

Introduction

1.1 FUNDAMENTALS OF MULTI-OMICS

Holistic understanding of human health and disease demands rendition of molecular complexity and variations at multiple levels such as the genome, epigenome, transcriptome, proteome, metabolome, and increasingly others. High-throughput technologies have revolutionized the medical research and biology has become increasingly dependent on data generated at these levels, which together is called as "multi-omics" data (Hasin et al., 2017). An abundance of multi-omics data has transformed the discipline of biology as well as medicine and generated opportunities for integrated system-level approaches. Extraction of meaningful information from the high dimensional multi-omics data have become feasible by applying approaches like, pattern recognition, machine learning (ML), and artificial intelligence. The COVID-19 pandemic has exhibited the pace with which information can be created and investigated by utilizing these approaches, and also revealed their limitations (D'Adamo et al., 2021). Translating "omics" into clinical practice has been initiated. Despite the endless potential of using big data for "personalized" or "precision" medicine, in reality, this remains very challenging. The only field of medicine that is extensively implementing these technologies is oncology, and even here, the usage are very limited. Several practical and ethical reasons are the cause of slow clinical translation of these low-cost technologies. An evident surge in medical big data generated from multiple sources will be witnessed in the future. It is perhaps the greatest challenge for the healthcare sector to integrate these resources in an effective, practical, and ethical manner.

Since the outset of sequencing and mapping of the human genome, novel techniques have come up with the possibility to acquire multiple molecular measurements within a cell or tissue. In Figure 1.1 interaction between four major entities are presented that are found across the cell. Each of them is constantly interacting, for instance, some proteins play the role of transcription factors and regulate the conversion of Deoxyribonucleic acid (DNA) to Ribonucleic acid (RNA). The cell metabolites regulate protein activity in the form of ligands but simultaneously gets modified via enzymatic action. The complex bio-molecular interactions suggest that investigating the biological system at a single level gives restricted insight into the cellular activities. Whereas, multi-omics studies help in digging the underlying biological information from complex biological systems of concern at a resolution that has never been explored. There can be multiple reasons for conducting multi-omics research, all with a common goal of achieving a comprehensive understanding of biological systems. For example, one might conduct an investigation on normal human lung tissues at the proteomics level to improve the understanding of protein activities, functional pathways, and interacting proteins in the lung. Another common reason for conducting an omics study can be to explore the connection between omics-based molecular measurements and a clinical outcome of concern, like survival time of cervical cancer, susceptibility of recurrence of breast cancer, or reaction towards a medication or treatment. In a nutshell, it can be said that taking advantage of omics-based measurements generated from multiple genomic platforms can help in the development of more reliable predictive or prognostic models that can be more accurate than using standard clinical approaches.

The multi-level understanding of human health and disease has started to regulate the disease management approaches and also how novel drugs and tests are being formulated and utilized. For instance, a multi-parameter genomics test known as Oncotype DX is designed to help the patients having early-stage breast cancer to predict the risk of recurrence and thus to decide whether they will be benefited from chemotherapy or not (Schildgen et al., 2019). Patients having lower risk of recurrence can securely forgo

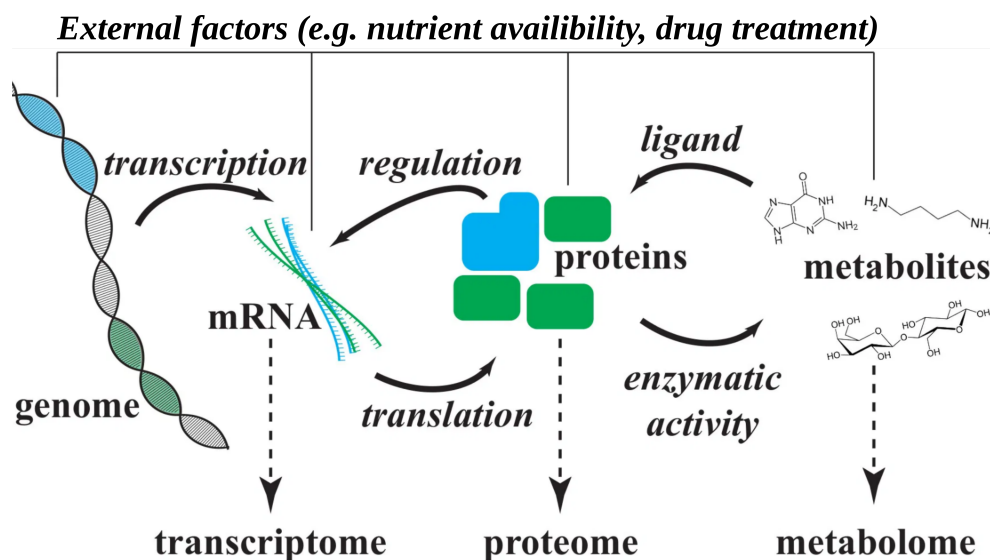


Figure 1.1. : A glimpse of multi-omics interactions

chemotherapy. This can save the patients from expenses, suffering, unhealthiness, and toxicities involved with these therapies. The surge in the formulation of drugs that can aim at a particular disease subtype or mutation are also witnessed. Along with drug discovery, associated diagnostic tests are also being designed that can help in the recognition of those patients who are most likely to respond or who are susceptible to severe side effects.

1.1.1 Different Types of Omics Data

Multiple types of omics data can be utilized to formulate an omics-based study or test. Next, some of them are discussed in detail. Relevant high-throughput technologies to profile omics-types are also presented in Table 1.1.

1. **Genomics:** The complete sequence of DNA in a cell or organism is called genome. This genetic content is present in cell nucleus and also inside other cell organelles, like mitochondria. There are certain experimental platforms, like whole genome sequencing techniques (DNA-seq) and SNP chips that are utilized for full or selective DNA sequencing. They can be utilized for new genomic assembly and also for the identification of novel genetic variants for a re-sequenced organism. Sequence Read Archive (SRA) (Kodama et al., 2011) and Gene Expression Omnibus (GEO) (Clough and Barrett, 2016) are the two databases that store publicly available genomic data. SRA houses raw sequence data whereas, GEO contains processed data along with metadata. National Center for Biotechnology Information (NCBI) hosts another public repository, that is dbGap (The database of genotype and phenotype) (Shendure and Ji, 2008). It contains individual-level genotype, sequence data, and phenotype with controlled access.
2. **Transcriptomics:** The complete set of RNA transcripts from DNA in a cell or tissue is called as transcriptome. It is comprised of mRNA (messengerRNA), miRNA (microRNA), rRNA (ribosomalRNA), tRNA (transferRNA), and other ncRNA (non-codingRNAs). Only about 1 to 2% of the human genome is reflected in transcriptome as protein-coding genes. The two dominating techniques for measuring transcriptome are Microarrays (Heller and J., 2002) and RNA sequencing (RNA-seq) (Wang et al., 2009). mRNAs quantification, identification of novel transcript, and discovering new splicing sites, all can be done using raw data (Weirick et al., 2015). The popular repositories containing transcriptional profiling data-sets are GEO (Clough and Barrett, 2016), SRA

(Kodama et al., 2011), and ArrayExpress (Kolesnikov et al., 2014).

3. **Proteomics:** All the proteins that gets expressed in cell, tissue, or an organism pools up to form proteome. Proteome has a complex intrinsic property as the proteins has to pass through several posttranslational alterations like phosphorylation, glycosylation, ubiquitylation, acetylation, and many other changes that affects the amino acids comprising the protein. Apart from this proteins have multiple physical forms and are located in different intracellular regions. They keep interacting with other proteins and other biomolecules. All these complex properties make it difficult and challenging to develop proteomics-based tests. Mass spectroscopy (CH et al., 2010) and protein microarray (Wolf-Yadlin and et al., 2009) are the two different techniques to quantify proteome. Successful protein arrays require recognition and delineation of capture agents because proteins do not have obvious complementary binding partners similar to RNA transcripts. The widely used repositories for storing proteome profiles are ProteomeXchange (Vizcaíno et al., 2014), ProteomicsDB (Wilhelm et al., 2014), and PRIDE (Jones et al., 2006).
4. **Epigenomics:** All the reversible chemical alternations to DNA or the histones that alters gene expression without modifying the base sequence are known as epigenome. The epigenomic alterations can take place in specific tissues, or in reaction to a certain environmental agent, or in diseased condition, and can exist across generations. Single organism has varying epigenome across cell types. The high-throughput biochemical measurement of epigenetic changes are mainly of two types: 1) Quantification of methylation of DNA cytosine residues at CpG sites, and 2) Quantification of multiple modifications that occurs to specific histone proteins in the chromosomes, also known as histone marks. Another technique to estimate the epigenetic changes in gene expression is RNA editing that is mainly quantified using the transcriptomic methods S. (2010)
5. **Metabolomics:** All the small molecules in the range of 50 to 1500 daltons present inside an organism constitutes metabolome. It includes all the intermediate metabolites in carbohydrate, lipid, amino acid, nucleic acid, and other biochemical pathways. It also includes hormones and other signaling molecules, as well as exogenous substances such as drugs and their metabolites. Metabolome has dynamic characteristic, it can differ within an individual organism and also among the organisms belonging to the same species. There are several regulating factors like eating habit, stress, mental and physical activities, pharmacological effects, and diseases. Mass spectroscopy (Weckwerth, 2003), and nuclear magnetic resonance (Zhang and et al., 2011) are the two techniques used for estimating the constituents of metabolome.
6. **Interactomics:** The inter-dependencies of biomolecules inside a cell can be mapped as interactome. In other words, interactome is a assembly of genome-wide interplay across genome, transcriptome, proteome, and metabolome. Further, multiple types of interactomes can be defined based on the type of interacting molecules. One of the highly studied interactome is protein-protein interactions (PPIs). PPIs can be estimated by yeast-two hybrid detection method Baker (2012) and lately by sequencing technology. One of the other widely studied molecular interaction is between DNA and protein. It can be examined using ChIP-Seq and ChIP-Exo techniques (Rhee and Pugh, 2011; Furey, 2012; Johnson et al., 2007). STRING (Szklarczyk et al., 2014) and BioGRID (Chatr-aryamontri et al., 2016) are two popular repositories that store information on molecular interactions.

1.1.2 Omics Profiling and Data Throughput

This section will present a high-level snapshot of methodologies that are used to operate high-throughput raw data generated from multiple platforms. For an overview, three important categories of high-throughput technologies, namely Microarray, NGS, and Mass spectroscopy will be discussed. The processing for each of the platforms can be segmented into three stages, having quality control stage of their own as presented in Figure 1.2. The first stage manages the raw data, most of the data processing is performed in the intermediate stage, and the last stage deals with post-processing and performing the final

Table 1.1. : Types of omics-data

Omics-type	Platform	Applicability	Resources
Genome	DNA-Seq (Shendure and Ji, 2008)	Genome assembly (Identification of genetic variants)	GEO (Clough and Barrett, 2016), SRA (Kodama et al., 2011), dbGaP (Mailman et al., 2007)
Transcriptome	RNA-Seq (Wang et al., 2009)	Transcriptome profiling (novel transcript discovery and novel splicing events)	GEO (Clough and Barrett, 2016), ArrayExpress (Kolesnikov et al., 2014) , SRA (Kodama et al., 2011)
Metabolome	Mass Spectroscopy (Bantscheff et al., 2012)	Proteome profiling (quantification of PTMs and Novel protein discovery)	PRIDE (Jones et al., 2006), ProteomeXchange (Vizcaino et al., 2014), ProteomicsDB (Wilhelm et al., 2014)
Proteome	Mass Spectroscopy (Dettmer et al., 2007)	Metabolome profiling (novel metabolite discovery)	MetaboLights (Kale et al., 2016)
Interactome	ChIP-SeqChIP (Furey, 2012)	Genome wide protein mapping (DNA interactions and gene-regulatory network)	GEO (Clough and Barrett, 2016), SRA (Kodama et al., 2011), hmChIP
	exo (Rhee and Pugh, 2011)	Histone modification maps and nucleosome maps	-
	Y2H (Baker, 2012)	Protein-protein interactions	STRING (Szklarczyk et al., 2014), BioGRID (Chattri-aryamontri et al., 2016)

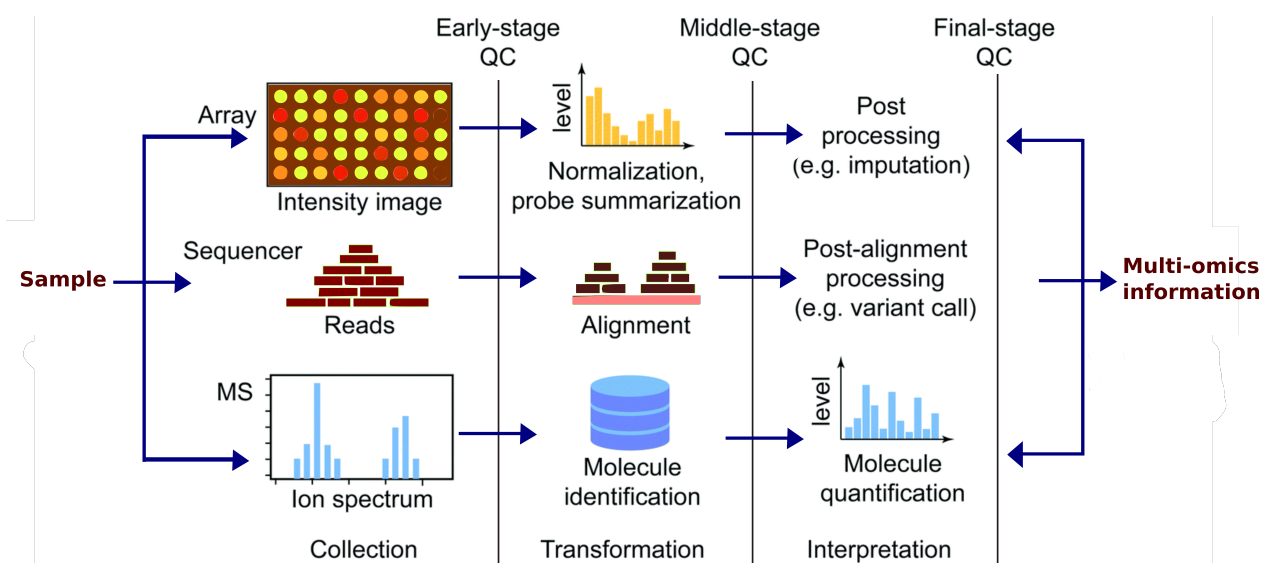


Figure 1.2. : Processing workflow for three main high-throughput technologies

estimation of molecular components.

1. **Microarray:** It is a technique based on the principles of hybridization. A small slide is used to examine huge amount of molecules (Leung YF, 2003; Heller and J., 2002). The hybridization approach provides multiple aspects of molecular profiling ranging from the estimation of genetic variants (DNA microarray) to transcripts quantification (cDNA microarray). The presence and abundance of molecules are quantified using intensity image. The raw intensity images gives the idea about the amount of hybridized molecules. Higher intensity of molecular hybridization is considered as that specific molecule exists in greater abundance. The intensity images are needed to be processed in multiple steps namely, background denoising, normalization, and probe-set summarization and so forth. Microarray technique for biomolecule quantification was introduced almost a couple of decades ago, since then, several pipelines for processing the raw data have been extensively developed (Ruskin, 2016; Quackenbush, 2002; Allison et al., 2006).
2. **Whole-genome sequencing (WGS):** WGS is a method to get complete insight of DNA/RNA of an organism at one single time. Next-generation sequencing (NGS) that has entered in the new era of genomics is the most promising recent advancement in the field of WGS. It has minimized the cost and time required to investigate the complete genome by characterizing short sequence reads in an immensely parallel mode (Metzker, 2010). The short sequence reads coming out of the sequencer as

raw data are usually aligned on a reference genome to identify short reads, then the post-alignment steps finalize the output. A wide range of omics information can be retrieved by applying different techniques at the post-alignment stage, like profiling of protein–DNA binding events for example, ChIP-Seq (Furey, 2012) and ChIP-exo (Rhee and Pugh, 2011).

3. **Mass-spectrometry (MS):** MS is a technique that utilizes the ionization property of chemical species to classify them based on their mass-to-charge ratio. This is a widely used technique for bio-molecular quantification especially for proteins and other metabolites. It is also used for creating modifications during the sample preparation step as in ^{13}C labeling that helps in the identification of metabolic fluxes (Yang, 2013). Unlike metabolites, first, the proteins are needed to be digested into small peptides using proteases so that they can be detected by the MS instruments. MS generates an ion spectrum that is used for molecular identification by matching the spectra from the other theoretical spectra deposited in databases (Han et al., 2008; Dettmer et al., 2007). For peptides, this step leads to the identification of peptide sequences. Finally, the target molecules are quantified based on the amount of identified small molecules.

1.2 IMPORTANCE OF COMPUTATION IN CANCER MULTI-OMICS RESEARCH

Integration of multi-omics data across different cellular function levels, including genomes, epigenomes, transcriptomes, proteomes, and metabolomes provide unmatched possibilities to interpret the inherent complexity associated with diseases, such as cancer. Cancer progression involves molecular modifications at different levels. With the advancement of affordable high-throughput techniques, lots of attempts are being made to untangle the mechanism behind cancer onset and progression (Alyass et al., 2015). The continuous reduction in the cost and time required for processing omics-based information has given rise to huge amount of data that has transformed the investigation approach from hypothesis-driven to data-driven. Nevertheless, investigation approaches that use single-level of omics information mostly lack in establishing the causal interactions between the molecular modification and its phenotypic reflection. In contrast, systems biology approaches help in holistic and systematic understanding of biological interactions (Yan et al., 2017). Fusion of the information coming from multiple regulatory layers can be specifically helpful in deciphering aberrant cellular functions behind diseases like cancer (Hasin et al., 2017). Analyzing cancer data at multi-omics level enhances the grouping of samples into biologically meaningful clusters, which in turn offers an improved understanding of prognostic and predictive phenotypes, elucidates cellular reactions towards a therapy, and aids translational research by integrative models (Mangul et al., 2019).

Although there are several advantages of integrating multi-omics data for cancer research, its entry to everyday clinical practice is prolonged. A major hindrance is the inconsistency in the maturation of different omics approaches. This leads to an increase in the gap between the volume of data being generated and the amount of data being processed. Here, the roles of computational techniques in multi-omics analysis is summarized, with a particular focus on the novel insight brought by multi-omics to cancer modeling. Computational techniques have four significant roles to play in multi-omics studies 1) Extraction of raw data from high-throughput instruments 2) Conversion of raw data into quantitative data 3) Managing the data storage in massive omics data repositories and 4) Implementation of analytical models. Next, these roles will be discussed in detail focusing on multi-omics in cancer research.

1.2.1 Multi-omics Data Extraction and Processing

The way in which multi-omics experiments are designed for generating high throughput data, dedicated bioinformatics workflow is required for their extraction and processing. With the increase in the complexity of these datasets and with the evolution of technologies, the challenges associated with them have also increased. Some of the examples of bioinformatics pipeline for multi-omics studies are Nextflow (Tommaso et al., 2017), Galaxy (Goecks et al., 2010), Snakemake (Köster and Rahmann, 2012), and other general purposes Common Workflow Languages (Chapman et al., 2016). The Bioconductor project (Huber

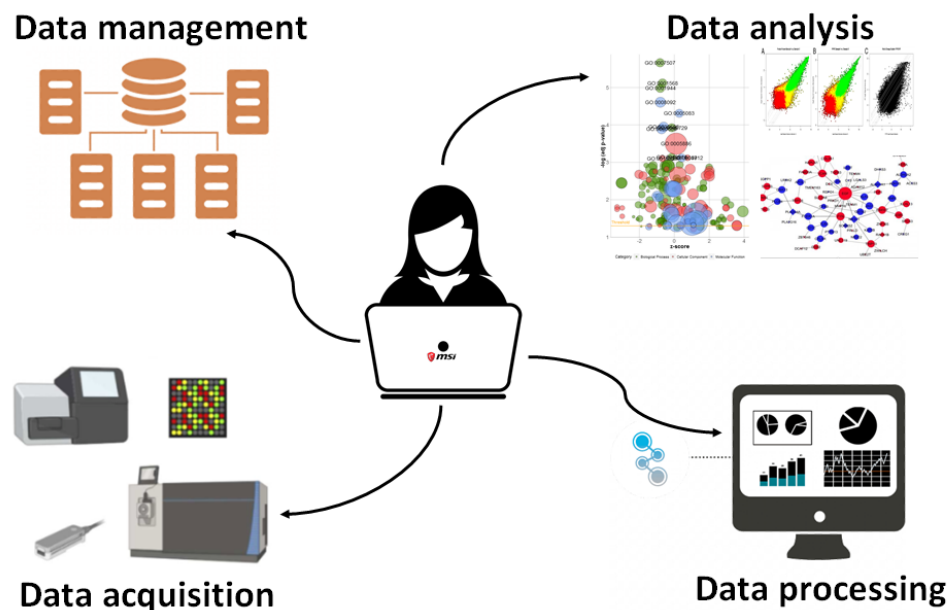


Figure 1.3. : Role of computation in multi-omics studies

et al., 2015) also contains a number of packages in R language for multi-omics data analysis, like MOVICS (Lu et al., 2020), OmicsARules (Chen et al., 2019a), mixOmics (F et al., 2017) CancerSubtypes (Xu et al., 2017), and others. It is a diverse open source repository for bioinformaticians. Biopython project (Cock et al., 2009) is another software ecosystem for the analysts in this field. Apart from these, there are huge amount of standalone computational programs and tools designed for various types of multi-omics data analysis, visualization, and interpretation, like MiBiOmics (Zoppi et al., 2021), MinOmics (Maes et al., 2018), Mergeomics (Ding et al., 2016), multiSLIDE (Ghosh et al., 2019), and others. Designing a suitable pipeline by developing novel algorithms or by choosing suitable combinations of existing tools and methods are very crucial for conducting multi-omics data analysis.

Transcriptomics Data Acquisition and Processing

Transcriptomic analyses are performed to quantify the amount of RNA present in a given physiological condition (Zhang et al., 2018a). Measuring gene expression is possibly the most common application of transcriptomic technologies. The biological state of a phenotype can be bar-coded using its gene expression profile. Further, it can be decoded via differential expression analyses to identify the cellular changes in cancers (Jeong et al., 2017). The gene expression measurements are an indirect approach to capture the critical regulatory changes for tumor progression. Measuring gene expression is technically and economically more feasible than measuring the protein content W. et al. (2003). Hence, the methods developed for estimating biological activities are designed by keeping transcriptomic data in mind (García-Campos et al., 2015). Exploring the molecular phenotype of cells using the transcriptome level of data has become an indispensable approach that can offer insight into the functional state of cells and the regulatory pathways, both in healthy as well as cancerous phenotypes (Li et al., 2015). Moreover, due to the lack of perfect correspondence between genes and proteins, need for multi-omics studies came into the picture.

Apart from gene expression, complete transcriptomic analyses also include the quantification of other non-coding RNAs, namely, miRNA, lncRNA, ncRNA, enhancer RNAs, and others. Transcript measurement can be performed using either expression microarrays or RNA-sequencing (Sultan et al., 2008; Mortazavi et al., 2008). Both of these technologies have distinctive technical considerations, but

follows some common steps for their analyses, like data acquisition and preprocessing; removal of technical artifacts; quality check and data normalization. The resulting data can be represented as an expression matrix: an $M \times N$ matrix where rows represent transcripts and columns represent samples (or observations). The omics experiments designed for cancer studies are primarily performed using tailor-made analytical pipelines.

Epigenomics Data Acquisition and Processing

Advancement in high-throughput omic technologies has enabled the researchers to examine the tumor genomes for chemical modifications. Occurrence of such epigenomic modifications have a noticeable effect on many cancers, and they are found to play critical roles in tumor progression. These modifications comprise DNA methylation and histone modifications occurring in either oncogenes or other cancer-related genes. Mutations in genes involved in epigenetic regulation have also been found in several tumor types. Investigation of epigenomic data using computational techniques may cater to some new perspective regarding initiation and progression of cancer. Studies based on computational analysis may facilitate the identification of genetic and epigenetic biomarkers for diagnosis and prognosis and thereby may speed up the advancement of new therapeutic approaches.

Multiple mechanisms are involved in regulating the epigenomic changes therefore, different types of high throughput molecular techniques are required for their quantification (BE et al., 2007). That, in turn, requires a dedicated set of computational methods or tools for handling data generated from each of these techniques. For instance, in the case of DNA methylation, the methylation state of a DNA region can be quantified using either array-based methods or sequencing methods, such as the popular whole-genome bisulfite sequencing (WGBS) (Fan and Chi, 2016). Data from array-based methods can be processed using dedicated packages, such as methylationArrayAnalysis (Maksimovic J, 2016). Similarly, data coming from sequence-based methods can utilize dedicated tools, such as the bsseq (Hansen et al., 2012) or methyAnalysis packages (Du and Bourgon, 2021). Apart from DNA methylation other epigenomic mechanism includes chromatin remodeling and chromosome conformation that can also be measured using other high-throughput techniques. A high throughput NGS technique like ATAC-seq and DNase-seq can be utilized for measuring the chromatin accessibility. Studies have shown an association between higher-order chromatin rearrangements and chromosomal modifications in cancer. Recently, chromatin conformation capture has allowed the identification of putative pharmacological targets in breast cancer (Baxter et al., 2018).

Proteomics Data Acquisition and Processing

Studying cancer at the proteomic level is propitious as it can identify biomarkers and therapeutic targets (Yakkioui et al., 2017). It can provide deeper insight into disease progression as the proteins are the molecular unit of cellular structure and function. Compared to the genomics and transcriptomics technologies, high throughput proteomics technologies have developed slower. Microarray-based technique for estimating the protein content in particular biological condition have several applications and shown a varied level of success (Sutandy et al., 2013; Atak et al., 2016). However, the more significant breakthroughs have come through the use of mass spectrometry (Cho, 2017). Cancer studies at the proteomic level require data analysis at several steps. At the data acquisition step, the molecular fragments detected by the instrument have to be identified (Schmidt et al., 2014). A general procedure to do this is by comparing the fragments to databases in real-time (Graumann, 2012). Next, the identified poly-peptide chains have to be assembled, that requires another group of computational approaches (Hoopmann and Moritz, 2013). Development of computational tools for these purposes is an active domain for research. Bioconductor provides an efficient group of tools for managing proteomics data, from data processing to functional analysis (Kopczynski et al., 2017; Gatto and Christoforou, 2014). Another good substitute for protein estimation is the Maxquant toolset (Cox and Mann, 2008).

1.2.2 Multi-omics Data Management

Since the conception of the Human Genome Project (HGP), a thrust for open data in the field of biomedical genomics has been observed. This has led to the emergence of a rich Genomic Commons (Contreras and Knoppers, 2018). For smooth and rapid scientific discoveries, it is essential to make these data available in public repositories. Although, there are several legal, ethical, and technological challenges associated with this (Cook-Deegan and McGuire, 2017). Difficulties in handling multi-omics data involve defining the type of data to be stored and how to store it; the policies for data access, sharing, and re-use; and long-term archiving policies. Perhaps, the richest and the most successful repository for multi-omics cancer data is NIH Genome Data Commons (GDC). It is comprised of all the data generated from TCGA project, but all of these data are not publicly available. The data is structured into a directed graph consisting of interrelated entities as presented in Figure 1.4 Every entity has a specific characteristic and connections. There are several ways to access the TCGA data available in public domain, one is through their Application Programming Interface that requires acquaintance with command-line tools. Though the project's data portal supports easy to use interfaces, they have limited application for detailed analyses. Another way to access TCGA data is through some devoted packages like RTCGA (Kosinski and Biecek, 2021) and TCGAbiolinks (Colaprico et al., 2015). Moreover, a more extensive collection of datasets can be accessed through the Broad Institute's Firehose (Deng et al., 2017). Also, cloud computing-enabled data access is provided through the Cancer Genome Collaboratory (Yung et al., 2017). Some subscription-based commercial softwares are also available for accessing TCGA data, like Qiagen's OncoLand database and the cloud-based analytics solution, Seven Bridges. Apart from being paid services, they are not customizable, and one cannot perform the desired analysis. Possibly, the best amalgamation of versatility and usability available in the TCGAWorkflow suite available as an R/Bioconductor package (Silva et al., 2016).

TCGA has an inarguable impact on the cutting-edge multi-omics research. It provides publicly accessible data that can be utilized for method development and validation. In this thesis, the publicly available multi-omics cancer data from the TCGA project are utilized to benchmark novel computational approaches designed for the identification of miRNA-mRNA regulatory modules in cancer and for the prediction of cancer subtypes. At present, there are lots of projects that utilizes TCGA data for cancer studies. Several wet-lab researchers are involved in projects that contribute to cancer multi-omics community. This provides the opportunity to integrate the data generated from the local experimental projects and large collaborative endeavors like TCGA. This kind of research practice allows contrasting specific hypothesis for the different research groups with the statistical power obtained via the much larger datasets generated by international multicentric collaborative projects. There is great scope of knowledge discovery from comprehensive and systematic re-analysis of the data available in large public repositories. A list of data repositories available to multi-omics cancer data is presented in Table 1.2.

Table 1.2. : Cancer Multi-omics Data Repositories

Repository	Available multi-omics data	Availability	Cancer types
TCGA	DNA-seq, RNA-seq, miRNA-seq, RPPA, DNA methylation, SNV, and CNV	https://cancergenome.nih.gov/	Multiple cancer
CPTAC	Proteomics data related to TCGA cohort	https://cptac-data-portal.georgetown.edu/cptacPublic/	Multiple cancer
ICGC	WGS, genomic variations	https://dcc.icgc.org/	Multiple cancer
CCLE	Gene expression, CNV, sequencing data, pharmacological profiles of 20 anticancer drugs	https://sites.broadinstitute.org/ccle	Multiple cancer cell lines
METBARIC	Gene expression, SNP, CNV, and clinical traits	http://molonc.bccrc.ca/aparicio-lab/research/metabarc/	Breast cancer
TARGET	Gene and miRNA expression, CNV, and sequence data	https://ocg.cancer.gov/programs/target	Pediatric cancer
Omics discovery index	Genomics, proteomics, transcriptomics, and metabolomics	https://www.omicsdi.org/	Consolidated data sets from 11 repositories

1.2.3 Multi-omics Data Analysis

One of the commonly asked questions is why it is required to integrate multi-omics information utilizing intricate mathematical models. The simplest and most probable answer to this is because the biological processes are not regulated by independent biological entities at an individual omic layer and integrating the multiple layers will help in better representation of the system under study. With the availability of multi-omics technologies, the researchers are continuously trying to get a better picture

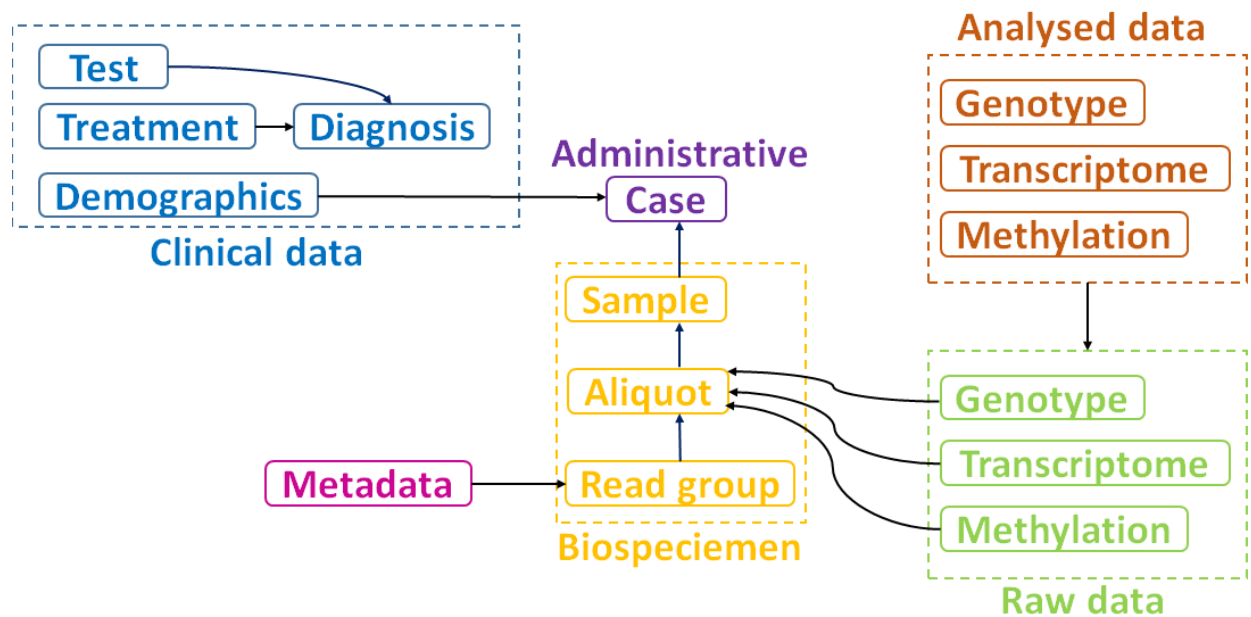


Figure 1.4. : A simplified illustration of data structure used in for TCGA data storage within GDC

of biological phenomena. One of the most important goals of biomedical research is to enhance the present cancer diagnostic capacities, and advancement in multi-omics integration tools have long been acknowledged for that (Parkinson et al., 2012). Multi-omics integration studies have the potential to offer better predictive tools than individual molecular technologies, as the individual technology is only capable of harnessing just a piece of the complete complex pathological system. Multi-omics data integration can add significant value to both basic and clinical research. Selection of an optimum omics combination is a trivial task, as there are many practical limitations (technical and economic) in the clinical setup, where the diagnostic tool has to be implemented (Yoo et al., 2018). Bioinformatic strategies and computational tools play a crucial role in designing such studies.

There are numerous ways to integrate information coming from different omics layer based on the study design (Palsson and Zengler, 2010). The most commonly used approach includes co-mapping or correlation. It suggests that if there is some common driver between two of the omic views or perturbation in one affects the other than there are chances that they are associated or correlated. Several tailor-made statistical methods that are primarily dependent on conditioning have also been developed for the integration purpose. The statistical models are designed to identify whether a single biological entity is independently affecting a diseased condition or one is the function of the other. For example, "mediation analysis": a regression-based method was designed to merge gene expression and SNP data considering gene expression as the facilitator in the causal mechanism from SNPs to disease (Huang et al., 2014; Sun and Hu, 2016). Approaches like these have also been applied for integrating other omic views (Sun and Hu, 2016; Yugi et al., 2016). Modeling multi-omics information as networks is another commonly used approach. It is a data-driven approach and is supported by the knowledge of molecular networks. An important biological concept suggests that genes and their products together form a complicated, and interconnected networks, rather than following a liner path Collins et al. (2003). Therefore, one of the best approach to model, them is as graphs, where the interconnected elements represent different biological entities (laszlo barasasi and bonabeau, 2003; Huan et al., 2013; Zhang et al., 2013a; Subramaniam and Nadeau, 2010). Several studies are conducted related to metabolic pathways modeling using graphs, where the metabolites are treated as nodes, and the edges represent enzymatic conversions (Jeong et al., 2000; Ravasz et al., 2002). Inspired by

these graphs, co-expressed genes are also modeled considering a series of modifications, where the nodes represent the genes that encode transcript and the edges represent the correlation between them (Ghazalpour et al., 2006; Langfelder and Horvath, 2007; EC et al., 2010). For studying a diseased condition, networks comprising of co-expressed genes can be constructed based on fluctuation in gene expression between normal and affected individuals separately (EC et al., 2010; Emilsson et al., 2008; Romanoski et al., 2011). Comparing the network structure between the normal and diseased condition helps in deciphering the entities closely correlated with the diseased state. Further, these networks can be experimentally validated, however the number of predicted interactions are so big that makes it difficult for testing. For instance, Marbach and group integrated transcriptomics, genomics, and epigenomics information to explore tissue-specific regulatory circuits in 394 human cell types (Marbach et al., 2016). They also mapped the results obtained by disease-specific genome-wide association studies onto the tissue-specific regulatory networks and discovered some regulatory modules enriched with genetic variants associated with that disease. All these studies illustrate that integrating the multi-omics data using mathematical models help in narrowing down the focus of disease research into specific functional aspects of particular cell types or tissues, considerably facilitating downstream mechanistic efforts and hypothesis generation.

Multi-omics data analysis can be looked from a data-science perspective, contemplating that every method is fundamentally solving a ML task. The mapping is tried to be presented with the help of Figure 1.5. The categories shown here are not strict, most of the time, the results from one of the analysis is fed to others for a complete exploration of biological knowledge. For instance, a supervised classifier can be based on the analysis performed by an unsupervised clustering algorithm having its application in disease diagnosis and prognosis. Considering the precision medicine framework, ML has been widely used for investigating and interpreting interactions between different types of omics data with the development of diagnostic, prognostic and predictive tools (Dias-Audibert et al., 2020; Mamoshina et al., 2018; Sonsare and Gunavathi, 2019). Though multi-omics analysis using ML is at its embryonic stage, a wide range of applications has already been explored for cancer studies.

Understanding the mechanistic details of disease biology lies central to diagnosis and developing novel interventions for the disease. This thesis deals with the development of computational approaches that are designed to decipher biologically significant information from multi-omics cancer data in an unsupervised manner. The methods developed here derive inference from input datasets without labeled response variables. There are several approaches for integrating multi-omics data in an unsupervised manner. A categorization is presented in Figure 1.6. Further analysis can be performed on this integrated information as per the biological question under consideration. Herein, the multi-omics data integration is performed to address some challenges related to cancer and its mechanism. The biological questions that are dealt with in this study are:

1. Prediction of miRNA-mRNA regulatory modules in cancer.
2. Cancer subtyping or identification of homogeneous groups of cancer patients.
3. Deriving insights into cancer biology.

A detailed mathematical and algorithmic insight into the available tools/methods for multi-omics data integration and further analysis aligning with the scope of this thesis is presented in Chapter 2.

1.3 CHALLENGES IN MULTI-OMICS STUDIES

Analyzing omics at the individual level (single omic) exhibits a multitude of challenges (Gomez-Cabrero et al., 2014). Inclusion of multi-omics for comprehensive studies inherit the challenges of single omics, and confounds the study with other new challenges like fusion/integration, clustering, visualization, and functional depiction (Pinu et al., 2019; Jamil et al., 2020). For example, before the

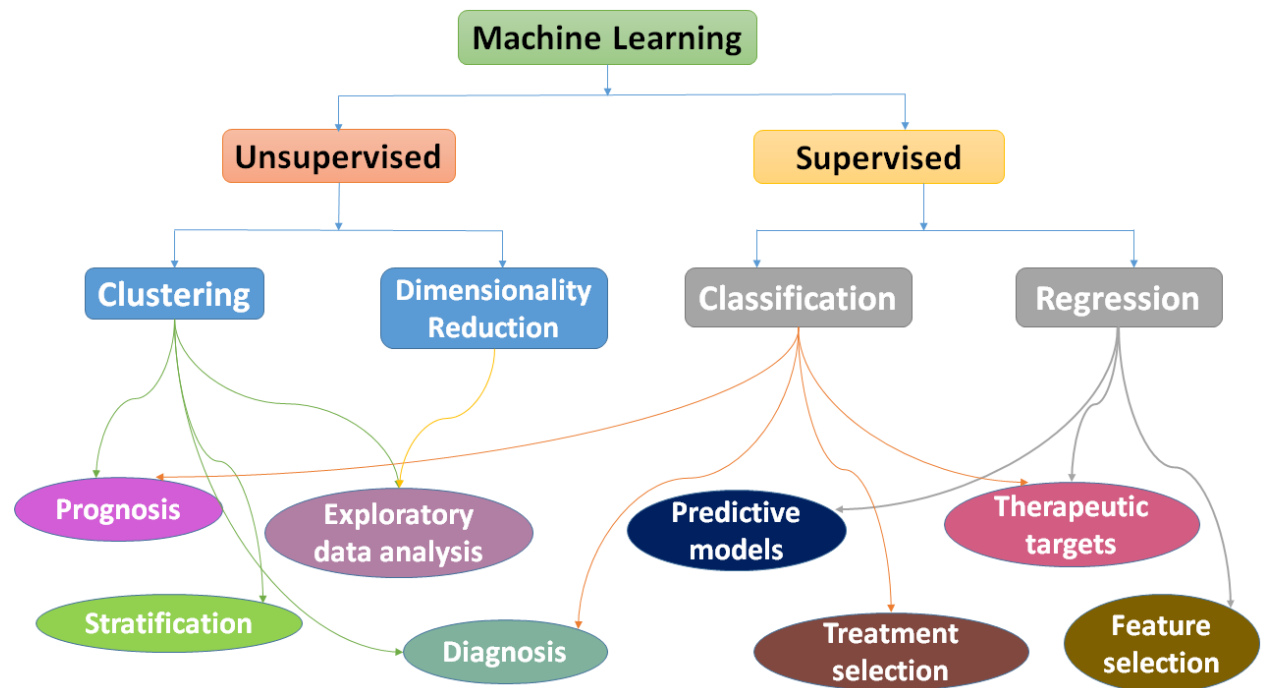


Figure 1.5. : Application of machine learning in multi-omics cancer studies

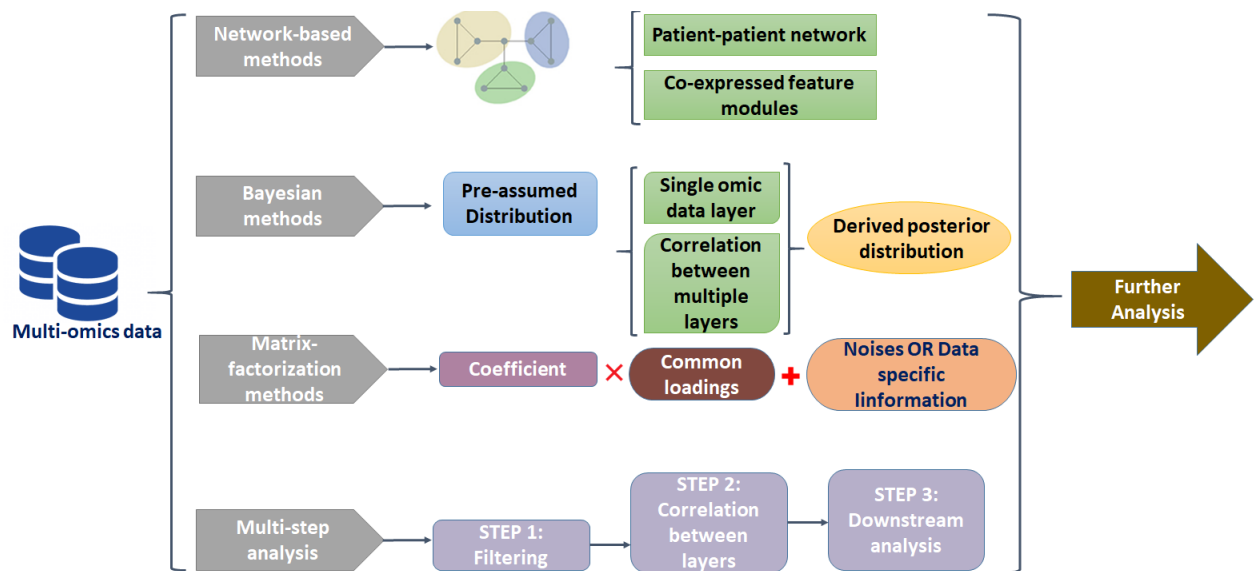


Figure 1.6. : Unsupervised approaches for multi-omics data integration

integration of two or more omics, the analyst faces the hurdles of data standardization like scaling and normalization. Further, integration of large multi-omics data-sets brings the curse of dimensionality and the constraints imposed by it. For some studies, demand for storage space and the computational load becomes a limiting factor. Sometimes, even mapping of identifiers (IDs) that is a prerequisite for multi-omics integration becomes a daunting task. For example, most of the time there is no one-to-one correspondence between genes and their associated transcripts or proteins. Also, annotating the omic entities (transcripts, proteins, and other metabolites) with extra information for the enrichment purpose is a cumbersome task. Considering the challenges faced during statistical analysis, the unsupervised approach for the multi-omics study can intensify any information like systematic batch effect that can be acquired during the process of sample collection, transporting, or processing. If the analyst fails to rectify or remove such technical artifacts coming from unwanted sources it may possibly misguide during the process of multi-omics integration and further downstream analyses. All these challenges can be summarized as follows:

1. **Data Wrangling:** The affect is also sometimes referred as "data munging". It includes the process of "transformation" and "mapping" at multiple levels of multi-omics data analysis. Data transformation is achieved by normalization, scaling, pruning, and imputation that helps to harmonize omics data coming from different platforms. In the category of "mapping" comes the daunting task of harmonizing IDs (entities) across the omics. It requires annotating data across meta-data, which is a labor-intensive process and needs enormous one-to-one or one-to-many relationship operations. Mindful enrollment of samples and strong metadata documenting tables, with participation of data generation and analysis teams can help to handle this challenge and minimize errors.
2. **Data Heterogeneity:** The multi-omics datasets are produced through various technologies and platforms that leads to data heterogeneity which is another obstacle while dealing with them. For instance, consider the scenario were the analyst has to deal with the data coming from some sequencing platform and mass-spectroscopy or microarray versus mass spectroscopy scenario, the pre-processing steps addressing the problems of individual datasets may not suit for others, especially when all of them are supposed to be used under a single framework. Moreover, some tools have helped in dealing with this problem very efficiently, for example Similarity network fusion (SNF) Bo et al. (2014), Multi-omics factor analysis (MOFA) Argelaguet et al. (2018), mixOmics F et al. (2017) and many others. These tools are based on a wide variety of mathematical and statistical approaches like canonical correlation, factor analysis, matrix factorization, graph theories, and so forth. They are used for the downstream analysis of multi-omics data for the usage like phenotypic predictions.
3. **Data representation and dimensionality reduction:** Most of the time, after the preprocessing step, data representation is performed to identify the variations coming from technical sources like the batch effects and other evident biological patterns at each level of analysis. The representation step may involve feature identification, selection, or extraction. This practice facilitates the learning of biological knowledge coming from the data in an unbiased manner. For instance, suppose an analyst has to design an algorithm to predict biomarkers responsible for a certain kind of tumor progression. For doing so, the analyst has to depend on patient data solely like expression profiles of mRNAs, miRNAs or other biomarkers coming from multiple genomic platforms. Here, every feature in the dataset describes the characteristic of the patient and the algorithm needs to predict the outcome by learning the relationship or inter-dependencies present within the features. Therefore in such scenarios presence of any noise or discrepancies in the data can totally manipulate the result, and the prediction can go wrong. The solution to this problem is to provide the algorithm with more abstract representation of the data. It means either select the most suitable set of features (feature selection) or map the data into better and lower-dimensional feature space (feature extraction) before providing it to the prediction algorithm. It helps in dimensionality reduction without losing important or relevant information. The redundant information is removed from the dataset and the algorithm's efforts in building variable combinations (features) facilitates in speeding up the learning and generalization steps in the prediction process. The methods like Principal component analysis (PCA)(Jolliffe and

Cadima, 2016), t-distributed stochastic neighbor embedding (tSNE) (Kobak and Berens, 2019), and uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) are most commonly used techniques for dimensionality reduction to explore both the linear and non-linear relationships available in the dataset. Moreover, this step is often challenging, given the complexity of the analytical space and the study goals due to latent patterns encoded in the input samples originating from different omics layers, technologies and platforms. There is a scarcity of mathematical and statistical models/algorithms that can optimally tune the representational complexities of multi-omics data. Investigation of noisy and high-dimensional data often leads to wrong interpretations, which is a common problem in multi-omics data analysis provided the heterogeneity present in every layer of the multi-omics data.

4. **Benchmarking limitations:** With the increasing dependence on computational tools, comes the need for their systematic evaluation (benchmarking) (Charles et al., 2018). The major challenges for benchmarking computational tools utilizing omics data are 1) availability of "gold-standard" datasets, 2) surge of new methods, 3) assuring the reproducibility in the context of increasing complexity of the software involved. All of these challenges gets amplified in the multi-omics setup as the availability of sample matched measurements becomes very difficult. On the top of that every novel method depends on a unique combination of omics, limiting the chances for employing the known benchmarks. The researchers tries to improve the quality of multi-omics tools by combining information retrieved from multiple single-omics tools, this adds on the complexity and benchmarking becomes even more difficult.

TCGA is one the the widely used sources for benchmarking multi-omics methods. It includes genomic, epigenomic, transcriptomic, proteomic, and clinical data for 33 cancers. Very limited amount of noteworthy multi-omics benchmarks are accessible for multi-omics dimensionality reduction algorithms Cantini et al. (2020), multi-view clustering algorithms Rappoport and Shamir (2018), and survival prediction algorithms Herrmann et al. (2020). All of these benchmarks are conducted utilizing TCGA cancer data. The integration approach cannot be applied on other multi-omics data like, microbiome or metabolome as they are not covered in TCGA repository. A lack in comprehensive comparison between existing tools is also not possible because of non-availability of standard datasets. Apart from widely used TCGA datasets there are some studies that used simulated dataset, these datasets are also derived from data dependence structure from the TCGA cancer studies (Chalise et al., 2018). Because of the constraints in the systematic characterization of multi-omic tools, methods, and algorithms, researchers need to choose tools that are either well benchmarked in appropriate scenarios and/or evidenced in multiple observational studies and systemically evaluated.

5. **Fair sharing:** Apart from the challenges discussed above, there is one another problem in the field of multi-omics data analysis, which is "fair sharing". Huge amount of multi-omics data are generated that can be integrated using various tools/methods/algorithms. This task can be computationally intensive. Also, a massive amount of processed data are generated that is needed to be taken forward for exploration, inferencing, and interpretations. Sharing both the data and code alongside compute environment allows interoperability and non-reinventing the wheel. Requirements of open and fair sharing of various components of a multi-omics study is presented with the help of the flow diagram in Figure 1.7.

1.4 THESIS STATEMENT

Cancer is by now widely accepted to be the quintessential complex disease. An in-depth understanding of its pathological phenotype can only be achieved by adequately integrating the myriad of interconnected biological elements (de Anda-Jáuregui et al., 2014). Cancer genomics programs like TCGA have generated a huge amount of multi-omics data for multiple cancer types that can be utilized for the analysis. Therefore, this thesis aims to develop computational methods and frameworks that utilize more

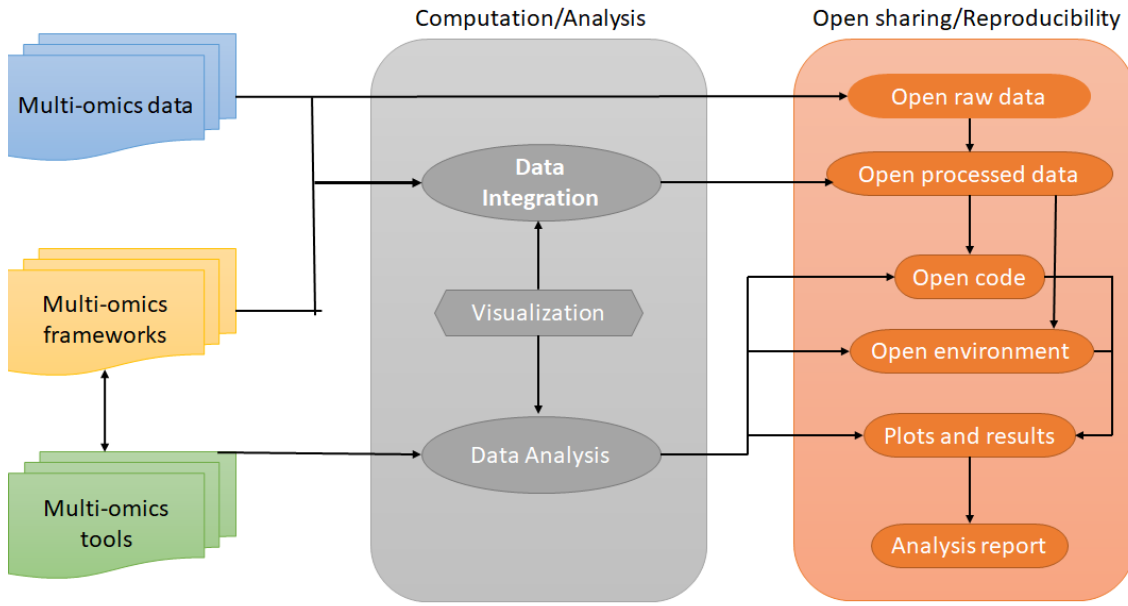


Figure 1.7. : Workflow for fair and open sharing of information in a multi-omics study

than one omic data at a time to understand the etiology of cancer. Two major objectives of this thesis are: 1) To identify miRNA-mRNA regulatory modules (MMRMs) in cancer and 2) To identify cancer subtypes. The computational objective associated with these biological objectives are 1) Simultaneous clustering of miRNAs and mRNAs (clustering of features) whose expression are measured across same set of cancer patients and 2) Clustering of cancer patients (clustering of samples) by integrating multiple levels of biological information coming from different omic platforms. The novel computational methods and frameworks designed in his study explores four different approaches of ML, namely unsupervised clustering, feature representation, feature selection, and multi-view learning (Xu and Jackson, 2019). The methods/algorithms/framework proposed in his study try to overcome the existing limitations of the methods available for M/FMRMs identification and cancer subtypes identification. The proposed methods outperforms other popular methods in this domain. Figure 1.8 presents the overall workflow of the study presented in this thesis. The different layers of the workflow presents the multiple aspects of this study.

1.5 ORGANIZATION AND SCOPE OF THIS THESIS

The study presented here addresses the requirement of computational approaches for multi-omics data analysis. Some novel clustering, feature selection, and multi-omics integration approaches have been designed for the identification of miRNA-mRNA regulatory modules in cancer and cancer subtypes discovery. The outline of the proposed thesis is presented in Figure 1.9 The thesis contains seven chapters. Chapter 1 introduces the work by emphasizing multi-omics data and the need for computational approaches for their analysis. It also emphasizes the roles of multi-omics studies in cancer research. The thesis objectives and outcomes are also presented here. Chapter 2 presents the literature survey. The recent developments and shortcomings in computational approaches for multi-omics studies in cancer research are discussed. The chapter focuses on a detailed survey of the computational methods available for M/FMRMs identification and cancer subtype discovery that are the two biological questions dealt with in this thesis from the multi-omics perspective.

Referring to the first objective, an algorithm named as *RFCM*³, is presented in Chapter 3. *RFCM*³ integrates miRNA and mRNA expression data for the identification of condition-specific miRNA-mRNA

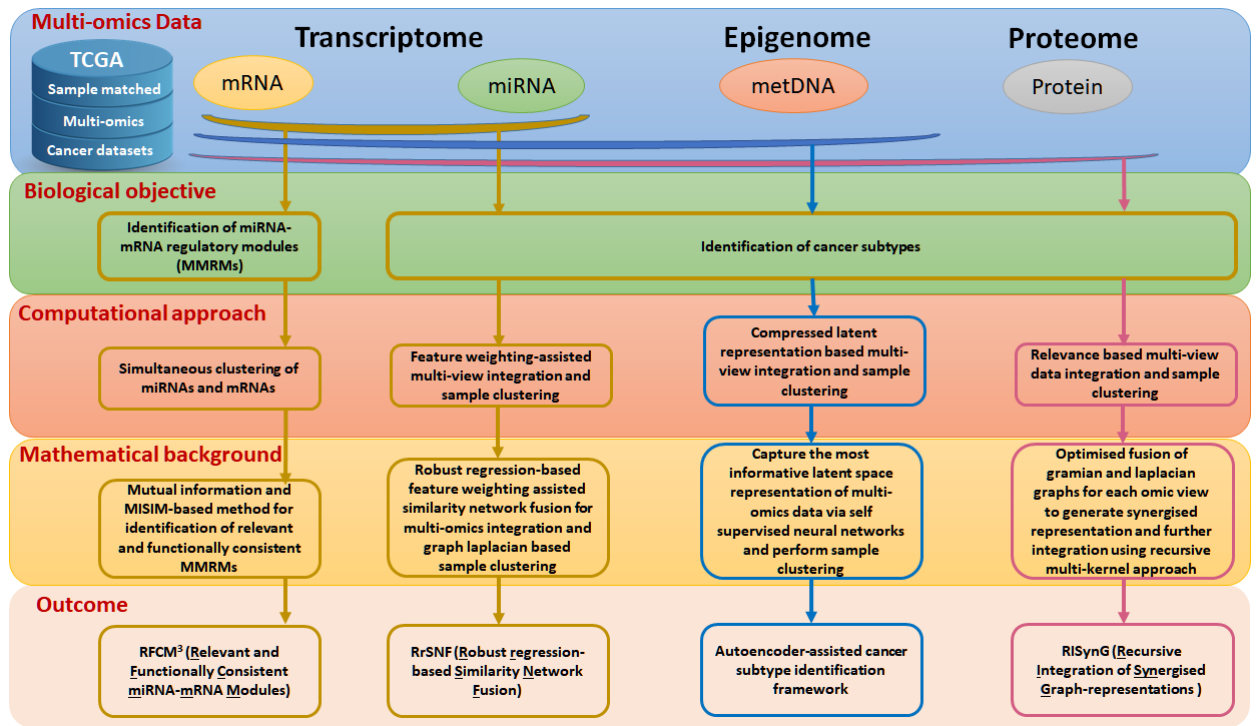


Figure 1.8. : The overall study presented in this thesis

regulatory modules. The method is designed to identify FMRMs in Cervical cancer. Cervical cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death in females all around the world. Therefore, exploring the molecular mechanisms related to tumor progression of cervical cancer is very important and might provide novel biomarkers to predict and improve the prognosis of patients, which ultimately would lead to better therapeutics. Identification of FMRMs will aid in deciphering the aberrant transcriptional regulatory networks and will help to understand the role of already known biomarkers like miRNAs and mRNA in various pathways involved in cervical cancer. Although, a few algorithms for finding FMRMs have been proposed, improvements are still needed. In this respect, unlike other methods *RFCM³* does not take the help of sequence-based interaction information from sequence-based databases to avoid bias. It uses miRNA-disease association information to identify functional similarities between the interacting miRNAs.

Cancer is a disease characterized by uncontrolled cell growth having a potential to develop in various cell types and tissues. The complexity of the disease increases even more because of the presence of multiple subtypes within one type. These subtypes are usually distinguished on the basis of variable histology, specific mutations, and molecular profiles. These variabilities further lead to different clinical outcomes. Identification of cancer subtypes can help in the segregation of cancer patients into separate groups with similar clinical phenotypes, prognosis or response to treatment. In recent years, advancement in multi-omics technologies have provided key insight into defining cancer subtypes within a tissue. However, there is still a lack of methods that can help in identification of subtypes present within cancer patients. This type of grouping can help to explore the variation present in the molecular profiles that are common or different across tissues and a deeper insight into it can further help in the development of broadly applicable and highly versatile therapies. In this respect, novel pipelines and algorithms have been designed in this thesis work to judiciously integrate multi-omics data for cancer subtypes identification.

The first work towards the goal of cancer subtypes identification is presented in Chapter 4. Here two transcriptome level of data that is miRNA and mRNA expression are utilized to group cancer patients. The

study introduces a novel method of feature (miRNA and mRNA) weighting based on robust regression fit. Weighting of features is important from biological perspective. Different metabolic pathways are triggered in a cell under different biological conditions. All the mRNAs and miRNAs in the cell do not have equal role in activation of these pathways. The biomarkers have differential expression patterns under different biological processes. Therefore, it is very important to consider the differential importance of every feature while clustering. The proposed weighting approach improves the clustering solutions. The method calculates the importance of every biomarker and assigns a weight to it, higher the weight means more important is the feature. Calculated weights of each biomarker are used to compute similarity between samples. Later, the samples similarity are used to build the SNF network Bo et al. (2014) that is clustered into respective subtypes using spectral clustering Ng et al. (2002).

The second work towards this goal is presented in Chapter 5. In this study, a novel algorithm named RISynG is presented. The proposed approach treats multi-omics data clustering as multi-view clustering, where information from multiple omics platforms are integrated to identify clinically important sub-groups within a cancer. In order to judiciously capture the variation present across the multi-omics dataset, the proposed approach works in three steps. At first, for each view, two sample similarity matrices are computed using graph representation matrices, namely, Gramian matrix and Laplacian matrix. This step acknowledges the statistical diversity in the multi-view omics data, which directly influences the quantification of similarity between the samples. Later, it involves the integration of representation matrices for the respective omic-view using a parameterized combination function to generate synergy matrices. In the second step, the variation captured through synergy matrices for each of the omic-view are fused. The proposed approach first arranges all the synergy matrices based on their relevance. Then, a recursive function is designed to merge each of the synergy matrices in a manner that less relevant matrix has only a slight influence on the final cluster structures. At last, k -means clustering is applied to the integrated, most relevant subspace to obtain cancer subtypes.

The third and last work towards this goal is presented in Chapter 6. In this study, the deep learning framework of Autoencoder architecture is utilized to capture the best latent representation of integrated multi-omics data that can be utilized for cluster identification. A novel AE-assisted workflow is designed to integrate the sample matched transcriptomic and epigenomic data for cancer subtypes identification. The proposed framework first performs a supervised feature selection based on the survival status of the patients. The selected features from each of the omic data are concatenated and passed to the AE. The information embedded in the latent space of the trained AE neural networks are then used for cancer subtyping using Spectral clustering. The AE architecture designed in this study exhaustively searches for the best compression for multi-omics data by varying the number of neurons in the hidden layers and penalizing activations within the layers. Clustering performed on this latent representation identifies compact and clinically significant patients clusters having significant differences in their survival pattern.

All the computational approaches developed in this thesis are benchmarked using multiple types of multi-omics cancer datasets taken from the TCGA database. Also, comparative analysis with the other popular and relevant methods in this domain is performed to demonstrate the robustness of the proposed approaches. Several metrics are calculated to quantify the results generated by different computational methods. Detailed biological analysis of the results is also provided to demonstrate the biological relevance of the results obtained from proposed mathematical models. All the scripts and codes related to the study are made publicly available on the author's website and Github pages. At last, the concluding remarks for this thesis work are presented in chapter 7. Some future research directions are also provided here.

...

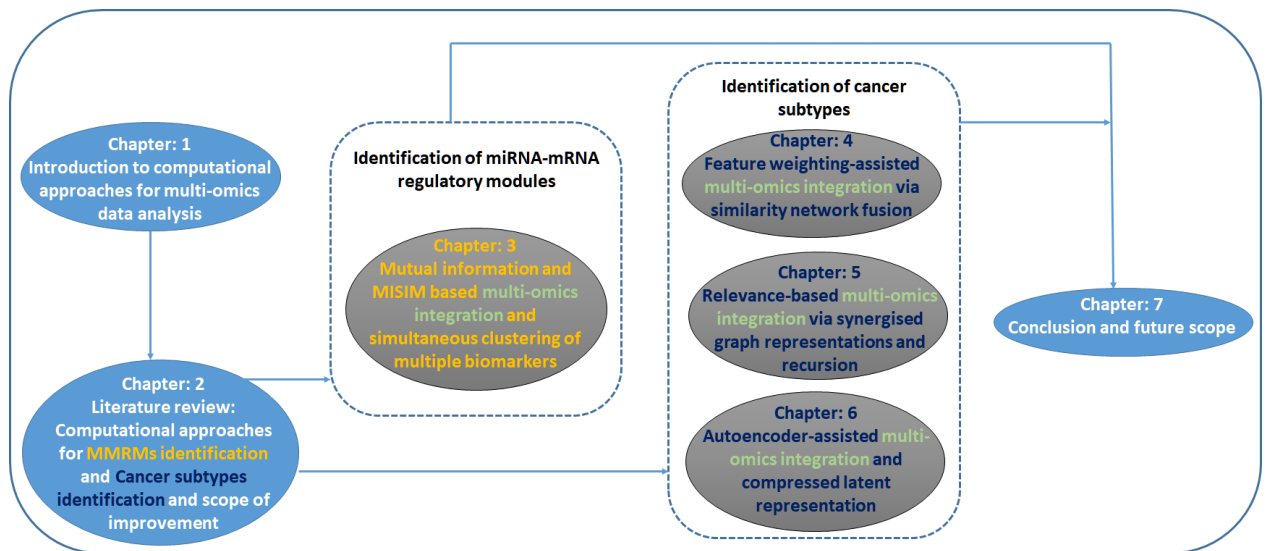


Figure 1.9. : Outline of the thesis

