

Chapter 1

Introduction to Omics



Priyanka Narad and S. V. Kirthanashri

Abstract Omics technologies also referred as high-dimensional biology encompasses the cells, tissues, and organisms in a manner that integrates the data from various platforms and helps in its interpretation. It primarily detects the genes (genomics), mRNAs (transcriptomics), proteins (proteomics), and metabolites (metabolomics) in a nontargeted and non-biased manner. The integration and inter-relationships between networks of biological processes is termed as systems biology. The approach provides hope for unravelling the intricate details in various aspects of biology and accelerates innovation in healthcare. Understanding the various dimensions encompassing not only the three levels constituting the central dogma of life but also the intermediate metabolites is significant for the scientists to cover new horizons in drug discovery and disease regulation. This chapter outlines the scope of omics, experimental design in omics research, and its applications. It will also provide an overview to the usage of languages like R for analyzing high-throughput data from all branches of “omics” technologies.

The primary focus is to understand omics approaches that enable the validation of large-scale data that is generated from various experimental platforms. Systems biology and omics data are way apart from hypothesis-driven traditional studies. The systems biology experiments generate hypothesis by employing all data that needs to be further analyzed.

Omics technology applied majorly for accurate understanding of normal physiological processes and gaining knowledge related to disease processes which involves screening, diagnosis, and prognosis that provides an understanding of the etiology of diseases.

Keywords Omics · Systems biology · R language

P. Narad

Amity Institute of Biotechnology, Amity University, Noida, Uttar Pradesh, India
e-mail: pnarad@amity.edu

S. V. Kirthanashri (✉)

Amity Institute of Molecular Medicine & Stem Cell Research, Amity University, Noida, Uttar Pradesh, India
e-mail: svkirthanashri@amity.edu

1.1 Background

In biology the suffix -omics refers to huge biological molecules; the broad analysis of large biological molecules was needed to be studied in detail as the conclusion of human genome project (HGP) in 2001. The HGP revealed that the human genome contained lesser number of genes and biological process were regulated not particularly on DNA sequence but involved various other processes, and with this evolved the new branch of study termed the omics (Hood and Rowen 2013).

This technology deciphered the cell, tissue, and organism in a holistic way around central dogma for the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) in the samples (specific biological component). Since the technology is non-biased, they are also referred to as high-dimensional biology, while the integration of these is the systems biology. Following the discovery of DNA structure by Watson-Crick in 1953, a series of inventions and discoveries followed. The development of PCR by Kary Mullis opened all possible channels in molecular biology research. The progress in *Omics* started from the development of genomics further followed by transcriptomics and finally the proteomics, and the term was coined in 1994 by Marc Wilkins. This was possible because of advanced development in techniques like high-resolution two-dimensional electrophoresis. The cascade of events in *Omics* is depicted in Fig. 1.1. The advantage of the omics study is that they reveal specific results that promote understanding. As the omics technology is of immense potential, they have been explored in various branches of medical and health science. This technology can help to understand the etiology of disease condition through the process of screening, diagnosis, and prognosis and also for the biomarker discovery to be made easy as they involve simultaneous investigation of multiple molecules (Poisot et al. 2013). Further *Omics* is of great use in drug discovery and toxicity assessment. Pharmacogenomics deals with the connection of genomics and pharmacology to examine the role of inheritance in individual variation in drug response utilized to individualize and optimize drug therapy. They help in the field of oncology to evaluate rigorous systemic toxicity and unpredictable efficacies that are

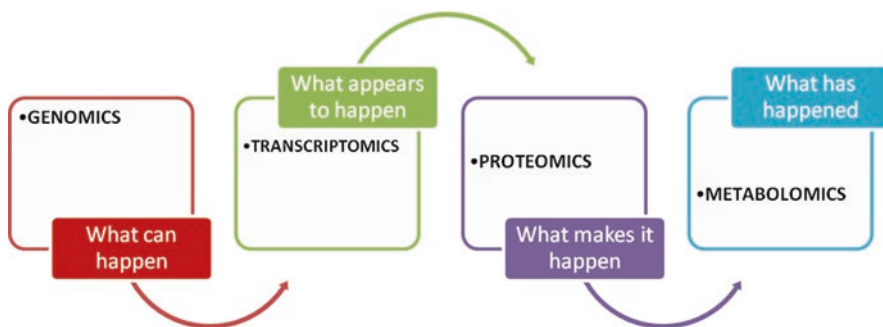


Fig. 1.1 Cascade of *Omics* development

hallmarks of cancer therapies. These technologies are helpful in selecting novel targets for the treatment including conditions like cancer, cardiovascular disease, and obesity. In the future, systems biology promises to develop new approaches that will be predictive, preventive, and personalized (Sagner et al. 2017). Research in the field of obstetrics and gynecology is currently taking advantage of these possibilities which can be used to solve the problems related to fertility. This review aims to provide a complete overview of various omics technologies available.

1.2 Overview of Omics

The omics technology can be classified into various types depending on their function. Figure 1.2 highlights a few of various omics technologies that are presented in detail in the following chapters.

1.2.1 Genomics

This refers to the interdisciplinary study based on evaluating the structure and function and mapping of the genomes. In short this is the study of a set of genes, the inheritance substance. The term genomics was coined by Tom Roderick in 1986 on mapping the human gene. The possible and highly researched areas under genomics include the functional genomics, metagenomics, and epigenomics (Feinberg 2010).

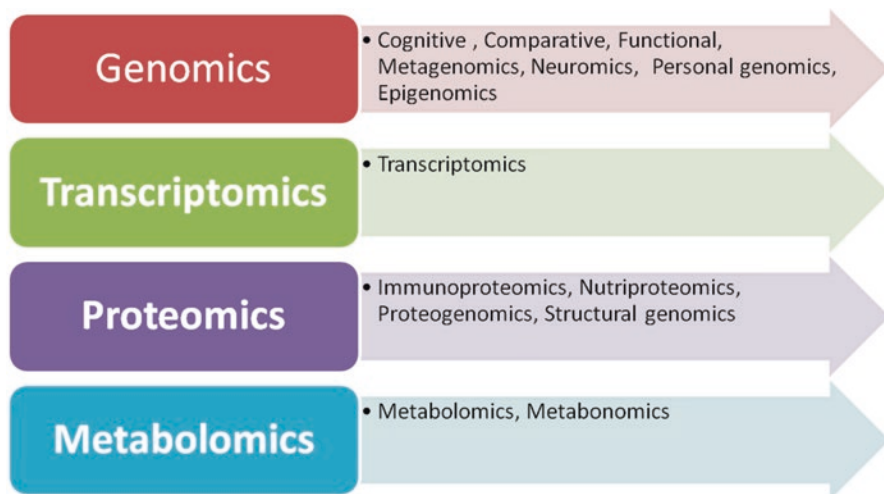


Fig. 1.2 Various omics technologies

1.2.2 Transcriptomics

The presence of mRNA in the sample reflects the abundance level of the corresponding gene. Gene expression involves the detection and classification of mRNA mixture in a specific sample. The goal of gene expression profiling is to differentiate the mRNA mixtures from different samples. Contrary to genotyping, gene expression categorizes the level of gene expression. The variation of the transcriptome can be seen over time between cell types and change according to environmental conditions (Hubank 2004).

1.2.3 Proteomics

The proteome refers to the total proteins expressed by a cell, tissue, or organism. The proteome is highly variable over time, shows species variation, and alters depending on environmental conditions. Proteomics is defined as the study that provides insights to protein functions in biological systems. Due to the variability and abundance of proteins in biological samples, there is a requirement to develop technologies to detect a wide range of proteins in samples of different origins. Currently exploited proteomic technologies are mass spectrometry (MS) and protein microarrays using capturing agents such as antibodies. However, the high dynamic range (abundance and concentration) of proteins complicates this type of proteomic analysis (Chandramouli and Qian 2009).

1.2.4 Metabolomics

The small molecules (e.g., lipids or vitamins) referred to as metabolites constitute the metabolome. The interaction between genetic, environmental, lifestyle, and other factors results in metabolic phenotypes. Interaction of metabolome with other biological macromolecules in the cell results in metabolic pathways. The metabolic profiles of biological sample represent the metabolomics which are changeable and time dependent and had a wide range of chemical structures (Bino et al. 2004).

1.3 Overview of Systems Biology

1.3.1 Systems Biology

The genome is the total DNA of a cell in the organism. The human genome contains about 3.2 billion bases with 30,000–40,000 protein-coding genes. The microarray technique enables quick analysis of the genes and also helps in examining the

differences in the DNA sequences and expression of genes, which help to analyze the chromosomal abnormalities. Variation in DNA sequence leads to single nucleotide polymorphism (SNP), which plays an important role in pharmacogenomics to explore individual patient responses to drugs. The total mRNA in the cell or organism is called the transcriptome, and they are the template for protein synthesis and are known as translation. The transcriptome reflects the genes that are actively expressed at any given moment. The advent of microarray techniques has led to the advancement of the genomics and transcriptomics. Microarrays measure changes only in mRNA that makes it complex for data interpretation. Most of the mechanistic and mathematical models are used in systems biology approach for data interpretation. Due to the large number of parameters, variables, and constraints in cellular networks, numerical and computational techniques are often used (Likić et al. 2010).

1.4 Techniques Involved in Systems Biology

1.4.1 Analytical Techniques

Reproducibility of the samples is the major concern for omics experiments. Expression profiling is one of the dominant modes of DNA microarray. Complementary DNA (cDNA)/oligonucleotide is the probe to estimate the amount of mRNA in gene expression microarray. The probe gets amplified by polymerase chain reaction (PCR) and immobilized on a solid support (glass slide) by spotting them. Extraction of RNA from the sample is carried out followed by reverse transcription along with the addition of fluorescent dyes where cDNA are generated which are hybridized in the microarray slide. The chips which are the microarray glass slides are scanned by ultraviolet laser to detect the fluorescent signal produced by each gene to carry out image analysis. Samples for analysis in metabolomic experiments require fractionation (chromatography or electrophoresis) utilizing various chemical/physical properties of molecules that fasten the separation of the metabolites in liquid or gas phase. The commonly used tool for analysis of the metabolite is the mass spectrometry. The analytical techniques had their own limitations and advantages in terms of instrument sensitivity, resolution, mass accuracy, and dynamic range, while various techniques are researched to analyze the entire proteome or metabolome. For instance, the proteomics study involves characterization of proteins using electrospray ionization (ESI), matrix-assisted laser desorption/ionization (MALDI), and surface-enhanced laser desorption/ionization (SELDI) though reproducibility, accuracy, and mass range are always a limitation. The use of fluorescent tags in gel-based techniques like differential image gel electrophoresis (DIGE) and isotope-coded affinity tag (ICAT) labeling is employed with mass spectrometry to achieve better resolution. The nuclear magnetic resonance (NMR) spectroscopy and infrared spectroscopy have been routinely used for metabolite identification (van der Greef et al. 2004). Thus each technique has its unique

and meritorious way of detection compared to the other techniques; it all depends also on the sample and the objective of the experiment.

1.4.2 Data Analysis

The analysis techniques generate huge data that mandates sophisticated software (bioinformatics and statistics). The results of the genomics and transcriptomics microarray are often huge and complicated that often conclude in false-positive results, if not accurately analyzed. Proteomics theoretical database is often matched with the experimental analysis to enable protein identification and/or quantification, while in metabolomics, raw data processing is carried out to generate meaningful and interpretable data. Thus, the prime aim of the data analysis is to represent the data in readable/understandable format which can be used to generate further hypotheses for testing with no false-positive results (van der Greef et al. 2004).

In the following text, we will discuss few of these packages using R language and their utility for analysis of “omics” data.

1.4.3 R Language in Omics Analysis

R is a statistical language which is fully featured and equipped with several packages useful for the “omics” and other life sciences research. It has an interactive and user-friendly interface where one can make plenty of debugging. The use of the language is coherent, and there is an extensive documentation available on the Internet to perform the data analysis. Integration to the Bioconductor platform has extended the ability of performing analysis and an easy approach for high-throughput “omics” data analysis. Within the last decades, huge amount of data has been generated through various sophisticated techniques of genomics/proteomics and metabolomics. There has been an array of new technologies in the past which have made new discoveries and research easier. It is a common practice to analyze each of the “omics” data like proteomics, genomics, and transcriptomics through statistical approaches like *t*-test and ANOVA. The task at hand is to make sense of the sea of data; else data generation is of no use. Toward this, R and Bioconductor platforms together provide packages for the interpretation of high-throughput data generated from “omics.” There are numerous data analysis packages which offer great features to the person working on these samples. These include the packages which are computationally highly efficient for the purpose of handling large sample data; secondly these packages are able to perform reduction of the dimension by creating smaller spaces and analyzing the data; thirdly they are helpful in providing better insights to the biological system under observation. When we talk about the integrative approach for systems biology, analyzing both the datasets together is required for the understanding of the different levels of “omics.” For instance, now it is clear that any integration would need inputs

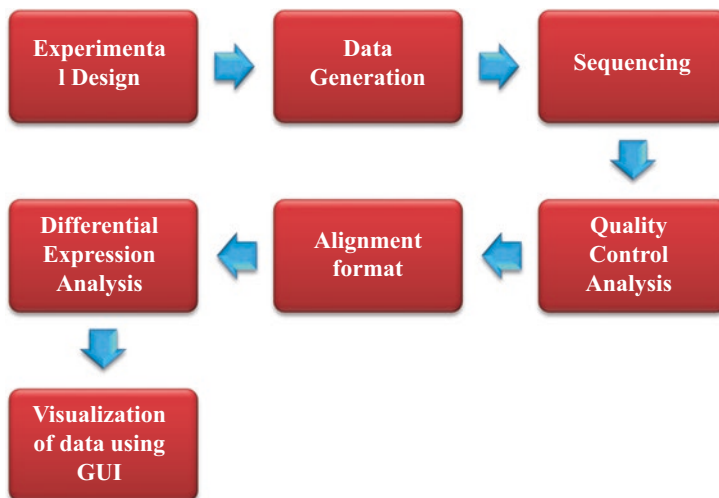


Fig. 1.3 General workflow for “omics” data analysis

from all branches of “omics” like transcriptomics, genomics, metabolomics, and proteomics in order to understand the biological processes in a comprehensive manner. Figure 1.3 provides a general workflow of an omics data analysis.

R and Bioconductor play an important role in extracting useful information from large-scale high-throughput “omics” data. R console is an interactive and user-friendly coherent language for data analysis. What makes R not the same as other programming dialects is its GUI for fast and simple transfer of information. Bioconductor (www.bioconductor.org) contains in silico software packages for interpretation of “omics” data which is generated from a number of experiments like microarray, SAGE, MS, and MS-MS. The packages at the Bioconductor platform can be split into three branches like the Annotation Data, the Experiment Data, and the Software. Some of the important packages are listed in Table 1.1.

1.4.4 *cpma*

This package performs phenotype analysis. Numeric values from the data are treated as the input.

1.4.5 *mlm*

This package is useful for fitting multiple linear models together. The argument consists of a formula, which consists of the description of the models, and data, which consists of the variable of the model.

Table 1.1 List of the packages for quality control and analysis of gene/protein microarray data

S. no.	Name of the package	Description
1.	Affy package	Oligonucleotide array analysis
2.	affymGUI	GUI for analysis of one-color affymetrix data
3.	ExpressionView	Visualization of possibly overlapping biclusters
4.	annmap	Genome annotation and visualization package
5.	DEGseq	Differential expressed gene analysis from RNA-Seq data
6.	Dexus	Differentially expressed genes in RNA-Seq
7.	MiChip	Differentially expressed data of miRNA for multiple species

1.4.6 *mixOmics*

This package describes a multiple variable-based system for the “omics”-based data and its analysis of information examination to the scientist who wants to give a few appealing properties. Mostly, the package is computationally productive to deal with huge informational collections, where the quantity of sample sets is significantly bigger than the quantity of tests. Furthermore, the package performs measurement lessening by anticipating the information into a littler subspace while catching and featuring the biggest feature selection from a variety of information, bringing about great perception of the natural framework under investigation. Finally, the information appropriation makes it profoundly adaptable to answer topical inquiries over various science-related fields (Rohart et al. 2017). *mixOmics* multivariate strategies have been effectively connected to factually coordinate informational collections created from contrast sciences ranging from the field of “omics” comprising of transcriptomics, proteomics, and metabolomics.

1.4.7 *integrOmics*

integrOmics productively performs integrative investigations of two kinds of “omics” factors that are estimated on similar examples. It incorporates a regularized form of standard connection investigation to illuminate relationships between two datasets and a scanty rendition of incomplete slightest squares (PLS) relapse that incorporates synchronous variable choice in both datasets. The helpfulness of the two methodologies has been shown already and effectively connected in different integrative examinations (Lê Cao et al. 2009).

1.4.8 *supraHex*

supraHex is an R package for preprocessing, normalizing, and imagining omics information. This console package devises a supra-hexagonal manual to process the information, and it gives a versatile functionality for after-dissection of the guide

and, all the more imperatively, takes into consideration overlaying extra information for multilayer omics information examinations. The application of supraHex was exhibited through its ability to apply DNA replication timing data, and it performs the same level of grouping and provides a real-time picture of the natural process. The scientist also observed that CpG overlaying to the replication map resulted in demonstrating the ability of supraHex to establish connections between CpG thickness and late replication data. Being part of the Bioconductor venture, supraHex is useful in making available to a wide network basically what might somehow or another be an unpredictable structure for the ultrafast comprehension of any forbidden omics information, both deductively and aesthetically (Fang and Gough 2014).

1.4.9 OmicCircos

OmicCircos is an R programming bundle used to create great roundabout plots for envisioning genomic varieties, including change designs, duplicate number varieties (CNVs), articulation examples, and also methylation designs. This method can be used to generate scatterplots for the abovementioned examples. Using the factual and realistic capacities in an R/Bioconductor condition, OmicCircos performs measurable examinations and presentations that come about utilizing bunch, boxplot, histogram, and heatmap groups. Moreover, OmicCircos offers various one of a kind capacity, including free track drawing for simple adjustment and combination, zoom capacities, connect polygons, and position-autonomous heatmaps supporting small representation (Hu et al. 2014).

References

- Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., & Trethewey, R. N. (2004, September 1). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9), 418–425.
- Chandramouli, K., & Qian, P. Y. (2009). Proteomics: Challenges, techniques and possibilities to overcome biological sample complexity. *Human Genomics and Proteomics: HGP, 2009*, 1.
- Fang, H., & Gough, J. (2014, January 3). supraHex: An R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. *Biochemical and Biophysical Research Communications*, 443(1), 285–289.
- Feinberg, A. P. (2010, October 13). Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nature Biotechnology*, 28(10), 1049.
- Hood, L., & Rowen, L. (2013, September). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, 5(9), 79.
- Hu, Y., Yan, C., Hsu, C. H., Chen, Q. R., Niu, K., Komatsoulis, G. A., & Meerzaman, D. (2014, January). OmicCircos: A simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer Informatics*, 13, CIN-S13495.
- Hubank, M. (2004, March 1). Gene expression profiling and its application in studies of haematological malignancy. *British Journal of Haematology*, 124(5), 577–594.

- Lê Cao, K. A., González, I., & Déjean, S. (2009, August 25). integrOmics: An R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21), 2855–2856.
- Likić, V. A., McConville, M. J., Lithgow, T., & Bacic, A. (2010). Systems biology: The next frontier for bioinformatics. *Advances in Bioinformatics*, 2010, 1.
- Poisot, T., Péquin, B., & Gravel, D. (2013, April 1). High-throughput sequencing: A roadmap toward community ecology. *Ecology and Evolution*, 3(4), 1125–1139.
- Rohart, F., Gautier, B., Singh, A., & Le Cao, K. A. (2017, November 3). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11), e1005752.
- Sagner, M., McNeil, A., Puska, P., Auffray, C., Price, N. D., Hood, L., Lavie, C. J., Han, Z. G., Chen, Z., Brahmachari, S. K., & McEwen, B. S. (2017, March 1). The P4 health spectrum—a predictive, preventive, personalized and participatory continuum for promoting healthspan. *Progress in Cardiovascular Diseases*, 59(5), 506–521.
- van der Greef, J., Stroobant, P., & van der Heijden, R. (2004, October 1). The role of analytical sciences in medical systems biology. *Current Opinion in Chemical Biology*, 8(5), 559–565.