technical nature introduced by the measurement techniques and the sensitive experimental steps necessary for sample preparation. The purpose of normalization is to remove this unwanted variability in order to make the data more uniform and comparable. This is especially important for transcriptomics and proteomics where data is often generated in a comparative setting using multiple technical and/or biological replicates.

A large and common source of bias in sequence-based transcriptomics is the varying sequencing depth between the samples. The number of counts for a particular transcript or protein thus cannot be directly compared within or between samples without first relating it to the total number of generated sequence reads. The RNA-seq counts can be transformed into RPKM values (reads per kilobase per million mapped reads) by normalizing the number of counts with the total number of fragments and the length of the transcript [37]. Studies have, however, demonstrated that the total number of reads is not robust against larger changes in the transcriptome. Highly expressed genes, such as actin or the ribosomal proteins, constitute a sizeable part of the total transcriptome, and differences in their abundance between the samples will affect the total number of fragments and thereby introduce biases. It is therefore recommended that the total number of reads is replaced with a robust alternative such as the upper quartile (UQ) of the transcript abundance distribution [38]. Another alternative is to apply the trimmed mean of M-value algorithm (TMM), which robustly estimates sample-specific scaling factors to ensure that the vast majority of the genes are not differing in expression between the samples [39]. Even though the UQ and TMM algorithms are more robust than normalizing with the total number of reads, they still explicitly assumes that only a small proportion of the transcripts are differentially expressed between the samples (e.g., <30% for TMM) and may otherwise perform suboptimally.

Considerable systematic and random biases are introduced between the MS runs, which makes normalization of proteomics data important [55, 56]. A common approach is to correct the ion intensities in the spectra, either based on specific peaks from housekeeping genes ("housekeeping peaks") or based on the data of all or the majority of the quantified peptides detected in the sample.

Evaluations have shown that methods that use regression-based correction of peak abundance [40, 55], normalization of the empirical quantiles between the replicates [44], and scaling based on the total ion intensity in relation to the total protein length all have good and robust performance for several types of proteomics datasets [56]. Many of the normalization techniques originally developed for correcting gene expression microarrays have also been shown to be suitable for proteomics data [57].

1.3.4 Statistical Analysis

Omics data is high dimensional where thousands of features (e.g., genetic variants, transcripts, or proteins) are measured simultaneously and where only a minority of them contain information that is valuable for the study. Statistical analysis aims to reduce the dimension of the data by distinguishing between the biologically relevant features and the features that mainly contain noise. However, the high levels of technical and biological variability, in combination with the often limited number of replicated samples, make the statistical analysis challenging, and dedicated methods are therefore necessary.

In genome resequencing, statistical analysis is based on calling of genetic variants, that is, identifying the true differences between the sequence and reference genome. Resequencing data contains millions up to billions of data points while the number of true variants may be very few, and a major challenge is therefore to avoid false positives. For resequencing of single individuals or clones, calling is done in discrete steps depending on the ploidy of the investigated organism (e.g., homozygosity and heterozygosity variants for a diploid genome). A statistical score is calculated for each variant depending on the likelihood of the mutation being true and not explained by random sequencing errors. Commonly used variant callers are UnifiedGenoTyper or the HaplotypeCaller in the GATK package, which use Bayesian models to calculate the most likely genotype [41], but alternative methods have also been developed [58]. For the analysis of cancer tissues, where the sample is often heterogeneous and contains multiple clones with different configurations of somatic mutations, calling is instead done based on the relative frequency of each variant. This is in essence a harder problem since somatic mutations may occur at very low abundance, and, in order to minimize the number of false positives, the calling of variants is often done in a pairwise setting using both tumor and normal tissue from the same individual [42].

For RNA-seq and proteomics data, the analysis is done featurewise, where the transcripts or proteins are statistically assessed individually. Statistical methods that specifically describe the complex variance structure of the data are necessary to distinguish between features with a true effect and features with random patterns caused by the biological and technical variability. Furthermore, the counting of sequence fragments and peptides results in discrete data, which needs to be described by non-normal statistical models (e.g., generalized linear models). The

estimation of the feature-specific variability is also hard, especially when only a few replicates are available, and robust estimators, often implemented in an empirical Bayesian setting, are typically applied. These modeling approaches have been implemented in several methods for the identification of differentially expressed genes from RNA-seq and proteomics data (e.g., edgeR, deSeq2, QPROT) [39, 43, 44].

1.4 Data Integration: From a List of Genes to Biological Meaning

The data processing and analysis described in the previous section results in a "gene list" consisting of variants, transcripts, or proteins that are deemed statistically significant and thus contains relevant biological information. The next main step of the integrative analysis, namely data integration, is to analyze this gene list from a biological and biochemical point of view. This is, however, a nontrivial undertaking because the data still has a high dimensionality where the number of significant features often is in the range of 100–1000. The aim of the data integration is to systematically put the list of features into context and identify specific biological functions that are of importance in the experiment. In this process, the dimension of the data is further reduced, which enables a more detailed and refined biological interpretation. For example, in a transcriptome experiment, the integrative analysis can combine the differentially expressed transcripts with information from external data resources to identify activated pathways or regulated functional subsystems. As an example, Väremo *et al*. [59] integrated transcriptome data from patients with type 2 diabetes into a human metabolic network in order to find markers for diabetes. In another study, Delmotte *et al*. [60] analyzed transcriptomics and proteomics data from the bacterium *Bradyrhizobium japonicum* living in symbiosis with the soy bean plant *Glycine max*. The combined transcriptome and proteome data was integrated into a database with 15 different gene functional categories to facilitate the interpretation of the data. Integrative analysis is also commonly used to interpret resequencing data, and one example is the analysis of the genetic signatures in the Greenland Inuit population in relation to diet and climate adaptation [61].

Gene set analysis (GSA) is the most common type of integrative analysis, where predefined gene sets are used as the basis of integration. A gene set is a collection of genes that share a common attribute, property, or function. The gene sets are defined *a priori* from information collected from various biological databases and resources. Many of the methods for finding significant gene sets were originally developed for microarray analysis, before the era of NGS, but are also applicable for integrating omics data generated by newer measurement platforms [62]. In this section, we will first describe how gene set collections can be constructed from different databases, and then present and explain the different methods for GSA.

1.4.1 Data Resources for Constructing Gene Sets

Gene sets contain the *a priori* information in the GSA and are typically defined from biological information present in databases and biological networks (Table 1.2). We will describe the properties of the most commonly encountered data resources in more detail.

1.4.1.1 Gene Ontology Terms

The Gene Ontology (GO) is a bioinformatics resource that provides a classification of genes into different terms based on three main categories: biological process (BP), molecular function (MF), and cellular compartment (CC) [63]. The GO terms are organized in a hierarchical structure according to a directed acyclic graph (DAG), meaning that a more specific GO term has one or more parent terms defining related and more general classes. The terms are general in the sense that they are the same regardless of the organism. For many model organisms, such as human or yeast, the gene to GO term association can be downloaded from www.geneontology.org, or from organism-specific databases (such as www.yeastgenome.org). For organisms that are lacking a gene to GO term relationship, the GO term annotation can be inferred by comparing the gene sequences with gene sequences from closely related model organisms, using, for example, Blast2GO [69]. GO provides a many-to-many mapping, where several genes can be classified into a single GO term and a single gene can be annotated with multiple GO terms. The genes associated with a specific GO term defines a corresponding gene set.

1.4.1.2 KEGG and Reactome

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [65] (www.kegg.jp) is a resource where genes have been organized into different biological pathways. The main focus of KEGG is on metabolic pathways, but it also contains descriptions of other biological functions such as transporters and pathways involved in the cellular response to stress. KEGG has also defined KEGG orthologies (KO), which are coupled to the pathways and defined based on orthologous groups of genes,

Table 1.2 Examples of data resources containing gene sets and biological networks.

Biological resource	Type of gene sets	Supported organisms	References
Gene ontology terms	Ontology	Many	[63, 64]
KEGG pathways	Pathways	Many	[65]
Reactome pathways	Pathways	Only human	[66]
Genome-scale metabolic models (GEMs)	Subsystems/pathways, metabolites	Limited	[67]
Transcription factor binding	Transcriptional regulation	Many	[68]

that is, genes in different organisms with the same origin which are therefore likely to share the same function. This makes the KEGG pathways applicable to a wide range of evolutionarily distant organisms. From a KEGG pathway, gene sets can be formed by considering all its associated genes defined in a specific organism. In addition to KEGG, the Reactome database also can categorize human genes into pathways [66].

1.4.1.3 Genome-Scale Metabolic Reconstructions

Genome-scale metabolic reconstructions (GENREs) are detailed descriptions of a cell's metabolism [70] containing genes, metabolites, and reactions. A reaction normally represents a biochemical reaction that converts reactants to products, and it can be catalyzed by one or more enzymes (either as a complex or as isoenzymes). The genome-scale reconstruction can be converted into a mathematical model, called the genome-scale metabolic model (GEM), which can be used to simulate the metabolism under different conditions. There are many genome-scale metabolic models available, for many different species including human [71], yeast [72], and the bacterium *Escherichia coli* [73]. A comprehensive list of reconstructed GEMs for different species can be found at http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms [74] and http://biomet-toolbox.org/index.php?page=models [75]. The interpretation of high-throughput data using integrative analysis is an important application for GEMs. For human metabolism, different types of omics data (e.g., transcriptomics, proteomics, and metabolomics) have been integrated into a general metabolic network to create tissue-specific models [76, 77]. There exist several tools for constraint-based simulations of GEMs [78, 79]. However, in order to use the GEM as a scaffold for integrative analysis and GSA, simulations are not necessary. Instead, the topology of the metabolic network as described by the GEM can be used. Since GEMs are detailed descriptions of the metabolism, the genes, metabolites, and reactions can be divided into parts corresponding to different pathways and model subsystems. Gene sets for the integrative analysis can then be created from the associated genes [80]. Another way to form gene sets is to consider genes that share a common metabolite in the metabolic network (either as a reactant or product) [81]. For a review on omics data integration into GEMs, we refer the reader to Hyduke *et al*. [82].

1.4.2 Gene Set Analysis

The general approach to GSA is to analyze gene sets based on the significance of their associated genes. The process is demonstrated in Figure 1.4. A list of significant genes is obtained from the statistical analysis together with a gene-specific measure of significance, typically a p-value. If information about the direction of the effect is available (e.g., up- and downregulation of transcripts), it can also be used together with the p-value as input to the GSA, which enables separate analysis of up and down responses. A gene set significance value is then derived based

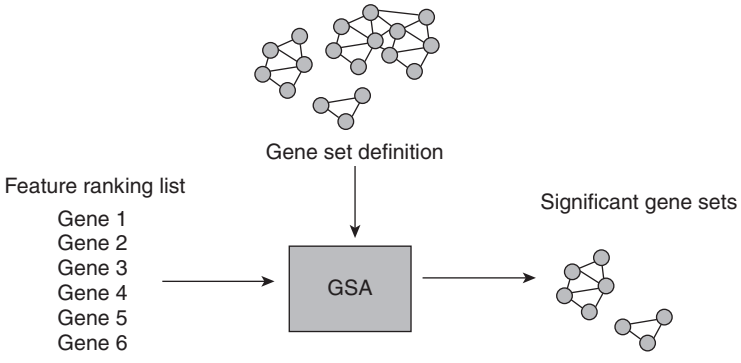

Figure 1.4 Concept of gene set analysis (GSA), which is the most common type of integrative analysis. The input to the GSA is a ranked list of significant genes identified in the statistical analysis part of data processing. During the GSA, each gene together with its *p*-value is integrated into an *a priori* defined gene set. A significance score is used to identify gene sets with biological functions that are significantly enriched in the gene list.

on the value of the score compared to a null model (e.g., a model assuming no effects between the studied conditions). A wide range of methods for GSA have been proposed applying different ways to calculate the gene set score and its significance (Table 1.3) [91]. These methods can be divided into list-based methods, which use a nonordered list of significant genes without any quantitative information about their significance, and rank-based methods, which use all genes analyzed in the experiment together with their respective *p*-values.

Table 1.3 Examples of methods for GSA.

Method	Type	References	Online resources/ packages
Overenrichment test	List-based	[83]	David, GOstats
Gene set enrichment analysis (GSEA)	Rank-based	[84]	www.broadinstitute.org/gsea/
Combination of *p*-values	Rank-based	[85]	Piano, www.biomet-toolbox.org/
Minimum hypergeometric score (GOrilla)	Rank-based	[86]	http://cbl-gorilla.cs.technion.ac.il/
GeneTrail	List-based or rank-based	[87]	http://gene-trail.bioinf.uni-sb.de
Enrichr	List-based	[88]	http://amp.pharm.mssm.edu/Enrichr/
SAFE	Rank-based	[89]	https://www.bioconductor.org/packages/safe/
MaxMean statistics	Rank-based	[90]	—

1.4.2.1 Gene Set Overenrichment Tests

One of the most common ways of finding significant gene sets is to perform an overenrichment test to assess whether there is an over-representation of genes from the gene set among the set of significant genes. In its most basic form, the enrichment test requires the user to provide a *p*-value cut-off to define which features that are considered significant. Then, for each gene set, the proportion of significant and nonsignificant features associated with the gene set is calculated and, under the assumption of independence between genes, a gene-set-specific *p*-value is derived from a hypergeometric distribution or by applying Fisher's exact test [92]. The resulting *p*-value describes the probability that an enrichment at least as extreme as what is observed in the data happens purely due to chance. Several tools that apply the overenrichment test are available, especially for GO term overenrichment, such as the web-based tool DAVID [93], the Cytoscape plugin BiNGO [94], and the R-package GO-stats [95]. Several generalizations of the overrepresentation test have also be developed, for example, topGO, which implements procedures that utilize the dependences from GO DAG to reduce the false positive rate, and GO-Bayes, which improves the inference of each gene set by incorporating prior information from functionally similar GO terms [96–98].

1.4.2.2 Rank-Based Enrichment Tests

Rank-based methods take advantage of all genes analyzed in the experiments together with their significance score (e.g., *p*-value or effect estimate); a user-defined significance cut-off is thus not necessary. Rank-based methods make therefore explicit use of the complete rank list, including both significant and nonsignificant genes. The general approach of rank-based GSA is to calculate a score for each gene set based on the significance scores of all its genes compared to the significance scores of the genes that are not part of the gene set. A high gene set score means that the genes in the gene set have, compared to the other genes in the gene list, high significance scores and are thus enriched at the top of the gene list. One of the first methods for rank-based GSA is the gene set enrichment analysis (GSEA), which was first introduced by Mootha *et al.* [99] and further refined by Subramanian *et al.* [84]. Mootha and coauthors used GSEA for integrative analysis of the transcriptional response of human muscle cells in diabetics, where they showed that the genes involved in oxidative phosphorylation were coordinately downregulated relative to the controls. For each gene set, GSEA calculates an enrichment score by walking down the ranked gene list and summing the contribution of each gene. The contribution of each gene is either positive or negative depending on whether it is a part of the gene set or not. A enrichment *p*-value is derived based on the highest value of the enrichment score using permutations of samples or genes [84].

Another way of obtaining the significance of a gene set is to combine the *p*-values of the involved genes using Fisher's method [100]. The *p*-values for all

genes in the gene set are combined using the following formula:

$$X = -2\sum_{i=1}^{k} \log(p_i)$$

where p_i are the individual p-values for each gene, and k is the number of genes in the gene set. Under the assumption of independence, a gene set p-value can be calculated from the chi-square distribution with $2k$ degrees of freedom. Stouffer's method is similar to Fisher's but it first converts the p-values into Z-scores, which are then summed into a gene set score: a corresponding gene set p-value can then be calculated from the normal distribution. Fisher's and Stouffer's methods are implemented in the R-package Piano [85], which can also be accessed as an online resource at www.biomet-toolbox.org [101].

There are also several other online tools that can be used for GSA together with both predefined and user-defined gene sets. For example, the web service GOrilla (http://cbl-gorilla.cs.technion.ac.il/) [86] performs GSA based on a ranked list of genes using a generalization of Fisher's exact test and provides a visualization of the significant gene sets in the GO structure. GeneTrail (http://genetrail.bioinf.uni-sb.de/) employs a wide range of databases for defining the gene set and lets the user choose between a hypergeometric list-based test or GSEA [87].

1.4.3 Networks and Network Topology

The gene-to-gene set relationships can also be considered as a biological network, or a graph, where genes (nodes) are connected (with edges) if they share a common property or function. Biological networks play an important part in systems biology and are often used to describe protein–protein interactions [102], interactions between transcription factors and genes, and transcriptional coexpression relations between genes. The organization and topology of many biological networks have been shown to follow a scale-free distribution, meaning that there are both highly connected genes ("hubs") and genes with only a few connections [103]. The genes that serve as hubs are often of primary importance in the system. By constructing an interaction network from the omics data, or by integrating the data to an existing network, the condition-specific properties of the system can be identified. One important application is to find key components or highly connected modules in the data, that is, genes or proteins that interact with a large number of other components. This has, for example, been used to identify genes crucial to the development of disease [104].

One way to identify network modules is to use unsupervised clustering to identify genes that are coexpressed or contain genetic variants in a large number of samples. However, this approach does not take advantage of any *a priori* information about the properties of the biological networks. Instead, modules can also be identified using integrative analysis where data is integrated on top of a biological network constructed from resources such as KEGG or GEMs. In contrast to GSA, network-based integrative analysis provides information about important

parts of the network instead of only finding significant sets of genes. Several different methods for identification of context-specific, highly connected modules in biological networks have been developed. For example, the reporter subnetwork algorithm [81] aims to find differentially expressed modules by converting the *p*-values from transcriptomics experiments to *Z*-scores and overlay the scores on top of a genome-scale metabolic model. ActiveModules [105] is a Cytoscape [106] plugin that finds differentially expressed modules in protein–protein interaction networks or protein–DNA interaction networks, also by using *Z*-scores. GiGa [107] is another Cytoscape plugin that uses the ranking list of individual genes to find differentially expressed modules. We refer the reader to the papers by Ideker and Krogan [108] and Kristensen *et al*. [109] for more complete reviews of the available methods.

1.5 Outlook and Perspectives

The development of high-throughput techniques, in particular NGS, has resulted in an explosion of biological information contained in biological databases. Genome databases, such as GenBank (NCBI), have grown in terms of the number of contained sequences, and in early 2015 GenBank reported that their database contains genomic sequences from at least 300 000 species [110]. In addition to submitted genomic sequences, other forms of experimental data are accumulating in various repositories. One example is the cancer genome atlas, TCGA [111], which collects genomics, transcriptomics, proteomics data, and other types of data relevant for the study of a wide range of cancer types. The development of NGS techniques has also facilitated the study of new organisms by providing cost-efficient approaches for *de novo* characterization of the genomes and transcriptomes. As a consequence, model organisms can, in many situations, be replaced by a more relevant nonmodel organism. Despite the many challenges in storing and handling large amounts of data, the availability of high-throughput data opens up new possibilities for researchers to gain new insights into their fields. Systems biology, as a field where understanding of the biological system is the main focus, provides the means to take advantage of the vast amount of biological information that is currently generated. This makes integrative analysis an increasingly important tool for data-driven biological and medical research, which will provide biological interpretations to the rapidly accumulating volumes of omics data.

The increasing amount of information and data available in public repositories leads to two important considerations when it comes to biological knowledge retrieval and data mining. The first consideration is related to data reliability. Gene sets for integrative analysis are constructed under the assumption that the gene associations are true. It is therefore of essence that the databases used for gene set construction are of high quality and provide information as error-free as possible. Many databases ensure a low error rate through manual curation or

by incorporating only components, annotations, and interactions that have been experimentally verified. There are, however, also databases that have less strict quality criteria and may contain putative interactions and hypothetical functional information. For integrative analysis, it is of vital importance to use information and associations that are as correct as possible. Otherwise, false relationships can be introduced, which can result in erroneous interpretations and incorrect biological conclusions. An important future challenge is therefore to grow the databases by incorporating all the newly generated information and at the same time keep and further improve the quality and veracity of the data.

The second consideration is the need for standardization of databases and their content. Today, genes and other features may be associated with different types of nomenclatures and therefore have different names in different databases. In some cases, there are translations available between databases and nomenclatures, meaning that, for example, the gene identifier in one database can be mapped to a gene identifier from another database. However, if such translations are missing, it is not possible to utilize a large part of the information that is available in databases or to use the information as scaffolds in integrative analysis. This is also of importance when constructing GEMs where information of genes and reactions needs to be extracted from multiple data sources. For genes, only a few model species (e.g., human and yeast) have adapted universally accepted nomenclatures. Similarly, for small molecules, such as metabolites, there is no standard naming, but chemical identifiers such as ChEBI [112] or INCHI codes [113] are often used to identify metabolites, which can be useful when comparing or merging several different models. Thus, increased standardization of molecular databases is crucial to ensure easy and painless access of biological information and, hence, efficient integrative analysis of omics data in as many types of biological systems as possible.

References

1. Cvijovic, M., Almquist, J., Hagmar, J., Hohmann, S. *et al.* (2014) Bridging the gaps in systems biology. *Mol. Genet. Genomics*, **289**, 727–734.
2. Kitano, H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.
3. Kitano, H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
4. Robinson, S.W., Fernandes, M., and Husi, H. (2014) Current advances in systems and integrative biology. *Comput. Struct. Biotechnol. J.*, **11**, 35–46.
5. Kim, S.J., Kim, S.H., Kim, J.H., Hwang, S. *et al.* (2016) Understanding metabolomics in biomedical research. *Endocrinol. Metab. (Seoul)*, **31** (1), 7–16.
6. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
7. Liu, Z., Liu, L., Österlund, T., Hou, J. *et al.* (2014) Improved production of a heterologous amylase in Saccharomyces cerevisiae by inverse metabolic engineering. *Appl. Environ. Microbiol.*, **80**, 5542–5550.
8. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K. *et al.* (2011) Exome sequencing as a tool for Mendelian

disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
9. Martincorena, I. and Campbell, P.J. (2015) Somatic mutation in cancer and normal cells. *Science*, **349**, 1483–1489.
10. McGettigan, P.A. (2013) Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.*, **17**, 4–11.
11. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
12. Noble, W.S. and MacCoss, M.J. (2012) Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.*, **8**, e1002296.
13. Yates, J.R., Ruse, C.I., and Nakorchevsky, A. (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.*, **11**, 49–79.
14. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.
15. Altelaar, A.M., Munoz, J., and Heck, A.J. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.*, **14**, 35–48.
16. Heather, J.M. and Chain, B. (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics*, **107**, 1–8.
17. Quail, M.A., Smith, M., Coupland, P., Otto, T.D. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
18. Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
19. Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013) The advantages of SMRT sequencing. *Genome Biol.*, **14**, 405.
20. Richards, A.L., Merrill, A.E., and Coon, J.J. (2015) Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.*, **24**, 11–17.
21. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M. *et al.* (2011) The quantitative proteome of a human cell line. *Mol. Syst. Biol.*, **7**, 549.
22. Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J. *et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, **7**, 548.
23. Shi, L., Reid, L.H., Jones, W.D., Shippy, R. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
24. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
25. Rung, J. and Brazma, A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.
26. Verducci, J.S., Melfi, V.F., Lin, S., Wang, Z. *et al.* (2006) Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiol. Genomics*, **25**, 355–363.
27. Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
28. Patel, R.K. and Jain, M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.
29. Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K. *et al.* (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, **6**, 2086–2094.
30. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
31. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
32. Trapnell, C., Pachter, L., and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

33. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
34. Eng, J.K., McCormack, A.L., and Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.*, **5**, 976–989.
35. Cottrell, J.S. and London, U. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
36. Tanner, S., Shu, H., Frank, A., Wang, L.-C. *et al.* (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, **77**, 4626–4639.
37. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
38. Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf.*, **11**, 94.
39. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
40. Callister, S.J., Barry, R.C., Adkins, J.N., Johnson, E.T. *et al.* (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.*, **5**, 277–286.
41. McKenna, A., Hanna, M., Banks, E., Sivachenko, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
42. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
43. Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
44. Choi, H., Kim, S., Fermin, D., Tsou, C.C. *et al.* (2015) QPROT: statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. *J. Proteomics*, **129**, 121–126.
45. O'Rawe, J.A., Ferson, S., and Lyon, G.J. (2015) Accounting for uncertainty in DNA sequencing data. *Trends Genet.*, **31**, 61–66.
46. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.
47. Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P. *et al.* (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.*, **9**, e1003031.
48. Lin, W., Wang, J., Zhang, W.-J., and Wu, F.-X. (2012) An unsupervised machine learning method for assessing quality of tandem mass spectra. *Proteome Sci.*, **10**, 1.
49. Bern, M., Goldberg, D., Mcdonald, W.H., and Yates, J.R. (2004) Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, **20**, i49–i54.
50. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
51. Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.
52. Jiang, H. and Wong, W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.
53. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq

data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
54. Nesvizhskii, A.I. and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today*, **9**, 173–181.
55. Kultima, K., Nilsson, A., Scholz, B., Rossbach, U.L. *et al.* (2009) Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell. Proteomics*, **8**, 2285–2295.
56. Griffin, N.M., Yu, J., Long, F., Oh, P. *et al.* (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.*, **28**, 83–89.
57. Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics*, **4**, 419–434.
58. Pabinger, S., Dander, A., Fischer, M., Snajder, R. *et al.* (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings Bioinf.*, **15**, 256–278.
59. Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A. *et al.* (2015) Proteome-and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell Rep.*, **11**, 921–933.
60. Delmotte, N., Ahrens, C.H., Knief, C., Qeli, E. *et al.* (2010) An integrated proteomics and transcriptomics reference data set provides new insights into the *Bradyrhizobium japonicum* bacteroid metabolism in soybean root nodules. *Proteomics*, **10**, 1391–1400.
61. Fumagalli, M., Moltke, I., Grarup, N., Racimo, F. *et al.* (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*, **349**, 1343–1347.
62. Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2015) Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Briefings Bioinf.*, **17** (3), 393–407.
63. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
64. Consortium GO (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
65. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
66. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
67. O'Brien, E.J., Monk, J.M., and Palsson, B.O. (2015) Using genome-scale models to predict biological capabilities. *Cell*, **161**, 971–987.
68. Oliveira, A.P., Patil, K.R., and Nielsen, J. (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.*, **2**, 17.
69. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
70. Oberhardt, M.A., Palsson, B.Ø., and Papin, J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.*, **5**, 320.
71. Mardinoglu, A., Gatto, F., and Nielsen, J. (2013) Genome-scale modeling of human metabolism – a systems biology approach. *Biotechnol. J.*, **8**, 985–996.
72. Sánchez, B.J. and Nielsen, J. (2015) Genome scale models of yeast: towards standardized evaluation and consistent omic integration. *Integr. Biol.*, **7**, 846–858.
73. McCloskey, D., Palsson, B.Ø., and Feist, A.M. (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.*, **9**, 661.
74. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L. *et al.* (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, **7**, 129–143.

75. Cvijovic, M., Olivares-Hernández, R., Agren, R., Dahr, N. *et al.* (2010) BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acids Res.*, **38**, W144–W149.
76. Shlomi, T., Cabili, M.N., Herrgård, M.J., Palsson, B.Ø. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.
77. Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N. *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.*, **8**, e1002518.
78. Schellenberger, J., Que, R., Fleming, R.M., Thiele, I. *et al.* (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.*, **6**, 1290–1307.
79. Agren, R., Liu, L., Shoaie, S., Vongsangnak, W. *et al.* (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput. Biol.*, **9**, e1002980.
80. Joyce, A.R. and Palsson, B.Ø. (2006) The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, **7**, 198–210.
81. Patil, K.R. and Nielsen, J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2685–2689.
82. Hyduke, D.R., Lewis, N.E., and Palsson, B.Ø. (2013) Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.*, **9**, 167–174.
83. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
84. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
85. Väremo, L., Nielsen, J., and Nookaew, I. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.*, **41** (8), 4378–4391.
86. Eden, E., Navon, R., Steinfeld, I., Lipson, D. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinf.*, **10**, 48.
87. Backes, C., Keller, A., Kuentzer, J., Kneissl, B. *et al.* (2007) GeneTrail – advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.
88. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.*, **14**, 128.
89. Barry, W.T., Nobel, A.B., and Wright, F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.
90. Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1** (1), 107–129.
91. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
92. Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
93. Jiao, X., Sherman, B.T., Huang da, W., Stephens, R. *et al.* (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
94. Maere, S., Heymans, K., and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
95. Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO

term association. *Bioinformatics*, **23**, 257–258.
96. Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
97. Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
98. Zhang, S., Cao, J., Kong, Y.M., and Scheuermann, R.H. (2010) GO-Bayes: gene ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics*, **26**, 905–911.
99. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A. *et al.* (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
100. Fisher, R.A. (1948) Answer to question 14 on combining independent tests of significance. *Am. Stat.*, **2**, 30–31.
101. Garcia-Albornoz, M., Thankaswamy-Kosalai, S., Nilsson, A., Väremo, L. *et al.* (2014) BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. *Nucleic Acids Res.*, **42**, W175–W181.
102. Wetie, A.G.N., Sokolowska, I., Woods, A.G., Roy, U. *et al.* (2014) Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.*, **71**, 205–228.
103. Barabasi, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
104. Segal, E., Friedman, N., Kaminski, N., Regev, A. *et al.* (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.*, **37**, S38–S45.
105. Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.
106. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
107. Breitling, R., Amtmann, A., and Herzyk, P. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinf.*, **5**, 34.
108. Ideker, T. and Krogan, N.J. (2012) Differential network biology. *Mol. Syst. Biol.*, **8**, 565.
109. Kristensen, V.N., Lingjærde, O.C., Russnes, H.G., Vollan, H.K.M. *et al.* (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, **14**, 299–313.
110. Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J. *et al.* (2015) GenBank. *Nucleic Acids Res.*, **43**, D30.
111. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
112. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
113. Sumner, L.W., Amberg, A., Barrett, D., Beale, M.H. *et al.* (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.