Chapter 2

Background

2.1 Gene Expression Data Analysis

Gene expression data analysis can be used to gain insight into the expressions of genes in various biological systems. By using these techniques, researchers are able to understand the roles of various genes in various processes, and can uncover new therapeutic targets for various diseases.

2.1.1 Gene and Gene Expression

Gene is a unit of genetic material that is made up of DNA or sometimes RNA and is passed from one generation to the next. The structure of a gene is shown in Figure 2-1. It consists of a sequence of nucleotides that code for specific proteins, enzymes, and other molecules that carry out specific functions in the organism. Genes are located in the nucleus of each cell, and they are made up of both exons and introns. Exons are the coding regions of the gene that are responsible for the expression of the gene product. Introns are non-coding regions of the gene that are spliced out of the mRNA molecule before it is translated into a protein. The human genome is made up of about 20,000 genes, each of which is responsible for a specific function in the body. Genes are responsible for the traits that we inherit from our parents, such as eye color, height, and intelligence. In addition, genes can also play a role in the development of diseases, such as cancer and diabetes. Scientists are still studying the role that genes play in the development of diseases, and they hope to use this information to better understand and treat diseases in the future.

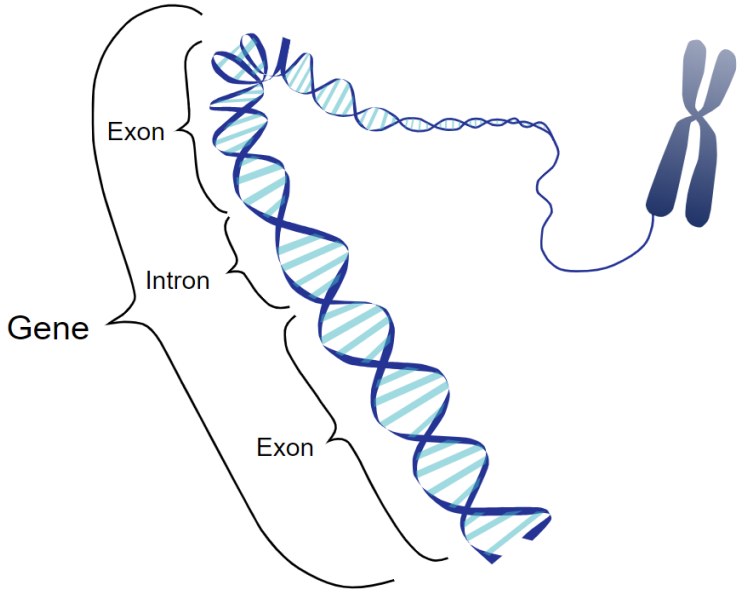

Figure 2-1: Schematic representation of a Gene. Source: File:Gene.png. In Wikipedia. https://commons.wikimedia.org/wiki/File:Gene.png, accessed on 21/07/2023

The process of utilising information from a gene in order to generate a functional gene product is known as gene expression shown in Figure 2-2. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA) and transfer RNA (tRNA) genes, the product is a functional RNA. The primary way gene expression is regulated through the use of transcription factors. These are proteins that bind to specific sequences of DNA and affect the activity of the gene. They can either increase or decrease the amount of mRNA produced from a gene, thus controlling the amount of protein that is produced. Transcription factors (TFs) can also control which proteins are produced by regulating the type of mRNA that is produced. Gene expression is essential for the development and functioning of all organisms, as it enables cells to respond to changes in the environment and to make decisions about how to respond. In addition, gene expression is a key factor in the development, progression, and outcome of diseases, as changes in gene expression can lead to changes in the structure, activity, and expression of proteins and other molecules that can have a major effect on the health and well-being of an organism.

2.1.2 Gene Expression Data

Gene expression data shows how much of a particular gene is being expressed in a given cell or tissue. It is typically measured as the amount of messenger RNA

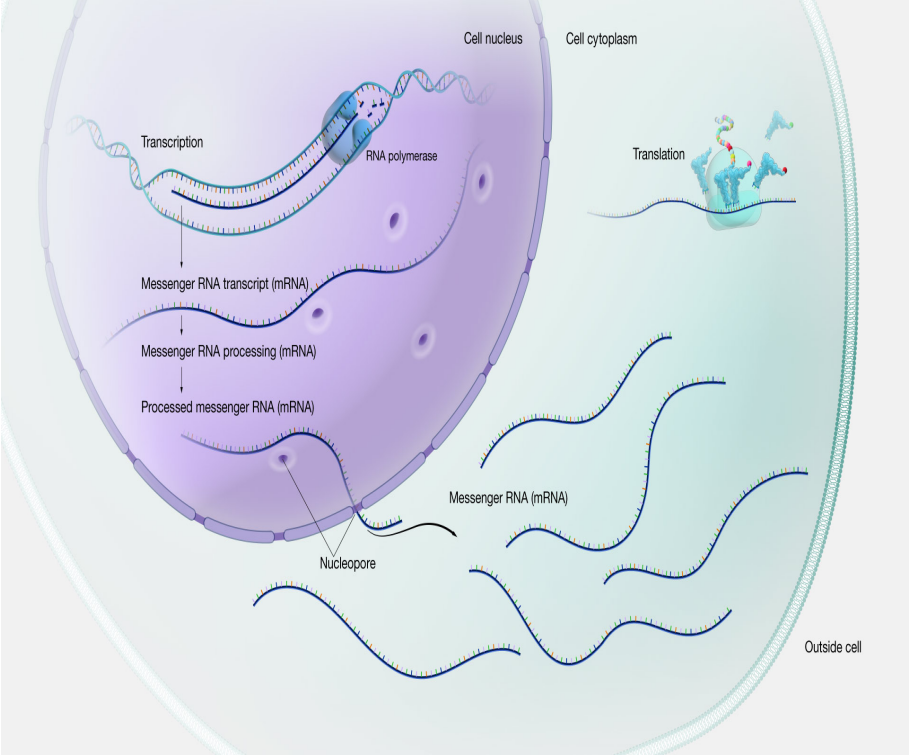

Figure 2-2: Schematic representation of Gene-expression. Source: https://www.genome.gov/genetics-glossary/, accessed on 21/07/2023

(mRNA) that is being produced from a gene. Gene expression data is typically represented as a matrix of gene expression levels, with rows representing genes and columns representing samples. Each cell in the matrix contains a numerical value that indicates the expression level of the gene in the corresponding sample. The numerical value can either be an absolute measure of gene expression, such as the number of transcripts per million (TPM) or a relative measure of expression. The gene expression profile refers to the cumulative expression levels observed for a gene across different experimental conditions. Sometime, in the gene expression matrix, the levels of annotation are also added to either the gene or the sample. For example, the function of the genes, or the additional details are provided on the biology of the sample, such as disease state or normal state. Gene expression data is typically collected through high-throughput DNA sequencing technology, such as RNA sequencing or microarray analysis. This type of data can help researchers understand the underlying biological processes that influence a cell's behavior, such as how it responds to environmental changes or diseases. It can also be used to discover new genetic pathways and develop new treatments. Additionally, gene expression data can be used to identify genes that are associated with certain traits, such as cancer or obesity. By analyzing gene expression data, scientists can gain insight into the inner workings of a cell and better understand how it functions. Gene expression data can be used to study how a gene is regulated over

time, how it responds to different environmental conditions, or how it interacts with other genes in a gene network. The data provide valuable insights into the extent of expression differences that are associated with malignancy. Moreover, it identifies certain genes that have the potential to be useful as diagnostic or prognostic markers.

2.1.3 Gene Expression Data Generation

In this section, three popular techniques to generate gene expression data such as microarray, RNA-seq, and scRNA-seq data have been discussed.

2.1.3.1 Microarray Data

High-throughput microarray technologies are more efficient for quantifying mRNA than low-throughput methods due to their ability to analyze a large number of genes simultaneously [1]. Microarrays, also known as DNA chips or biochips, allow researchers to measure gene expression levels of thousands of genes at once, providing an unprecedented level of information. Manufacturing a microarray and using it to measure gene expression is a complex laboratory process, consisting of four main steps such as (a) sample preparation and labelling, (b) hybridization, (c) washing, and (d) image acquisition (as seen in Figure 2-3). A microarray chip refers to a slide made of glass or silicon, containing a grid of spots. Each spot holds multiple copies of a gene sequence known as a probe. The mRNA molecules, called the target, are extracted and labeled using fluorescent dyes, typically red for Cy3 and green for Cy5 [8]. These labeled mRNA molecules are then applied onto the slide, allowing them to hybridize with the complementary gene sequences on the array. The degree of hybridization varies based on the concentration of mRNA molecules present in the sample, particularly those genes highly expressed in different conditions. The chip is subsequently scanned to determine the color intensity at each spot, creating a digital image of the array that can be stored on a computer. To measure the differential expression level of a specific gene, the ratio between the signal intensities of the two colors is calculated. By analyzing the fluorescent light emitted from each hybridized spot, the amount of mRNA in the sample can be quantified. The resulting images are processed using image analysis software, generating an intensity matrix for further analysis [9]. The data is usually displayed in the form of a heat map, with different colors representing different levels of gene expression which is then transformed into a matrix form called the

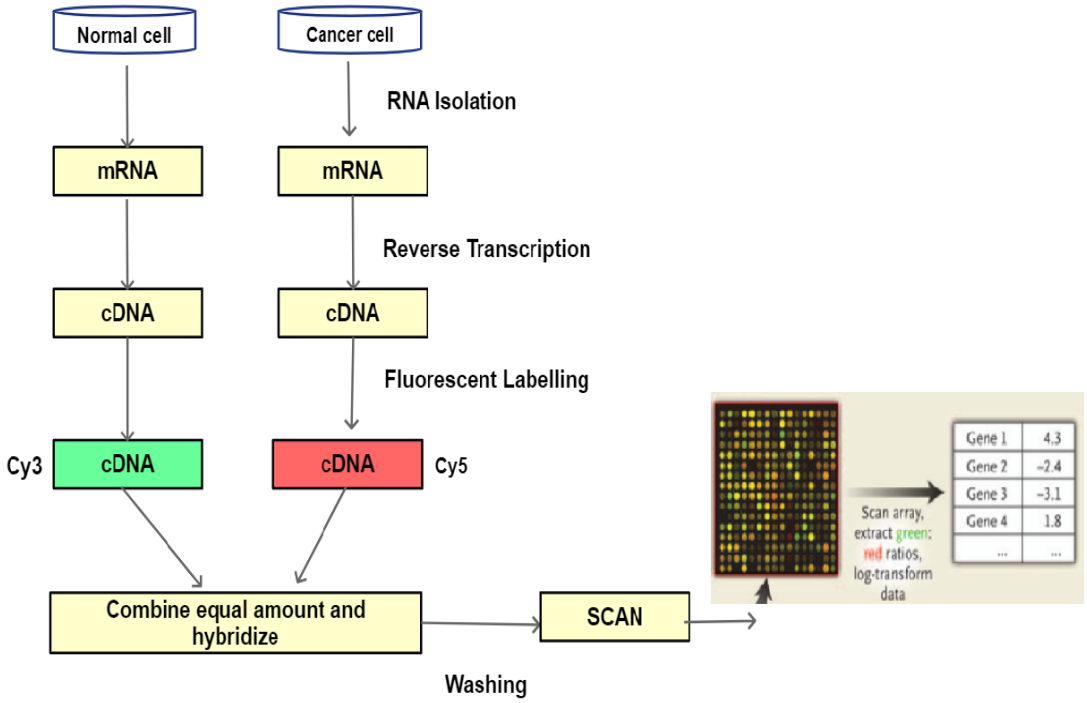

Figure 2-3: Schematic representation of DNA microarray technology.

gene expression matrix or gene expression data. This gene expression matrix or data, shown in Figure 1.4, has rows representing genes and columns representing samples or conditions, with values that represent the expressions levels of the gene in the sample.

2.1.3.2 High-throughput sequencing (HTS)

HTS is a newly invented cheaper, easier, and more powerful technology alternative to microarray. HTS offers many advantages over DNA microarrays [10]. In particular, it is more precise and not subject to cross-hybridization, thereby providing higher accuracy and a larger dynamic range. Next generation sequencing (NGS) is a powerful HTS method of DNA sequencing that uses massively parallel sequencing to sequence DNA at a much higher throughput. It uses high-speed, automated processes to sequence millions of DNA or RNA fragments at once, allowing researchers to gain a detailed understanding of genetic variation or expression of genes. High-throughput sequencing is used in a variety of applications, including genomics, epigenomics, transcriptomic, metagenomics, and clinical diagnostics.

A. RNA-seq Data: RNA-seq is an advanced technique that uses HTS or

NGS technologies to decode a transcriptome. A transcriptome includes the complete set of transcripts, such as protein-coding messenger RNA (mRNA) and non-coding RNA like ribosomal RNA (rRNA), transfer RNA (tRNA), and other non coding RNA in a tissue, organism, or a specific cell for a given physiological condition. The emergence of HTS platforms, such as the Illumina [11] or Solexa technology, Ion Torrent semiconductor sequencing technology, Single-Molecule Real-Time (SMRT) sequencing, and Oxford Nanopore Technologies in 2005, and revolutionized DNA sequencing [12]. RNA-seq is advantageous over other gene expression techniques due to its ability to identify novel transcripts and quantify expression levels of both known and unknown genes. Moreover, RNA-seq overcomes several limitations associated with previous technologies, such as microarrays, which often require prior knowledge of the organism being studied and failed to quantify the levels of the diverse RNA molecules that are expressed across a wide range of genomes. The main steps of the RNA-seq technique include sample preparation, library preparation, library amplification, sequencing, quality control, read alignment, transcript discovery, quantification and data analysis [13] (as seen in Figure 2-5 and 2-4). In the sample preparation step, total RNA is isolated from the sample and then purified to remove any potential contaminants. In the

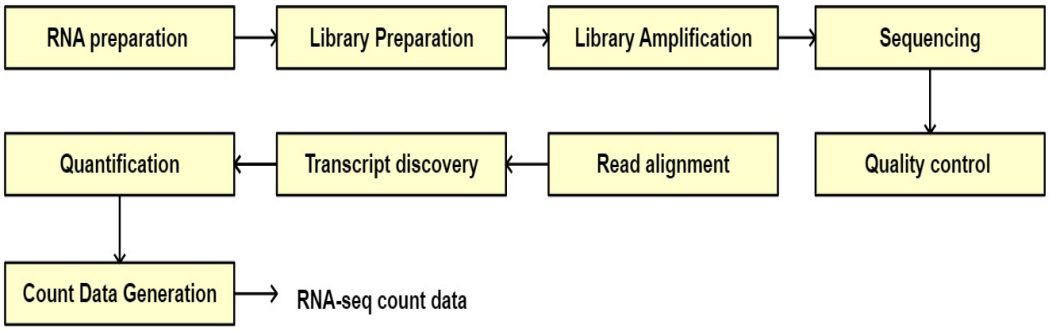

Figure 2-4: RNA-seq data generation steps

library preparation step, cDNA molecules are generated by reverse transcribing mRNA molecules from a given sample. The resulting cDNA molecules are fragmented into smaller pieces to enable sequencing, added sequencing adapters and amplified by PCR to enhance fragments. Next, they can be sequenced using one of the many available high-throughput sequencing platforms. A sequenced fragment is called a "read" and it can be classified as: exonic reads, junction reads and poly(A) end reads. The sequencing step involves passing the library of fragmented RNA molecules through a sequencing machine, which reads the sequence of the nucleotides in each fragment in a single direction (single-end) or both directions (paired-end). Single-end reads are typically cheaper and simpler to generate, but paired-end reads generate better alignments and assemblies for transcript isoform

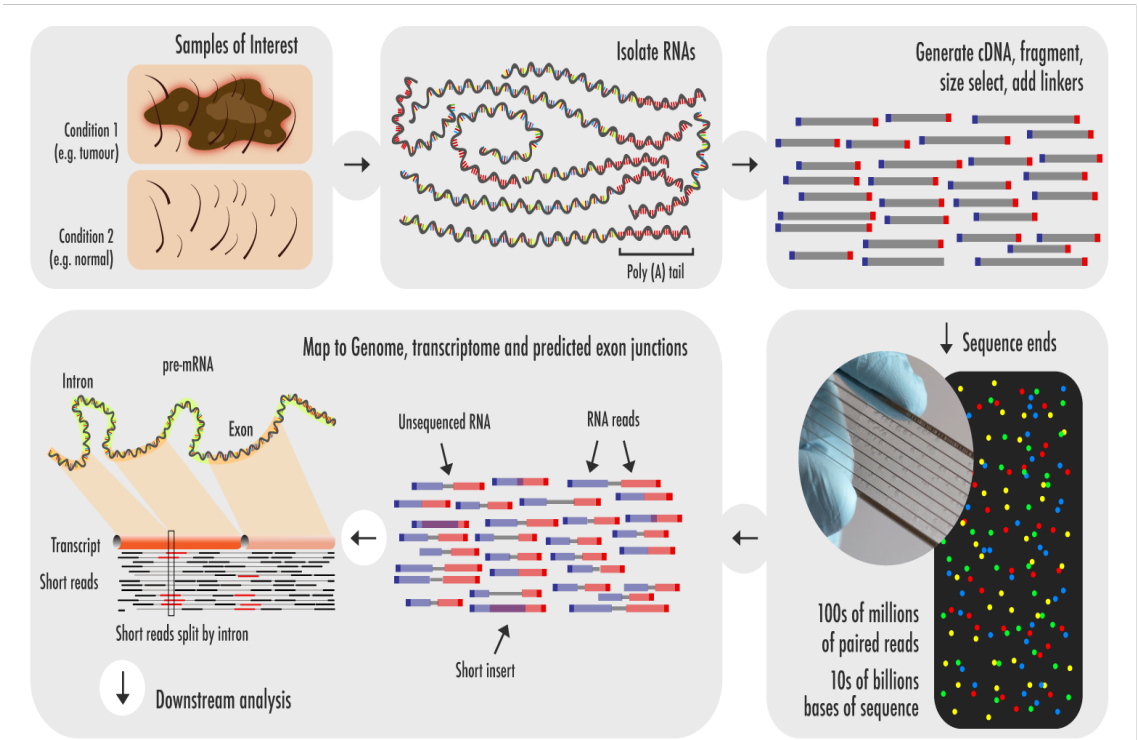

Figure 2-5: Schematic representation of RNA-seq technology. Source: https://en.wikipedia.org/wiki/RNA-Seq, accessed on 21/07/2023

discovery. The sequencing data is then processed for quality checking to assess the quality of the raw sequence reads and filtering out any reads that may be of poor quality. Once the quality of the data has been determined, the reads must be aligned to the reference genome or transcriptome. Alignment of the reads to the reference is done using a sequence alignment software, such as TopHat, Bowtie2, BWA, HISAT, AlignerBoost, GSNAP or STAR [14]. The main aim of alignment is to align sequence reads to intron boundaries accurately. If a reference genome or transcriptome is available, high-quality reads can be aligned to the reference using a reference sequence. However, if no reference is available, de novo transcript reconstruction is required to align the reads to the transcriptome. After alignment, the mapped reads for each sample are assembled into gene-level, exon-level, or transcript-level information to quantify expression levels. Each RNA-seq experiment produces a vector of read counts for genes, resulting in a matrix of count data. This matrix count data is the number of reads mapped to each gene, which can be used for downstream analyses, such as differential expression analysis [15]. Count data follows a discrete data distribution such as poisson and negative binomial. The datasets in RNA-Seq are larger and more complex, and the generated data cannot be interpreted easily without extensive bioinformatics intervention. Integration of results from RNA-Seq data with other biological data sources can help generate a better picture of gene regulation.

RNA-Seq helps to understand many biological phenomena, such as the underlying mechanisms and pathways controlling disease initiation, development, and progression. Initial transcriptomic studies were conducted using hybridization-based microarray techniques, but such RNA-Seq provides deeper coverage and resolution of the transcriptome, facilitating RNA editing, newly transcribed region detection, analysis of alternative splicing and allele-specific expression.

B. scRNA-seq Data: Single-cell sequencing is awarded as the method of the Year for 2013 by Nature Methods [16]. The first publication on next-generation sequencing platform for scRNA-seq analysis was published in 2009. It became more popular after 2014, when new protocols and lower sequencing costs made it more available. Single-cell RNA sequencing is used to understand gene expression patterns in individual cells. It allows scientists to study gene expression in a single cell instead of an entire population, providing a more detailed view of cellular biology. This technique enables researchers to identify gene expression differences between cell types and measure changes in gene expression over time or in response to an environmental stimulus. scRNA-seq uses a combination of high-throughput

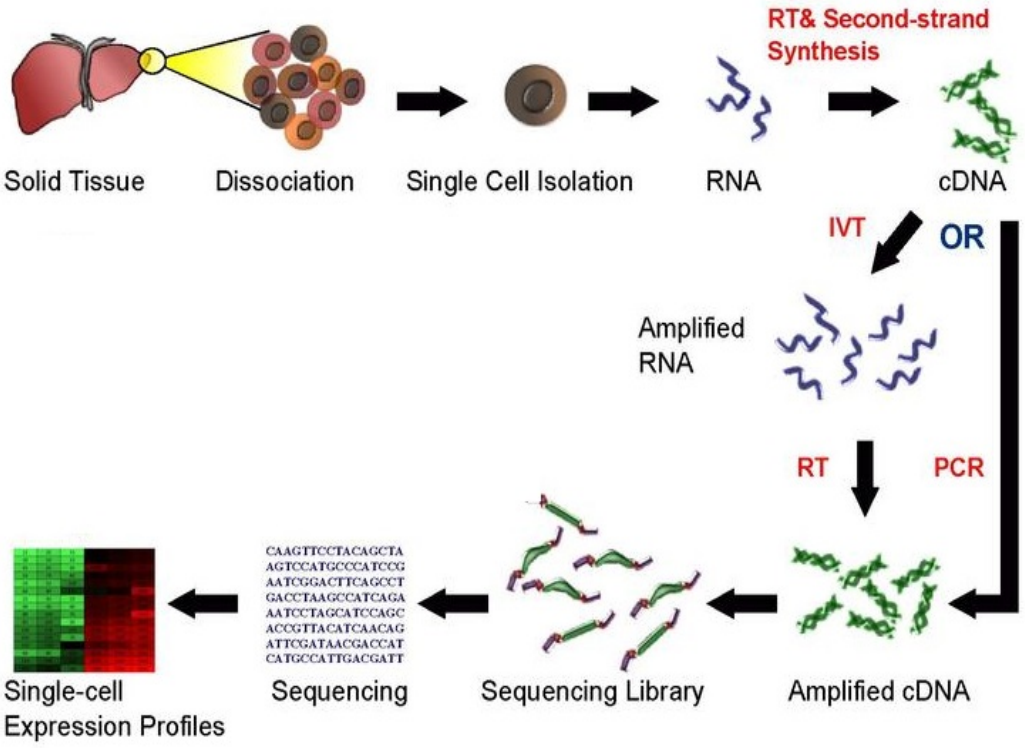

Figure 2-6: Schematic representation of scRNA-seq technology. Source:https://www.stephaniehicks.com/2018-bioinfosummer-scrnaseq/introduction-to-single-cell-rna-seq.html, accessed on 21/07/2023

sequencing and bioinformatics to analyze the transcriptome of individual cells. By

sequencing millions of cells, scientists can identify cell-specific genes and construct a gene expression profile for each cell. This allows researchers to group cells into cell types and identify distinct gene expression patterns. The process of scRNA-seq data generation begins with the isolation of single cells from a tissue sample. This can be done by either microfluidic sorting or manual sorting, depending on the type of cells being studied. In addition to isolating individual cells, scRNA-seq also requires the addition of barcodes or tags to each cell. These tags provide a unique identifier for each cell, which helps to distinguish it from other cells in the sample. Cell barcodes are used to identify the origin of a read, indicating which cell it came from, while UMI (Unique Molecular Identifier) helps to identify the specific mRNA molecule from which the read originated. This UMI information is useful in detecting and eliminating potential PCR duplicates in the data. This is especially important when analyzing large datasets, as it helps to ensure that each cell is accurately identified and analyzed. Once the cells are isolated, they are lysed to release their contents, including the mRNA molecules (as seen in Figure 2-6). The mRNA molecules are then reverse transcribed into cDNA and amplified to produce enough material for sequencing. The cDNA is then sequenced using a high-throughput sequencing platform such as Illumina. Thus short sequencing reads are obtained. The sequencing reads are then aligned to a reference genome to identify the expressed genes in each cell. Finally, the expression levels of these genes are quantified and used to generate a gene expression profile for each cell. The data generated by scRNA-seq can then be used for a variety of applications, including the identification of cell types, the discovery of novel genes, and the analysis of gene expression patterns. It can also be used to study the function of genes, as well as to identify potential therapeutic targets. scRNA-seq can be used to study gene expression in healthy cells as well as diseased tissue.

Microarray and RNA-seq technologies provide average gene expression values of cells, while scRNA-seq technology can quantify gene expressions at a cell level. Before conducting downstream analyses such as differential expression analysis, co-expression analysis, or differential co-expression analysis, each of these three technologies requires specific preprocessing techniques. Here, we will discuss some commonly used preprocessing methods for both RNA-seq and scRNA-seq data.

2.1.4 Preprocessing of RNA-seq data

2.1.4.1 Elimination of low read counts

Low read counts can be due to technical or biological reasons which affects the downstream analysis results. Low-expression genes of RNA-seq data may be indistinguishable from sampling noise, which can decrease the sensitivity of detecting DEGs [17]. Identification and filtering of these low-expression genes or low read counts may improve DEG detection sensitivity. The removal of low read count can be done in a number of ways, but a simple way is to discard any entries in the count matrix that have a read count less than a certain threshold. This is done to improve the accuracy of the results.

2.1.4.2 Normalization

Normalization is a process of adjusting the data to account for differences in library size, sequencing depth, and other factors that can affect the data. The goal of normalization is to reduce the effect of these factors and make the data more comparable across samples. Normalization is used to remove technical variation from noisy data. Normalization techniques for RNA-seq include Quantile [18], Reads Per Kilobase of Transcript Per Million Mapped Reads (RPKM) [19], logarithmic transformation, Counts Per Million (CPM), Transcripts Per Million (TPM) [20], Upper Quartile (UQ) [21], Trimmed Mean of the M-values (TMM) [22], PoissonSeq and DESeq [23]. For microarray data pre-processing, Min-Max normalization, Z-score normalization and Decimal scaling normalization are commonly used.

2.1.4.3 Transformation

Transformation of RNA-seq data, which are discrete measurements of gene expression, is necessary for the application of statistical methods that assume a normal distribution. Transformation techniques include log transformation, variance-stabilizing transformation (VST), rank-based transformation, and Box-Cox transformation [24]. Log transformation is the most commonly used technique to transform RNA-seq data. Log transformation involves taking the natural logarithm of each expression value. This transformation is used to normalize the data and reduce variance. Variance-stabilizing transformation (VST) is another commonly used technique for transforming RNA-seq data. This transformation is based on

the quantile normalization of expression values across multiple samples. It is used to normalize the data and reduce the variance of the expression values. Rank-based transformation is a transformation technique based on the ranking of the expression values across multiple samples. This transformation is used to normalize the data and remove any bias from the data. Lastly, Box-Cox transformation is a statistical technique used to transform data to a normal distribution. This transformation is based on the principle of maximum likelihood and is used to normalize the data and improve its statistical properties.

2.1.4.4 Imputation

Imputation of missing values is a statistical technique used in RNA-seq to fill in missing values in the data. Missing values often occur in RNA-seq data due to technical limitations in the sequencing process and can lead to inaccurate results. Imputation pre-processing works by using the data that is available to estimate the values of missing data. The most common methods used for imputation pre-processing in RNA-seq include k-nearest neighbors (kNN), singular value decomposition (SVD), and multiple imputation by chained equations (MICE). .

2.1.4.5 Removal of Batch Effect

Batch effects are a common challenge in RNA-seq data analysis. They can arise from differences in experimental conditions or in the sequencing process itself. Batch effects can cause spurious differences in gene expression levels between samples, and can complicate the identification of differentially expressed genes. COMBAT [25] and ARSyN [26] are two methods used in removing batch effects from RNA-seq data, as well as microarray data.

2.1.5 Preprocessing of scRNA-seq data

2.1.5.1 Elimination of low read counts

Low read counts can indicate that the cell has a low abundance of the targeted transcript or could indicate technical issues with the sequencing. By removing these cells, it can help improve the accuracy of downstream analysis.

2.1.5.2 Imputation in scRNA-seq data

ScRNA-seq datasets typically experience high levels of zero inflation and dropout events, which can lead to inaccurate downstream analysis. To combat this, several imputation methods such as scUnif [27], MAGIC [28], scImpute [29], DrImpute [30], SAVER [31], and BISCUIT [32] have been developed. These imputation techniques allow researchers to fill in missing values in scRNA-seq datasets and to reduce the effect of dropout events. scUnif and MAGIC are two unsupervised methods, while scImpute, DrImpute, SAVER, and BISCUIT are supervised methods. Unsupervised methods use the gene expression data itself to infer missing values, while supervised methods use external information and reference datasets. All of these methods have the potential to improve the quality and accuracy of scRNA-seq data. scUnif uses a univariate normal distribution to impute missing values, while MAGIC uses a non-parametric nearest-neighbor imputation. scImpute uses a low-rank matrix approximation to impute missing values, while DrImpute uses a deep learning approach to impute missing values. SAVER uses a probabilistic model to impute missing values, while BISCUIT uses a Bayesian nonparametric model to impute missing values. These methods help to improve downstream analysis by reducing the effect of dropouts and zero inflation, as well as by providing more accurate estimates of gene expression levels.

2.1.5.3 Removal of low quality and doublet cells

The removal of low-quality and doublet cells preprocessing step is an important part of single-cell RNA-seq (scRNA-seq) data analysis, and involves the elimination of low-quality cells and doublets. Low-quality cells are typically identified as those with low sequencing depth, low gene expression, or low signal-to-noise ratio. Doublets are two cells that are usually captured in the same droplet accidently, resulting in a combined pseudo-single cell. Removal of these cells is important to ensure the accuracy and reliability of downstream analyses. This process can be done using various software tools, such as Cell Ranger, Cell Ranger ATAC, or Seurat. These tools can identify and remove low-quality cells and doublets based on a variety of metrics, such as gene expression, principal component analysis, or a combination of both. This preprocessing step also helps remove unwanted cells, such as dead cells or cells from other species, which could introduce noise into the data set.

2.1.5.4 Batch Effect Removal in scRNA-seq data

This step can be used to correct the differences in sample preparation, library construction, sequencing platform, and other sources of technical variability. It is important to ensure that the data is not biased in any way and that the results are representative of the true biological state. This preprocessing step can be executed in a number of ways, including adjusting data normalization, specific to scRNA-seq data, or using a dimensionality reduction technique such as Principal Component Analysis (PCA).

2.1.5.5 Normalization

Normalization of scRNA-seq data is the process of adjusting the read counts for each gene across cells to ensure that technical biases are removed and that the expression values are comparable across cells. This is typically done using a scaling factor to make the sum of the read counts in each cell equal to a predetermined number, such as 10,000 or 1,000,000. Normalization also ensures that the overall expression distributions are consistent across all cells. This is important for producing accurate downstream analyses such as clustering, differential expression, and trajectory inference.

2.1.5.6 Highly variable gene selection

Highly variable gene selection is an important preprocessing step for scRNA-seq data that allows researchers to identify genes that are most informative and variable, and to reduce the dimensionality of the data. This process involves selecting a subset of genes that have the highest average expressions across all cells, as well as the higher amount of variability among the cells. The selection process is typically based on the coefficient of variation (CV) or dispersion measures such as the median absolute deviation (MAD). Genes with the highest CV or MAD values are selected as the most informative and variable across all cells. Once the highly variable genes are identified, one can perform downstream analyses such as clustering, biclustering or triclustering.

2.1.5.7 Transformation

Log transformation and variance stabilization transformation (VST) are commonly employed for the analysis of scRNA-seq data [33].

2.1.5.8 Dimensionality Reduction

Dimensionality reduction is a preprocessing step used to reduce the dimensions (or features) of scRNA-seq data. This is done to remove redundant and irrelevant features that do not contribute to the overall structure of the dataset, as well as to reduce the complexity of the data. Dimensionality reduction is typically performed using Principal Component Analysis (PCA) [34] or t-distributed Stochastic Neighbor Embedding (t-SNE) [35]. PCA is a linear transformation technique that creates a new coordinate system with fewer dimensions that still captures most of the variance in the original data. t-SNE is a nonlinear transformation technique that creates a low-dimensional embedding of the data in which similar samples are closer together and dissimilar samples are further apart. Both techniques can be used to reduce the dimensionality of scRNA-seq data and make it easier to visualize.

2.1.6 Gene Expression Data Analysis

This section presents three prominant gene expression data analysis approaches such as differential expresssion analysis, co-expression analysis and differential co-expression analysis.

2.1.6.1 Differential Expression Analysis (DEA)

DEA is used to compare gene expression levels between two or more biological conditions. It is used to determine how differentially expressed genes (DEGs) are related to specific biological processes or diseases. DEA can be used to identify genes that are up- or down-regulated in response to stimuli, such as drugs, hormones, or environmental conditions. Differential analysis between two conditions can be viewed as clustering into two categories, changed or not changed[36].

DEA is an important tool used in understanding the molecular basis of diseases and in the development of new therapeutic agents. By studying which

RNA-seq count data

	Normal					Disease				
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
G1	21	16	29	24	22	31	26	39	34	32
G2	13	8	21	16	14	10	5	18	13	11
G3	162	7	74	0	12	0	109	1	9	926
G4	57	26	77	91	51	34	50	31	49	9
G5	5	6	14	4	31	1	8	7	6	1

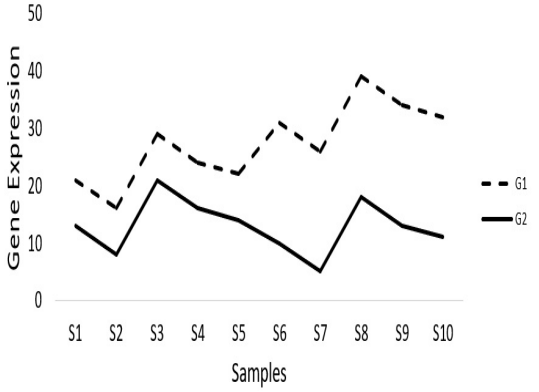

Figure 2-7: An example of differential expression pattern. The gene expression of G1 varies (increases) significantly in disease conditions. Whereas, gene G2 shows almost steady expression in disease condition. So, gene G1 may be considered as a differentially expressed gene. In this figure, first 5 samples are from normal condition and rest are from disease condition.

genes are expressed differently in different samples, one can gain insight into which genes are associated with certain diseases or processes. This can help to identify potential drug targets and guide drug development. Additionally, DEA can be used to identify novel diagnostics or prognostics biomarkers. In Table 2.1, various widely used DEA tools have been listed. There are different methods for DEA such as edgeR[37], DESeq[38], DESeq2[39] NBPseq[40], baySeq[41] and EBSeq[42] which are Bayesian approaches based on a negative binomial model. From the literature, it is found that the best performing tools[43] tend to be edgeR[37], DESeq/DESeq2[38][39], and limma-voom[44]. DESeq and limma-voom tend to be more conservative than edgeR (better control of false positives), but edgeR is recommended for experiments with fewer than 12 replicates[45].

2.1.6.2 Gene Co-expression Network Analysis (GCNA)

The co-expression network was first introduced in the year 1999 by Butte and Kohane [68]. A gene co-expression network (GCN) is an undirected graph extracted from a gene expression dataset, where the strength of associations between a gene pair is represented with the help of nodes and edges. The higher value of association corresponds to high biological significance. A GCN highlights the statistical correlations or semantic similarities among the genes. To find a GCN, a correlation matrix is computed on the pre-processed dataset. This matrix consists of all possible genes including those that are highly co-expressed, loosely co-expressed, and zero co-expressed. If a change in one gene tends to follow a change in another

Table 2.1: Differential Gene Expression Analysis Methods/Tools

Methods	Year	Type of software
limma[46]	2004	R package
DEGseq[47]	2009	R package
edgeR[37]	2010	R package
DESeq[23]	2010	R package
baySeq[41]	2010	R package
Cuffdiff[48]	2010	Command-line user interface
Cuffdiff2	2013	Command-line user interface
BBSeq[49]	2011	R package
DEXseq[50]	2012	R package
EDAseq[51]	2012	R package
Bitseq[52]	2012	R package
ShrinkSeq[53]	2012	R package
QuasiSeq[54]	2012	R package
SAMseq[55]	2013	R package
Ebseq[42]	2013	R package
rSeqDiff[56]	2013	R package
DSGSeq[57]	2013	R code
DESeq2[39]	2014	R package
ShrinkBayes[58]	2014	R code
edgeR-Robust[59]	2014	R code
limma-voom[44]	2014	R package
NBPSeq[60]	2014	R package
ImpulseDE[61]	2016	R package
SARTools[62]	2016	Standalone
NOISeq[63]	2016	R package
BNP-Seq[64]	2016	Standalone
DEApp[65]	2017	Web tool
iDEP[66]	2018	Web tool
ideal[67]	2020	R package

gene, it is assumed that the two genes are associated or correlated and this interdependence is called correlation or covariation, and the pair of genes are called co-expressed genes. A co-expressed gene may be either positively co-expressed or negatively co-expressed. Two genes G1 and G2 are called positively co-expressed, if an increase in G1 is associated with an increase in G2, exhibiting a positive correlation score close to 1. Two genes G1 and G3 are called negatively co-expressed if an increase in G1 is associated with a decrease in G2, exhibiting a correlation score close to -1 or 0 depending on the correlation strength. A GCN module M, is a dense sub-network of a GCN consisting of highly co-expressed genes with common biological processes and similar functions.

Figure 2-9 depicts the classification and types of GCN. To obtain the degree of correlation between a pair of genes, several similarity and distance measures

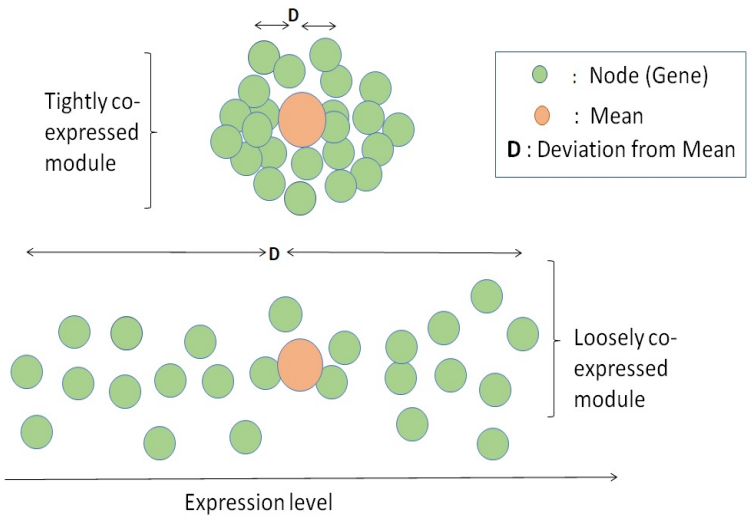

Figure 2-8: An example of co-expressed module.

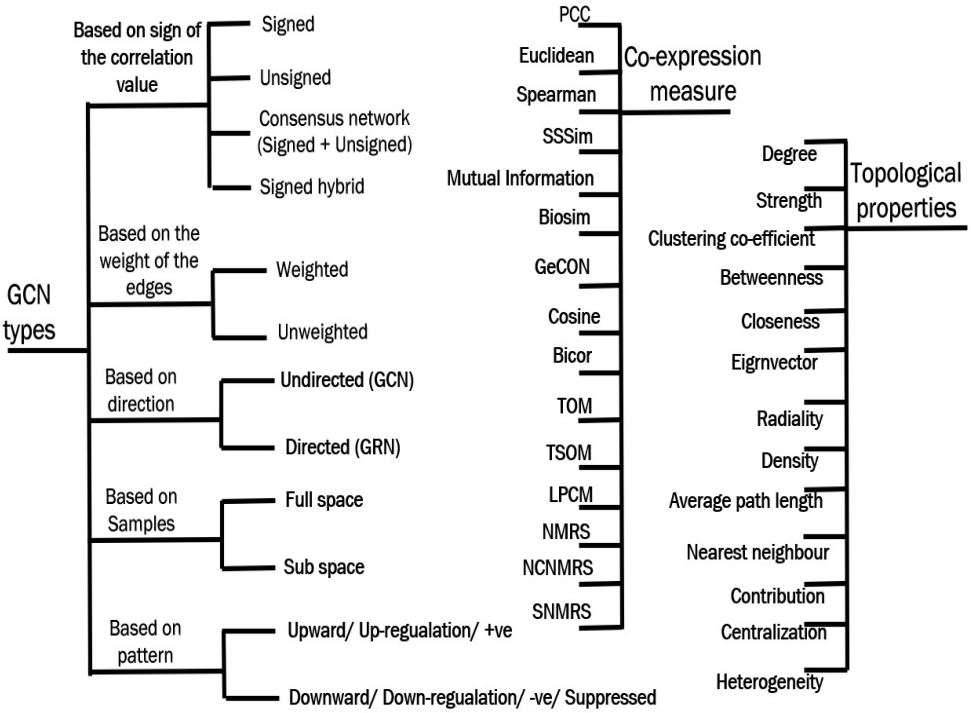

Figure 2-9: Types of GCNs and types of co-expression measures and topological properties

are available among which the Pearson Correlation Coefficient is the most reliable one. If the correlation value between two genes exceeds a certain threshold, they are considered to be related. A GCN is a single, large, complicated network, that helps to perform (i) downstream analysis, which includes the discovery of comparatively small groups of co-expressed genes, and hub genes, (ii) guilt by association analysis, (iii) transcriptome regulatory network construction, and (iv) preservation and disruption analysis. Hence, the importance of GCN analysis in

biological research can not be overstated. A GCN allows researchers to investigate the roles of undiscovered genes and their links to illnesses, prioritization of candidate genes, and identification of regulatory genes. The functional linkage between genes can be identified by observing and analyzing the coordinated behaviour of pairs of genes across samples [69]. Weirauch [70] reported that co-expressed genes are regulated by the same regulator, functionally linked, or are participants of the same pathway or protein complex. The patterns of co-expressed genes follow absolute, shift-or/and-scale, inverse/cross and re-wiring (means rearranging the ties in a graph) expression patterns across a variety of experimental samples [71]. A GCN analysis method must have the ability to handle all types of such expression patterns. A GCN can be of different types: weighted or unweighted based on the type of strength and directed or undirected based on the type of relationship. A GCN is characterized by topological properties, modularity, presence of intra/inter hub genes, and power-law degree distribution. Though GCNs and PPI are static in nature, they include a great deal of information regarding dynamic processes such as genetic network activity in response to DNA damage, protein function, and genetic interaction [72].

Several methods have been proposed in the past for GCN analysis. We have come across a few extensive surveys related to GCN analysis for all types of gene expression datasets. Extensive surveys on GCN, found in Aoki et al. [69], Van Dan et al.[73], Horvath et al.[74], and Chowdhury et al.[71]. Aoki et al.[69] and Horvath et al.[74] provide discussions discussion of co-expression networks with network architecture, terminologies and associated diagrams. Additionally, Van Dam et al.[73] and Chowdhury et al.[71] present tools and methods related to such networks. Another systematic literature survey on GCN is reported in [71]. Co-expressed genes are members of the same pathway or protein complex, are controlled by the same transcriptional regulatory mechanism, or are functionally related [70]. However, unlike a GRN, a GCN does not define the causation links between genes [75]. An interesting geometric interpretation for GCN was introduced by Horvath et al. [74]. Identifying disease-related modules leads to the development of improved diagnostic strategies and new drugs[76]. The gene expression network topology provides the basis for molecular characterization of the cellular environment [77]. The most extensively used tool for generating GCN, recognizing modules performing preservation analysis, identifying hub genes, and selecting potential genes as biomarkers is WGCNA [78]. WGCNA calculates a weighted adjacency matrix using a signed co-expression measure and soft thresholding. The adjacency matrix is turned into a Topological Overlap Matrix (TOM) to reduce the impact of noise and spurious associations. Then, the corresponding

dissimilarity matrix is calculated on which Hierarchical clustering is performed to get modules of highly co-expressed genes. To quantify co-expression similarities of entire modules, their eigengenes are calculated and clustered based on correlation. Hierarchical tree visualizing functions as well as functions for presenting the correlation matrix in heatmap form are available in the WGCNA package. It does not provide a function to produce a graphical view of the co-expression network but does allow exporting networks into Cytoscape, to visualize the networks [78]. Jianqiang Li et al. [79] modifed the original WGCNA working pipeline to study high-throughput genomic data. A combination of Signed and Unsigned WGCNA (csuWGCNA) is a modified WGCNA approach [80]. A signed WGCNA outperforms an unsigned WGCNA in terms of expression pattern detection in a module [81]. THD-Module Extractor is a method for detecting and extracting GCN modules from microarray datasets [82].

2.1.6.3 Correlation

In order to measure the strength of a linear relationship between two gene expression profiles, the correlation coefficients are used. The value of a correlation coefficient value which is greater than zero implies a positive relationship while a correlation coefficient less than zero indicates a negative relationship. Further, when the value of correlation coefficient is zero, it signifies there is no relationship between the two gene expression profiles being measured. Correlation score helps detecting two gene expression profiles whether they co-expressed or not.

A correlation pattern is obtained by adding (or subtracting) the same amount (the rate of change) each time to get from one state to the next. Correlation patterns can be positive or negative which may be absolute, shifting, scaling, shifting-and-scaling and scaling-and-shifting. An example is shown in Figure 2-10. From a random variable a=(3,7,5,2) other variables 'b', 'c', 'd', and 'e' variables are calculated to demonstrate inhibition shifting, scaling, and shifting-and-scaling correlations individually. Here, 'b' and 'c' are two different shifted patterns with the same amount of change. The types of correlation patterns discussed above are shown in Figure 2-10 which shows the mechanism for inducing changes during pattern transformation.

For GCN construction, various correlation measures have been suggested till date. We discuss some widely used measures used to construct a GCN. For easy comparison, we present Table 5.2 that shows the formulas and characteristics of different co-expression measures.

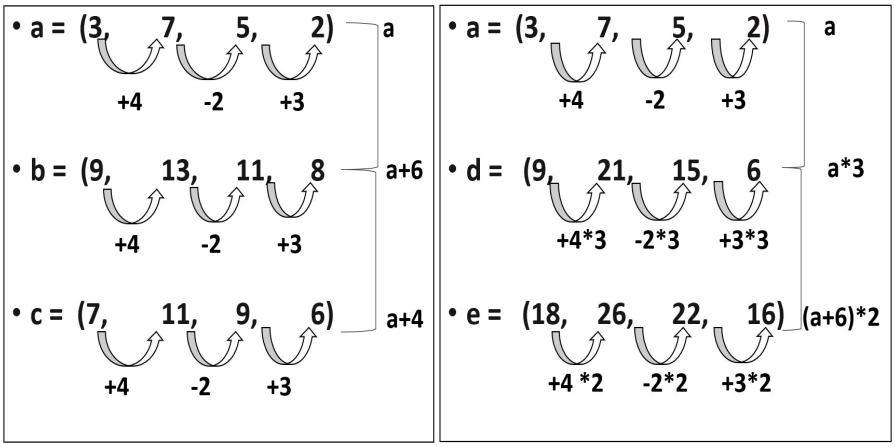

Figure 2-10: Different types of patterns with change. Here, 'a' is a random variable, 'b' and 'c' exhibit shifted pattern, d exhibits a scaling pattern, and e inhibits a shifting-and-scaling pattern.

Euclidean distance: The geometric distance between two gene profiles can be measured using Euclidean distance [83], which takes into account both the magnitude and direction of the gene profiles. The euclidean metric is not appropriate when the absolute amounts of functionally related genes differ significantly.

Pearson's correlation coefficient (PCC): PCC [84] is the most popular co-expression metric and is recommended by many authors in constructing GCNs. It measures the tendency of two gene profiles to rise or drop simultaneously in the range -1 to 1. PCC is calculated as the cosine of the angle between the mean-centered profiles. Mean centering maintains the shape of the profile[85]. PCC is robust to order sensitivity. The PCC measure has limitations in that it can only detect linear relationships and is susceptible to outliers. PCC does not work with identical vectors. For PCC, zero correlation indicates independence or no linear relationship between a pair of genes. However, in this case, there may be the possibility of nonlinear dependence and a curvilinear relationship. PCC is computationally intensive, particularly in the case of large matrices[86]. After standardization, Pearson correlation and Euclidean distance can be proven to be equivalent [87].

Spearman's rank correlation coefficient: Spearman's rank correlation [88] does not take into account the actual magnitude of the expression ratio between two genes but takes into account the 'rank' of the expression ratio between two genes. When the variables are measured on a scale that is ordinal, it is useful because it makes no assumptions about the distribution of the data. Also, it is resistant to data outliers. However, datasets with a small number of samples, it is less sensitive to expression values and may detect a large number of false positives.

Mutual Information: Mutual information (MI) [89] is a metric that identifies non-linear statistical dependence between two genes, similar to the correlation coefficient. Song et al. [90] discovered that MI did not outperform correlation in many circumstances.

Bicor: Biweight mid-correlation (commonly known as bicor) is a statistical measure of similarity between a pair of genes. Because bicor is median-based rather than mean-based, it is less vulnerable to outliers than other similarity measures like Pearson correlation or mutual information, according to [90]. It is used for weighted correlation network analysis to evaluate similarity in gene expression networks.

Topological overlap matrix (TOM): This measure is based on shared network neighbours [91]. It was designed to make networks less vulnerable to erroneous connections or connections that were lost due to random noise. The topological overlap between two nodes indicates how similar they are in terms of the nodes they link to. The higher the overlap between two substrates within a metabolic network, the more likely they belong to the same functional class, researchers have discovered.

2.1.7 Degree centrality

Nieminen [92] first introduced the concept of degree centrality for an undirected graph. As described by Freeman in 1979 [93], degree centrality is a count of the number of edges incident on a given node of a network. The more connected a node is in the network, the more essential it is according to this metric. The node or vertex is nothing but a gene in GCN. The number of genes adjacent to a gene determines its importance in a co-expression network. Only a small number of genes have high degrees in many real-life networks. The nodes with large degrees are also known as hub genes [94]. The limitation of degree centrality is that its value depends on network size. To overcome this issue, one can compare the relative centralities of points from different graphs [93]. In comparison to non-disease genes, disease genes are said to have a higher degree of adaptation [95].

2.1.7.1 Differential Co-expression Analysis (DCEA)

DCEA involves the identification of gene modules whose co-expression patterns exhibit significant variations across different conditions [71]. It involves compar-

ing gene expression data from two or more different conditions or experiments and then looking for statistically significant differences between them. DCEA is based on the assumption that the expression of genes is correlated within the same sample or experiment. Therefore, if two different conditions or experiments show different gene expression patterns, then it is likely that the genes involved in those differences are playing a role in the differences between the samples or experiments. Detecting the genes which change their expressions in different conditions (such as normal versus cancer) is an essential task and can help understand the causes of diseases [96]. This type of analysis can be useful for studying gene expression changes in response to treatments, or for identifying genes that are involved in disease processes. It can also be used to identify novel gene regulatory pathways and to study gene-gene interactions. It can also help to identify novel biomarkers and therapeutic targets. By understanding the changes in gene expression between different conditions or experiments, researchers can gain a better understanding of the underlying biological processes. In addition to the above three approaches, preservation analysis is commonly used to study the effectiveness of gene subsets or modules extracted from the network constructed using any of the above approaches.

The most commonly utilized tools for DCEA, such as WGCNA [78] and DiffCoEx [97], share a similar approach. They initially identify co-expressed modules across the entire set of study samples. These modules can then be correlated with predefined sample subpopulations, representing factors like disease status or tissue type. In the case of WGCNA, it evaluates the activity and significance of each module in each subpopulation. It calculates an eigen gene for each module, which represents the expression pattern of all genes within that module across the analyzed samples. It then prioritizes genes that behave similarly to the eigen gene or are intra-modular hub genes, as these are likely to be associated with the phenotype linked to the module. DiffCoEx, on the other hand, focuses on modules that exhibit differential co-expression with the same sets of genes. This can include sets of genes that "hop" from one correlated gene set to another in a coordinated manner. DiffCoEx clusters these "hopping" genes accordingly. DINGO [98], a more recent tool, resembles DiffCoEx by grouping genes based on their differential behavior in a specific subset of samples compared to the baseline co-expression observed across all samples. These genes are more likely to explain phenotypic differences attributed to the two different

2.1.7.2 Preservation Analysis

A preservation study of a module extracted from a GCN can reveal a lot of hidden information. Such analyses help quantify structural changes across conditions as well as the amount of preservation across states, using statistics such as Zsummary[99] and MedianRank[99]. These are the two commonly used preservation quantification techniques. Preservation analysis is a useful method for comparing co-expression networks and identifying modules that have been preserved by natural selection or have been disrupted by certain pathways or biological processes.

2.2 Biomarker Identification

Biomarkers are crucial genes that are used to identify and track changes in the body. Biomarkers can be used to diagnose, monitor, and predict disease progression as well as to evaluate the effectiveness of treatment [100]. They are also used to assess environmental exposures and to measure drug response [101]. Biomarkers are often identified using genetic sequencing technologies such as next-generation sequencing and targeted sequencing. These technologies enable researchers to identify and map gene sequences within a genome, and then use this information to identify the biomarkers that are associated with a particular disease or condition. Biomarker identification involves a combination of genetic and molecular techniques, including gene expression profiling, gene sequencing, and proteomics. These techniques are used to identify the biomarkers that are associated with a particular disease or condition [101]. Once identified, these biomarkers can be used in clinical trials to test the efficacy of treatments. The identification of biomarkers can be an important tool in the diagnosis and treatment of diseases [78]. By identifying biomarkers, doctors can determine if an individual is at risk of developing a certain disease, or if they have an existing condition. Additionally, biomarkers can help doctors determine the best course of treatment for an individual, and can also be used to monitor the progress of a disease. Many authors have identified critical genes or biomarkers associated with cancers using GCN analysis methods. These cancer types include alzheimer [102], gastric cancer [103], adenocortical [104], COVID-19 [105], and ovarian [106].

2.2.1 Cluster Analysis

Clustering is used in gene expression analysis to investigate the occurences of interesting patterns across the states or conditions. Clustering is the common way technique to analyze the gene expression data [107]. Clustering allows researchers to identify co-expressed gene expression patterns and to uncover relationships among genes, gene expression levels, and biological processes. Clustering methods identify co-regulated genes, identify pathways and networks, and to understand the dynamics of gene expressions.

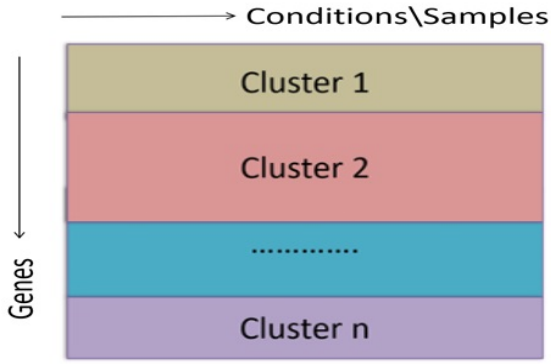

Figure 2-11: Clusters

Clustering algorithms are typically used to group genes with similar expression levels. This is done by grouping genes following similar trends or patterns into clusters. Clustering algorithms can also be used to identify genes whose expressions are significantly different from the rest of the group. This is useful for identifying novel genes that may be involved in a particular biological process. Clustering can also be used to identify pathways and networks by grouping genes that have similar expression profiles. This can help to elucidate the relationship between genes and biological processes, as well as their potential role in disease development. Additionally, clustering can help to identify gene expression signatures that are associated with different diseases. Clustering has been used in gene expression analysis for decades, and its utility continues to increase as new technologies and data sources become available. Clustering can be used to identify novel genes, uncover relationships between gene expression and biological processes, and to understand the dynamics of gene expression. Additionally, clustering can be used to identify gene expression signatures associated with different diseases. Some clustering techniques are - K-Means Clustering [108], hierarchical custering [109], DBSCAN [110], Self-Organizing Map [111], CLICK [112], etc. K-means is a widely used clustering algorithm that aims to partition data points into K distinct clusters. It works by iteratively assigning data points to the cluster cen-

troid that is closest to them and updating the centroids based on the newly formed clusters. K-means is known for its simplicity and efficiency, but its performance can be sensitive to the initial choice of centroids. Hierarchical clustering builds a hierarchical structure of clusters by iteratively merging or splitting them. It does not require specifying the number of clusters beforehand and can be visualized as a dendrogram. The two main types of hierarchical clustering are agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering starts with each data point as its own cluster and recursively merges the most similar clusters until a stopping condition is met.

2.2.2 Biclustering or Two-way Clustering Approaches

Nowadays biclustering is a well-known technique for the study of gene expression data, to discover functionally related set of genes under different subsets of experimental samples or conditions [113]. This subset similarity method has been named as biclustering [113], co-clustering or block clustering [114]. Biclustering is used to identify patterns of gene expression across multiple samples. It is used to identify both the genes and samples that are associated with a particular pattern of gene expression. Several unsupervised machine learning techniques have been developed to analysis the gene expression data obtained from DNA microarray experiments. These algorithms have helped us to understand conceptually and to visualize the basics of clustering and Biclustering approaches. Biclustering is a two-step process. First, it searches for genes and samples that are associated with a pattern of gene expression. It does this by examining the correlation between gene expression values across multiple samples and genes. Then, it uses the identified genes and samples to form clusters. These clusters can then be analyzed to identify patterns of gene expression that are associated with different biological processes. These patterns may indicate the presence of a disease, or they may provide insight into how a particular gene is involved in a biological process. To identify the existence of various types of correlations among the expressions of a group of biologically significant genes using biclustering technique is a challenging task for the researchers. Some examples of bilcutering techniques are Cheng and Church, OPSM [115], xMotif [116], Qubic [117], Bimax [118], IBBIG [119], SAMBA [120], Plaid [121], Spectral [122], ISA [123], CPB [124], FABIA [125], BBC [126], BIBIT [127], and COALESCE [128].

ChengCheng and Church (CC) [113] algorithm is the first biclustering algorithm developed to overcome the drawbacks of clustering algorithms in gene ex-

pression data analysis. This algorithm follows a greedy strategy and node-deletion approach. In CC algorithm the concept of a new measure named as *Mean Squared Residue*. Order Preserving Submatrix (OPSM) [115] is a deterministic greedy algorithm that seeks biclusters with ordered rows. Though OPSM can construct complete biclusters by extracting constant columns, shifting, scaling and shifting-scaling biclusters, but, this algorithm is based on the order of values and hence the algorithm is quite restrictive and it takes a long time for large dataset. Liu and Wang proposed Maximum Similarity Bicluster (MSB) [129] based on the similarity score. This method extracts additive biclusters using a greedy approach. It is a polynomial-time algorithm to find an optimal set of biclusters with the maximum similarity. Iterative Signature Algorithm (ISA) by Ihmels et al [115] is a non-deterministic greedy algorithm that seeks biclusters with two symmetric requirements and each column in the bicluster must have an average value above some threshold. Multiple biclusters can be discovered by running the ISA algorithm on several initial gene sets. Its drawback is that there is no evaluation of the statistical significance. xMOTIFs is a non-deterministic greedy algorithm that seeks biclusters with conserved rows in discretized dataset [130]. In Biclusters Inclusion Maximal (Bimax), the raw data is first converted into binary and applied a fast divide-and-conquer technique that partitions the discretized matrix into different sub-matrices, one of which contains only 0-cells and other contains 1-cell [118]. The algorithm is then recursively applied and the recursion ends if all the biclusters are extracted which contain only ones. The deterministic algorithm QUBIC [117] works in discrete data and looks for biclusters with nonzero constant columns.

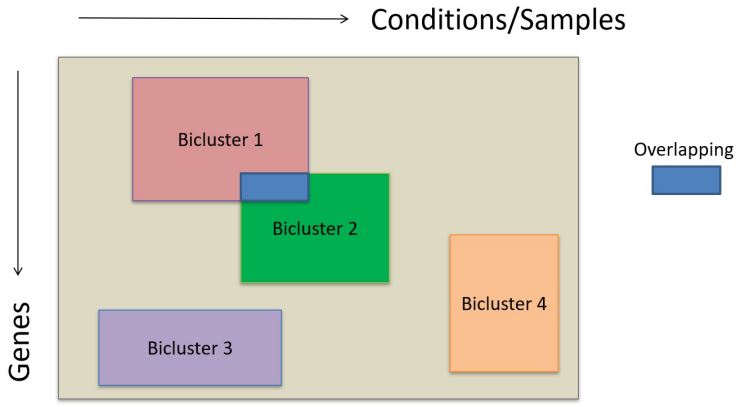

Figure 2-12: Biclusters

2.3 Gene Expression Data Repositories

Gene expression data repositories are specialized databases that store and manage large amounts of gene expression data generated from various experiments and studies. They are used to store, organize, analyze, and share data generated from a variety of high-throughput experiments such as microarray, RNA sequencing, and proteomics. These repositories provide access to data from a variety of species, tissues, and experimental conditions and are a valuable resource for scientists working in the fields of systems biology and personalized medicine. The data stored in these repositories can be used to investigate gene expression patterns, identify novel gene regulatory networks, and to understand the interplay between genes and the environment. Additionally, gene expression data repositories can be used to identify potential drug targets, compare gene expression profiles between different species, and to develop diagnostic and prognostic tests. Furthermore, the data is freely available and easily accessible, making it an ideal resource for researchers.

2.3.1 GEO

The Gene Expression Omnibus (GEO) [131] repository is a public repository for microarray and sequencing data. It is hosted by the National Center for Biotechnology Information (NCBI) and is funded by the National Institutes of Health (NIH). The GEO repository contains data from over 10,000 studies, including over 1.5 million gene expression measurements. It is searchable by gene name, probe set, or tissue type. The GEO repository is free to use and is open to submissions from any researcher.

2.3.2 Recount2

Recount2 [132] is an online resource consisting of RNA-seq gene and exon counts as well as coverage of bigWig files for 2041 different studies. For ease of statistical analysis, each study count data are created at the gene and exon levels and also extracted phenotype data, which are in raw formats as well as in RangedSummarizedExperiment R objects. The count tables, RangedSummarizeExperiment objects, phenotype tables, and mean bigWigs are ready to use and freely available here. By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, it makes finding and analyzing RNA-seq data

considerably more straightforward.

2.3.3 NCBI SRA

The NCBI is a public database of genetic and genomic information. The NCBI SRA (Sequence Read Archive) (www.ncbi.nlm.nih.gov/sra) is a repository of HTS data. The SRA contains raw sequencing data as well as processed data from a variety of sequencing platforms.

2.3.4 cancerSEA

The SEA (Single-cell Expression Analysis) database is an online resource for single-cell expression data available at http://biocc.hrbmu.edu.cn/CancerSEA/. It is a comprehensive database of single-cell expression data from a variety of cancer types, including breast, lung, and colorectal cancer. The database contains over 1,000 single-cell expression profiles from over 200 cancer samples. The data is organized into three categories: gene expression, gene copy number, and gene fusion.

2.4 Datasets Used

This section reports seven benchmark datasets of ESCC and other sources.

2.4.1 ESCC disease dataset: GSE20347

The ESCC dataset GSE20347 is a microarray gene expression dataset generated by the Affymetrix HG-U133A 2.0 gene expression arrays. RNA was extracted from 17 micro-dissected tumor and matched normal tissue pairs of 17 Esophageal Squamous Cell Carcinoma (ESCC) patients from a high-risk region of China. The dataset contains expression levels from 22278 genes in each sample. The goal of the study was to identify the gene expression patterns associated with ESCC, so that potential therapeutic targets could be identified. The dataset is freely available and can be accessed online and can be downloaded in various formats, including Excel and tab-delimited text. Additionally, the data can be accessed through the Gene Expression Omnibus (GEO) database.

2.4.2 ESCC disease dataset: GSE23400

The GEO GSE23400 ESCC microarray dataset is a publicly available dataset that was developed to study the gene expression profiles of ESCC. The dataset consists of gene expression data from 53 paired tumor and 53 normal tissue samples. The dataset contains a total of 22287 genes, with expression values for each gene in each sample. The dataset was generated using the Affymetrix U133A/B chip platform and is available through the Gene Expression Omnibus (GEO).

2.4.3 ESCC disease dataset: GSE32424

GSE32424 (GEO) or SRP008496 (SRA) RNA-seq dataset is generated by Illumina high-throughput sequencing. A total of 12 clinical samples from human ESCC (7 tumors and 5 non-tumors) are reported in the dataset and it contains a total of 58,037 gene profiles. The dataset is available at GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32424 and at Recount2 https://jhubiostatistics.shinyapps.io/recount/.

2.4.4 ESCC disease dataset: SRP064894

These samples consisted of 15 ESCC tissue samples and their 14 respective paired non-tumor tissues with 58000 genes. This count data for this dataset is available in recount2 with accession number SRP064894.

2.4.5 Yeast Sporulation dataset

The yeast sporulation microarray dataset quantifies expression levels of 6118 genes across 7-time points during meiosis and spore formation in Saccharomyces cerevisiae. A pre-processing was carried out by Sanghamitra et al.to exclude genes whose expression values did not change significantly using a threshold level of 1.6 for the root mean squares of the log2-transformed ratios. This resulted in a dataset of 474 genes over seven time points. The original dataset is available at http://cmgm.stanford.edu/pbrown/sporulation/ and the preprocessed dataset is available at http://anirbanmukhopadhyay.50webs.com/data.html.

2.4.6 Iris dataset

The iris dataset includes samples from three different Iris species (Iris versicolor, iris virginica, and iris setosa). The length and width of the sepals and petals, both in centimetres, are measured for each sample. The dataset is composed of four features, each with 50 instances, each referring to a different iris plant kind. This dataset can be downloaded from the UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/iris

2.4.7 ESCC disease dataset: GSE81812

The dataset scRNAseq GSE81812 is a collection of single cell transcriptomes for radio-resistance analysis of the esophageal squamous cell carcinoma (ESCC) cell line KYSE-180. It was generated by Next Generation Sequencing (NGS) and contains gene expression data from individual cells of the KYSE-180 cell line. The data includes information on the gene expression of each cell, as well as metadata such as cell type, radio-resistance, and treatment. Cultured KYSE-180 cells were subjected to accumulative irradiation doses of 0 Gy, 12 Gy or 30 Gy, respectively. Single-cell libraries were then generated using the Smart-seq 2 kit, and sequenced on an Illumina HiSeq 2500. The sequence reads that passed quality filters were analyzed for transcript expression levels with TopHat, followed by DESeq2/Monocle. The dataset is intended to be used for research into ESCC radio-resistance, providing insight into the molecular mechanisms underlying radio-resistance in this cancer type.

2.5 Software Tools Used

Different tools are employed in the various stages of this work. These tools are used for either the implementation of proposed techniques or for the implementation of existing techniques of a similar type. Additionally, some tools are employed to use an established validation framework to assess the results of the proposed techniques. The major tools used in this research are discussed next.

2.5.1 R

R [133] is a programming language and software environment for statistical computing and graphics. It is one of the most commonly used programming languages in data science and is a popular choice for statistical software with an integrated scripting language interface. R is an open-source language, meaning it is free to use and free to modify. It is known for its wide range of libraries and packages, which enable users to perform complex tasks quickly and easily. It also has a large community of users who are constantly developing and sharing new packages. R packages are collections of functions, data, and compiled code in a well-defined format. They extend the capabilities of R by providing additional functions, datasets, and tools that are often used by data scientists. R packages are typically created by developers and distributed through the Comprehensive R Archive Network (CRAN), a public repository of R packages. There are currently over 13,000 packages available on CRAN, providing a wide range of data analysis and statistical tools. R has a wide range of data types, including vectors, matrices, and data frames, and is able to manipulate them quickly and easily. It also has a wide range of graphical tools for creating plots, charts, and other visualizations. R is used for a variety of tasks, including data analysis, machine learning, and data visualization. It is also used for web development and can be used to create interactive web applications. R is an extremely powerful language and is used by researchers, data scientists, and software developers around the world. Its wide range of libraries and packages make it an ideal choice for those looking to quickly and easily manipulate and visualize data.

2.5.2 FuncAssociate

FuncAssociate (http://llama.mshri.on.ca/funcassociate/) is a web-based gene enrichment analysis tool designed for users to submit lists of genes or proteins. Using FuncAssociate, users can quickly search and analyze gene-disease associations to identify relationships between genes and diseases. It uses Fisher exact test and annotations from Gene Ontology to identify GO terms that significantly overlap with the provided list, providing the user with corresponding p values. This tool is designed to help researchers quickly identify potential gene-disease associations and explore existing relationships between genes and diseases. FuncAssociate uses a comprehensive set of data sources, including public gene-disease databases and literature-based information. It also integrates information from multiple sources, such as the Gene Ontology and the Human Phenotype Ontology,

to identify potential gene-disease associations. The tool also provides visualizations and insights into the relationships between genes and diseases, allowing users to quickly and easily explore the data and draw meaningful conclusions.

2.5.3 DAVID

DAVID (Database for Annotation, Visualization, and Integrated Discovery) [134] is an online bioinformatics resource available at https://david.ncifcrf.gov that provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. It helps to identify and categorize the biological themes and pathways that are associated with the input genes. It uses a powerful statistical algorithm to identify and assign gene functional categories from over 2000 available gene ontology categories. It also provides an interactive visualization interface to explore and interpret gene lists for biologists. It is used for many applications such as gene expression studies, gene set enrichment analysis, functional annotation clustering, and pathway analysis. DAVID is a free online resource and is widely used by researchers in the field of bioinformatics.

2.5.4 STRING

The STRING tool [135] (https://string-db.org/) is a powerful bioinformatics resource for exploring the functional relationships between proteins. It is a database of known and predicted protein-protein interactions that is based on scientific literature, as well as high-throughput experiments. The tool allows users to easily identify and visualize the interactions between proteins, and to explore the functional relationships among them. The tool also offers a variety of other features, such as the ability to create protein-protein interaction networks, to predict functional relationships between proteins, to assess the functional significance of specific pairs of proteins, and to identify potential drug targets. The STRING tool is an invaluable resource for those looking to better understand the complex relationships between proteins, and the overall dynamics of biological systems.

2.5.5 Cytoscape

Cytoscape [136] is an open-source bioinformatics software platform for visualizing molecular interaction networks and biological pathways. It is used for understanding the structure and dynamics of cellular processes and is widely used in systems biology research. Cytoscape offers a variety of features, including interactive visualization, automated layout and analysis. Cytoscape is an important tool for researchers studying the structure and dynamics of large-scale systems, such as protein-protein interaction networks, metabolic pathways, and disease pathways. It can be used to build and visualize networks from high-throughput experiments, such as gene expression data or protein-protein interactions. It can also be used to analyze the topology of the networks and detect patterns, trends, and subnetworks. In addition to visualizing and analyzing networks, Cytoscape can be used for a variety of other tasks, such as gene ontology enrichment analysis, gene set enrichment analysis, and pathway analysis. It also provides a rich set of plugins to extend its functionality.

2.5.6 GeneMania

GeneMania [137] is a bioinformatics tool designed to help researchers explore the relationships between genes and diseases. It is a web-based tool that allows users to search through gene, protein, and disease data from multiple databases. It is used to visualize, analyze, and interpret gene-disease relationships, and to quickly identify potential disease-causing genes. GeneMania is an interactive tool that allows users to explore the relationships between genes, proteins and diseases. It enables users to quickly search multiple databases and visualize the relationships between genes and diseases. The tool also provides information about known diseases, genes, and proteins, and allows the user to identify potential disease-causing genes. GeneMania is a useful tool for researchers in the field of bioinformatics. It provides users with an easy way to access and analyze large amounts of genetic and disease data. It is an invaluable tool for gene-disease research, and can help to quickly identify potential disease-causing genes.

2.5.7 GeneMalacard

GeneMalacard under GeneAnalytics https://ga.genecards.org website is a online tool for exploring gene-disease associations. It is a product of the GeneCards

Suite, a suite of software tools for exploring and analyzing human genes and gene-disease interactions. The GeneMalacard tool provides free access to a curated database of gene-disease associations from over 400 sources, including PubMed, OMIM, and ClinVar. It enables users to quickly and easily search for information on gene-disease associations, and provides interactive visualizations to help explore the data. The GeneMalacard tool allows users to search for gene-disease associations by gene symbol or disease name, or to browse associations by disease type. It also provides a powerful search engine, allowing users to filter results by type of association, type of evidence, and other criteria. The tool also provides interactive visualizations, allowing users to explore the data in a graphical format. In addition to providing access to curated gene-disease associations, GeneMalacard also allows users to explore related information, such as genetic variants associated with a disease, gene expression profiles, and more. It also allows users to export data in a variety of formats, making it easy to incorporate GeneMalacard data into other applications.

2.5.8 WGCNA

WGCNA (Weighted Gene Co-expression Network Analysis) [78] is a bioinformatics tool that can be used to identify relationships among genes and other molecular features. It helps to explore and analyze gene networks, allowing researchers to gain further insight into the biology of the system they are studying. The WGCNA package contains a number of modules, including network construction, network visualization, network-based gene selection, and module identification. WGCNA can be used in a variety of contexts, such as to identify networks of genes associated with a particular phenotype, to identify gene modules associated with a particular biological process, and to identify gene modules that are associated with a drug's response. WGCNA can also be used to identify candidate genes for further study, or to identify potential drug targets.

2.5.9 Python

Python [138] is an interpreted, high-level, general-purpose programming language. It was created by Guido van Rossum and first released in 1991. It is used for a wide variety of applications, from web development to software development. Python is known for its easy-to-read syntax and its ability to quickly solve complex problems. It also has a large standard library, which provides a wide range of useful

modules and functions. Python is a great choice for those just getting started with programming, as it is relatively easy to learn and understand. It also provides the flexibility to build applications with a variety of features and functions. Python is widely used for web development, scripting, game development, artificial intelligence, and scientific computing. Pythons syntax is designed to be intuitive and straightforward, allowing developers to write code quickly and efficiently. It also supports object-oriented programming, which allows developers to organize their code into logical blocks. This makes it easier to maintain and debug code.

2.6 Statistical and Biological Evaluation

2.6.1 Gene Enrichment Analysis

Gene Ontology (GO) is a standard convention for defining vocabulary terms associated with mainly genes [139]. The term annotation is used to describe associations between genes and available biological terms for various organisms. Biological terms in GO may correspond to cellular components, biological processes, or molecular functions. Thus, annotation terms are the main contents of the GO repository. Members of the GO Consortium are responsible for improving this knowledge base by submitting their findings for integration into the existing GO data. Associated P-values indicate how well a set of genes fits into various GO categories. The P-value is computed using the hypergeometric test or Fisher's Exact Test [140]. The Q-value for a particular gene is defined as the proportion of false-positive among all genes that are as or more differentially expressed [141]. It is nothing but a minimum False Discovery Rate (FDR) at which this gene appears biologically significant. The GO categories and Q-values from an FDR corrected hypergeometric test for enrichment can obtain using a tool called GeneMANIA [142] which has a web interface. P-values and Q-values are estimated using the Benjamini Hochberg procedure. In Table 2.2, a list of GO enrichment analysis tools is reported.

2.6.2 Pathway Analysis

Genes associated with a particular disease have a very high probability of being functionally connected within the processes or pathways associated with the corresponding disease. As a result, pathway analysis is required to validate the results.

Table 2.2: List of available GO Enrichment Analysis Tools

Tool	Type of software	Availability
g:Profiler[143]	web-based	biit.cs.ut.ee/gprofiler/gost
GSEA[144]	stand-alone	software.broadinstitute.org/gsea/index.jsp
Gonet[145]	web-based	tools.dice-database.org/GOnet/
GeneCodis[146]	web-based	genecodis.cnb.csic.es/analysis
KAAS[147]	web-based	www.genome.jp/tools/kaas/
KEGG[148]	web-based	www.genome.jp/kegg/
Enrichr[149]	web-based	amp.pharm.mssm.edu/Enrichr
FunRich[150]	stand-alone	funrich.org/download
Geneshot[151]	web-based	amp.pharm.mssm.edu/geneshot/
ShinyGO v0.61[152]	web-based	bioinformatics.sdstate.edu/go/
DAVID[134]	web-based	david.ncifcrf.gov/
FuncAssociate[140]	web-based	llama.mshri.on.ca/funcassociate/
GOrilla[153]	web-based	cbl-gorilla.cs.technion.ac.il/
MalaCards[154]	web-based	www.malacards.org/
GOMA[155]	stand-alone	goma.sel.is.ocha.ac.jp/
WebGestalt[156]	web-based	www.webgestalt.org/
pathfindR[157]	R package	cran.r-project.org/web/packages/pathfindR/index.html
WebGIVI[158]	web-based	raven.anr.udel.edu/webgivi/
ViSEAGO[159]	R package	bioconductor.org/packages/release/bioc/html/ViSEAGO.html
GO TOOLS	web-based	go.princeton.edu/
topGO[160]	R package	bioconductor.org/packages/release/bioc/html/topGO.html

Pathway analysis is a type of data analysis that looks at the interactions between different pathways in a system. It is used to study the relationships between different components of a system, such as the interactions between genes, proteins, and other molecules. By looking at the interactions between these components, researchers are able to gain insight into how a system works and the role that each component plays in the overall functioning of the system. Pathway analysis can provide valuable insights into the functioning of a biological system. It can help researchers identify pathways that are involved in a particular disease or condition, as well as pathways that can be targeted for therapeutic intervention. Additionally, pathway analysis can be used to assess the effects of environmental or genetic factors on the functioning of a system. This can be helpful in understanding how a particular disease or condition is caused or prevented.

2.6.3 Topological Analysis

Topological investigation of the discovered modules could lead to the finding of causal genes that are not otherwise linked to the disease. In a well-connected graph, two genes, say, G1 and G2, possess a strong association, and G1 might already be known to be a causative gene. In such a case, there's a good chance that G2 is a causal gene. Topological analysis is a useful tool for understanding gene expression data, as it can reveal patterns that may not be apparent from the raw data. Furthermore, it can provide insights into the relationships between genes, which can be used to develop new hypothesis and further our understanding of gene expression and its role in disease.

2.6.4 Literature Mining/Evidence

In the process of validating experimental results, researchers often turn to external sources or validated literature, a method known as literature mining [161]. This approach is particularly useful for confirming findings obtained through downstream analysis. For instance, if a study experimentally establishes that a specific gene is causally linked to a particular disease, literature mining can be employed to explore published wet-lab results that support the notion that this gene indeed has a high potential to be a causal factor for the disease. Investigating such types of well-established facts related to the concerned biological question can further strengthen the biological and topological validation approaches [162]. As an example, literature mining may involve reviewing relevant scientific articles, databases, or authoritative sources that corroborate the identified gene's role in the context of the disease. Such cross-referencing not only reinforces the validity of the experimental findings but also contributes to a more comprehensive understanding of the gene-disease association [163].

2.7 Discussions

In this chapter, a background of list of tools, repositories, and validation measures have been reported which have been used in subsequent contributory chapters towards identification of crucial genes for ESCC. It delves into the background of gene expression data analysis, covering techniques such as Microarray, RNA-seq, and scRNA-seq, along with discussions on ESCC disease, biomarker identification,

and pre-processing techniques for different types of gene expression datasets. The chapter emphasizes the use of data mining in these analyses, exploring various biclustering techniques, co-expression network analysis, and similarity measures for gene expression data. The research focuses on designing effective data mining techniques to address challenges in analyzing gene expression data and explores various tools and methods for the identification of crucial genes related to ESCC. In the next chapter, I introduce a framework called PD_BiBIM for biomarker identification of ESCC disease, which is based on biclustering method. Further, gene networks are constructed for topological, pathway and causal analysis. This method is found successful to identify potential biomarkers for ESCC disease.