Chapter 2

Background

2.1 Statistical Methods And Measures

Statistical methods and measures are essential for gene expression data analysis. A mixture of various statistical tests such as t-test [761, 615], moderated t-test [637], are employed to find differentially expressed genes (DEGs) while multiple testing correction methods such as Benjamini-Hochberg [43, 764] are used to regulate the false discovery rate (FDR). Fold change [443] is a simple method to measure the relative changes in gene expression level between two conditions and is used in conjunction with p-values [166] to identify biologically and statistically significant genes. Clustering methods such as k-means clustering [444] and hierarchical clustering [756, 291] are used in gene expression analysis to find groups of genes or samples. By lowering their complexity while keeping important links, dimensionality reduction techniques like principal component analysis (PCA) [546, 292] aid in the visualization and exploration of high-dimensional gene expression data. Various statistical measures and methods have been employed throughout various stages of this research.

2.1.1 *p-value*

p-value [166] is a measure frequently used to draw a conclusion about a population of samples or evaluate the statistical significance of an observed outcome. To evaluate the strength of evidence for or against a hypothesis, *p-value* is quite useful. With the assumption that the null hypothesis is true, *p-value* gives the likelihood of observing a test statistic that is as extreme or more extreme than the observed value. Following are the key points for *p-value* statistics.

(a) The probability value for the *p-value* ranges from 0 to 1. It measures the likelihood that the data or a more extreme result will be observed if the null hypothesis is accepted.

(b) The null hypothesis, H_0, is a declaration that there is no relationship, difference, or effect between the variables. The *p-value* aids in evaluating the opposing evidence to the null hypothesis.

(c) The significance level, α, is a pre-determined cutoff that is set to assess the amount of evidence needed to reject the null hypothesis, H_0, the most common significance values being 0.05 (5%) and 0.01 (1%).

(d) It is deemed statistically significant if the *p-value* $< \alpha$. As a result, the null hypothesis, H_0, is rejected in favor of the alternative hypothesis, showing that the observed result is unlikely to have happened by chance alone. In contrast, if the *p-value* not greater than α, the result is not statistically significant and the null hypothesis, H_0 is accepted.

(e) Based on the research problem and the kind of the hypothesis being evaluated a one-tailed or two-tailed test is chosen In a two-tailed test, the *p-value* takes into account extreme values in both distribution tails. In a one-tailed test, just one tail's extreme values are taken into account. The decision has an impact on how the *p-value* is interpreted and where the key area is for rejecting the null hypothesis, H_0.

In other words, *p-value* merely conveys the degree of support for the null hypothesis, H_0. *p-values* are susceptible to restrictions and inherent presumptions of the statistical test being employed. Larger sample sizes leads to more accurate estimates of the *p-value*.

2.1.2 False Discovery Rate

When testing several hypotheses, the False Discovery Rate (FDR) [43] statistic is used to determine the percentage of false positives among a set of significant results. When performing large-scale studies that simultaneously test several hypotheses, FDR is especially important. The key points behind FDR statistics are listed below.

(a) Multiple testing problem occurs when the significance thresholds, α, are adjusted inaccurately due to the simultaneous testing of multiple hypotheses or variables. This leads to a higher risk of discovering false positives simply by chance.

(b) While the *p-values* regulate the Type I error (false positives) rate for each individual hypothesis test, FDR statistics regulates the expected fraction of false positives among the hypotheses that are deemed statistically significant.

(c) FDR is the anticipated percentage of false positives among all the statistically significant results.

(d) The multiple testing issue has been addressed and several techniques have been devised to regulate the FDR. The Benjamini-Hochberg [43, 764] approach, which offers a step-up procedure to alter the significance levels for multiple comparisons, is one often employed technique. It evaluates the *p-values* to a crucial value, ranks them, and then modifies the significance levels accordingly.

(e) A decreased percentage of false positives among the significant results is indicated by a lower FDR value. A desirable FDR threshold (e.g., 0.05 or 0.01) is selected to assess the degree of confidence in the discoveries.

The underlying accuracy and inherent assumptions predicate the FDR statistics, and thus sample sizes, distribution of the data, and the correlation between the hypotheses being tested can have an impact on the accuracy of the FDR statistics. Although FDR relies on accurate interpretation of the results, it provides a valuable framework for regulating the anticipated number of false positives, thus enabling strong and reliable statistical tests.

2.1.3 *q-value*

FDR statistics [43], commonly referred to as *q-value* statistics [645], regulates the rate of false positives (type I errors) over multiple hypotheses, while *p-values* regulates the rate of false positives on a single hypothesis. The key points behind *q-value* statistics are listed below.

(a) *q-value* is an estimation of the FDR which defines the percentage of null hypothesis that were incorrectly rejected. The lowest FDR at which the test is deemed significant is the *q-value*.

(b) To adjust the significance threshold for *q-value*, the *p-values* are compared to a key threshold and sorted in ascending order.

(c) With the decrease in *q-value*, the likelihood that the observed result is statistically significant (i.e, true positive) increases.

q-value enables the management of FDR , thus offering a more robust and conservative solution to multiple testing as compared to *p-value* that primarily focuses on statistical significance of individual genes. The underlying accuracy and inherent assumptions predicate the FDR statistics, and thus sample sizes, distribution of the data, and the correlation between the hypotheses being tested can have an impact on the accuracy of the *q-value* estimations.

2.1.4 Local False Discovery Rate

Local false discovery rate (*lFDR*) statistics evaluates the proportion of false positives within the set of individual hypotheses. *lFDR* is also referred to as empirical Bayes methods and gives a local measurement for FDR that takes into account the properties unique to the data being examined. The following are the key points for *lFDR* statistics.

(a) Instead of regulating the overall FDR across numerous hypotheses, *lFDR* concentrates on the significance evaluation of specific hypotheses and forgoes the assumption that true null hypotheses are uniformly distributed. *lFDR* calculates the local FDR for each null hypothesis while considering the heterogeneity of the data.

(b) *lFDR* employs empirical Bayes framework and prior knowledge from the data.

(c) *lFDR* increases sensitivity to identify both strong and weak signals by concentrating on measuring the FDR for each distinct hypothesis. Global FDR statistics, on the other hand simultaneously regulate the overall false discovery rate across all hypotheses.

For a given hypothesis, a lower *lFDR* signifies a reduced percentage of false positives. Under certain circumstances *lFDR* more precise estimates when compared to FDR. *lFDR* makes certain assumptions about the distribution of the data and as such the quality and heterogeneity of the data can have an impact on *lFDR* estimates. It is crucial to verify the findings and take the statistical model's constraints into account.

2.1.5 Heirarchical Clustering

Hierarchical clustering [756, 291, 47] groups similar data points into clusters based on their pairwise similarity or dissimilarity and represents the relationships among the data points as a hierarchical structure known as a dendrogram. Agglomerative (bottom-up) and divisive (top-down) hierarchical clustering are the two primary approaches.

Agglomerative hierarchical clustering starts by treating each data point as a separate cluster. The similarity and dissimilarity between all pairs of clusters is determined using a distance metric and the clusters that are most comparable are combined into one cluster. The similarity or the dissimilarity of the new clusters and the other remaining clusters are repeated. The process of merging and recalculation is repeated iteratively until either the estimated number of clusters is attained or each data point belongs to a single cluster. The resulting dendrogram represents the hierarchical links of the clusters as well as their merging order. For divisive hierarchical clustering, the initial cluster consists of all the data points. The similarity or dissimilarity of the data points within the cluster is determined, and a data point or a group of data points is chosen that divides the cluster into two new clusters. The similarity or dissimilarity within the new clusters is recalculated. The splitting and recalculation process is repeated until either the estimated number of clusters is attained or each data point belongs to a single cluster. The resulting dendrogram represents the hierarchical links of the clusters as well as their splitting order.

The choice of distance measure and linking criterion is critical for both approaches. The cluster structure that is produced depends on the linking criterion, which also controls how similarity or dissimilarity between clusters is calculated. Single-linkage criteria (minimum distance between any two data points in separate clusters), complete-linkage criteria (maximum distance), and average-linkage criteria (average distance) are common linkage criteria. The number of clusters should not be specified in advance when using hierarchical clustering, which also offers a hierarchical representation that enables exploration at various granularity levels. The dendrogram also makes links and clusters between data points graphically apparent. Hierarchical clustering is sensitive to the choice of distance metric and linking criterion, and it can be computationally demanding, especially for big datasets. It might also have trouble scaling to highly dimensional data.

2.1.6 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [546, 292] is a statistical method that is used for exploration of the data, extraction of features and reduction of the dimensionality of the data. PCA seeks to change a group of correlated variables into a group of variables that have no correlation. Following gives a summary of the key steps involved in PCA.

(a) To ensure that each variable in the dataset has similar weights PCA standardizes the variables by removing the mean and dividing by the normal deviation.

(b) To analyze the connections between two variables with correlations, the standardized covariance matrix is determined while eigen analysis methods are employed to determine the eigenvectors and eigenvalues of the covariance matrix.

(c) The principal components (PCs) are represented by the eigenvectors, and the variance explained by each component is quantified by the corresponding eigenvalues.

(d) The eigenvalues are arranged in descending order with the highest eigenvalues corresponding to the PC that accounts for the majority of the variations in the data. A cumulative variance plot is frequently employed for the choice of the number of PCs to be retained.

(e) For the creation of a matrix of the chosen PCs the original standardized data are multiplied by the eigenvector matrix. Each PC also represents a linear combination of the original variables.

PCA can aid in identifying crucial variables, removing background noise, and identifying clusters in the data that exhibit hidden patterns. PCA makes the assumption that the data is linear, and as such, it is not appropriate for non-linear datasets. Pre-processing and removal of outliers is a crucial step before PCA, as outliers can significantly impair PCA.

2.1.7 Benjamini-Hochberg

Benjamini-Hochberg [43, 764] is a statistical method to manage the FDR (Section 2.1.2) in multiple hypothesis testing. It aims to limit the percentage of false positives that arise from running several statistical tests simultaneously. The Benjamini-Hochberg approach is described below.

- For each hypothesis test, the *p-value* associated with strength of evidence against the null hypothesis is calculated (Section 2.1.1). This is then followed by the listing of the *p-value* in ascending order.
- Thresholds for significance are established based on the required FDR.
- The largest *p-value* that is smaller than or equal to the crucial value corresponding to its rank (k) in the sorted list of *p-values* is identified and denoted by the notation $p(k)$.
- The *p-values* $\leq p(k)$ result in the rejection of all corresponding null hypotheses and these findings are deemed as statistically significant.
- The *p-values* of the rejected hypotheses. The step-up approach outlined by Benjamini-Hochberg is frequently employed to adjust the *p-values* of the rejected null hypotheses (false positives) with the aim to control the FDR. All hypotheses that are found significant after the adjustment are deemed statistically significant.

2.1.8 T-test

A t-test [761, 615] is a statistical method used to compare the means of two groups and determine if there is a significant difference between them. The key points behind the t-test are listed below.

(a) With the assumption that a) the observations within each group are independent of each other, b) the data within each group are approximately normally distributed, and c) the variances of the two groups are approximately equal, the two hypotheses are formulated. The null hypothesis (H_0) states that the means of the two groups are not significantly different from one another. In contrast, the alternative hypothesis (H_α) states that the means of the two groups are significantly different from one another.

(b) Depending on the presence or absence of correlation among the two groups being compared, paired t-test and independent samples t-test are utilized.

(c) The t-statistic is calculated for independent samples t-test using the following formula.

$$t = \frac{mean_1 - mean_2}{\sqrt{s_1/n_1 + s_2/n_2}} \tag{2.1}$$

where, s_1 and s_2 are the sample standard deviations, n_1 and n_2 are the sample sizes, and $mean_1$ and $mean_2$ are the means of the two groups. The formula for the paired samples t-test is as follows.

$$t = \frac{\text{mean of the differences}}{\text{standard deviation of the differences}/\sqrt{n}} \tag{2.2}$$

where, n is the number of paired observations and the standard deviation of the differences is the standard deviation of the paired differences.

(d) The degrees of freedom for the independent samples t-test is calculated using $df = (n1 + n2) - 2$, where $n1$ and $n2$ are the sample sizes of the two groups. The degrees of freedom for the paired samples t-test is equal to the number of paired observations minus 1.

(e) To determine the crucial value corresponding to the critical value of significance (e.g., $\alpha = 0.05$), a t-distribution table is consulted. Alternately, the t-distribution is calculated to determine the *p-value* associated with the t-statistic.

(f) If the t-statistics $\geq$ crucial value or *p-value* $\leq \alpha$, the null hypothesis, H_0, is rejected and it can be established that the means of the two groups are significantly different from one another (alternate hypothesis, H_α). Otherwise, the null hypothesis, H_0 is accepted and it can be established that the means of the two groups are not significantly different from one another.

2.1.9 Preservation Analysis

Module preservation analysis is a statistical analysis method to evaluate the module or cluster preservation across various datasets or conditions. Module preservation analysis is frequently used in network analysis as well as biclustering approach. Through module preservation analysis, the conservation of the modular structure seen in one condition or dataset (reference) is ascertained in another condition or dataset (test). In other words, with the assumption that control condition is the reference dataset and disease condition is the test dataset, preservation analysis ascertains the conservation of control modules in the test dataset. Similarity of preservation of modules is quantifies so as to determine whether the observed preservation is substantial.

With the $Z_{summary}$ statistic [327, 329], a summary statistic is computed, such as the mean or median of connectivity among modules, and it is compared to the null distribution created by permuting the module labels. The observed module structure's preservation versus random expectations is shown by the $Z_{summary}$ statistic. The module labels are permuted across samples in permutation-based testing, and the preservation statistics are assessed for each permutation. Statistical significance is then determined by comparing the observed preservation measure to the null distribution. By randomly selecting subsets of the original data, consensus module preservation [329, 328, 574] includes creating numerous datasets and computing preservation statistics for each subset. After combining the outcomes from various iterations, the consensus statistic is calculated, giving a more reliable estimate of module preservation.

2.1.10 Z-summary Statistics

The $Z_{summary}$ statistic [327, 329] is a statistical method that is used in module preservation analysis to assess how well modules or clusters are preserved across various datasets or conditions. To determine the statistical significance of the preservation it quantifies the conservation and similarity of the module structure between two conditions or datasets. The following discusses the key points of the Z-summary statistic.

(a) To measure how well the test dataset preserves the observed modular structures of the reference dataset, the module labels are permuted resulting in the creation of a null distribution by the $Z_{summary}$ statistic. This is then followed by comparison to the summary measure of within-module connectivity in the reference dataset.

(b) By random permutation of the module labels across samples in the test dataset the original modular structures are disrupted and the observed module preservation is ascertained. Through computation of $Z_{summary}$ statistic for each permutation , the observed statistic is compared to the null distribution.

(c) Typically, the within-module connectivity of the reference dataset is compared to the null distribution to compute the $Z_{summary}$ statistic. It shows how many standard deviations the observed within-module connection deviates from the null distribution's mean.

(d) While high positive $Z_{summary}$ statistic can be interpreted as the preservation of module structures between the two conditions or datasets, a low or negative $Z_{summary}$ statistic, on the other hand, can be interpreted as poor preservation of the modules structures between the conditions and datasets.

(e) The $Z_{summary}$ statistic is compared to the null distribution to ascertain its statistical significance. If the observed $Z_{summary}$ value exceeds a particular percentile of the null distribution or falls outside the range of values predicted by chance alone, it is deemed statistically significant and suggests that the module structure has changed or been preserved.

2.2 Gene Expression Analysis Programming/Tools

Various platforms and tools are used at various stages of this research. In essence, these tools are employed to implement proposed frameworks or to evaluate the results. This section discusses some of the key platforms and tools used in this research.

2.2.1 R

R [68] [1] is an open-source software environment and programming language made primarily for statistical computing and graphics. It offers an extensive selection of tools and libraries for manipulating, analyzing, visualizing, and modeling data. Statistical experts, data scientists, and researchers have embraced R because of its robust features

[1] https://www.r-project.org

and active user base. Some of the key features of R are discussed below.

(a) Numerous tools in R make it simple to do effective data wrangling operations like filtering, converting, summarizing, and merging datasets.

(b) Descriptive statistics, hypothesis testing, regression analysis, time series analysis, survival analysis, and other statistical analysis tools are all included in the extensive set of functions and packages that R offers.

(c) For producing excellent visualizations, such as scatter plots, bar charts, histograms, box plots, heat-maps, and interactive visualizations, R includes robust libraries.

(d) Machine learning methods are well supported in R and the tasks of classification, regression, clustering, and dimensionality reduction can be carried out by users.

(e) R provides tools to help with reproducible research and with the help of which users may produce dynamic documents that incorporate code, analysis, visualizations, and narratives, making it simple to share research findings.

(f) Users of R may build customized plots and graphs of publication quality using the language's built-in graphics features and packages which provide granular control over aesthetics, making data discovery and presentation easier.

(g) Because of R's extensibility, users can create custom functions and packages to expand their functionality and meet particular needs. Access to thousands of extra functionalities is made possible by the enormous repository of user-contributed packages.

(h) Other programming languages like Python, C++, and Java can easily be integrated with R. Data interchange, interoperability, and utilizing the advantages of many programming languages are all made possible.

A complete and adaptable environment is offered by R for data analysis, statistical modeling, and visualization. It is a well-liked solution for statistical computing and data analysis jobs due to its huge package ecosystem, interactive development environment (IDE) options like RStudio [2], and rich collection of tools.

2.2.2 Bioconductor

For the analysis and comprehension of high-throughput genomic data, Bioconductor [186] [3] is a widely used open-source software project and platform. Various types of

[2] https://posit.co/products/open-source/rstudio/

[3] https://www.bioconductor.org/

gene expression data can be processed, analyzed, and interpreted with the help of this tool. Key features and components of Bioconductor are discussed below.

(a) For the analysis of genetic data, Bioconductor offers a huge selection of packages (software libraries). Numerous activities, including data pre-processing, quality assurance, statistical analysis, visualization, and functional annotation, are covered by these packages. Within the R programming environment, users can install and use these packages.

(b) The R programming language, a potent and well-known statistical computing and graphics environment, serves as the foundation for Bioconductor. Users can take advantage of R's vast ecosystem of statistical and data manipulation tools as well as the specialized genomics packages offered by Bioconductor thanks to its seamless integration with R.

(c) To effectively handle and manipulate genomic data, Bioconductor includes specialized data structures and classes. These consist of classes that can display information about gene expression, genomic coordinates, DNA sequences, protein structures, and other things. Users can easily execute intricate computations on genetic data thanks to these frameworks.

(d) Comprehensive annotation resources, like as packages for different genomes, are available through Bioconductor. These sites offer details on the symbols for genes, genomic coordinates, functional annotations, pathway data, and other pertinent metadata for genes, transcripts, and genomic areas.

(e) Workflows and pipeline frameworks are provided by Bioconductor to help users through the analysis process. These workflows provide best practices and step-by-step guidance for particular analysis activities, improving reproducibility and making it easier to adopt common analysis techniques.

2.2.3 DAVID

A popular bioinformatics tool called DAVID (Database for Annotation, Visualization, and Integrated Discovery) [628, 253] [4] enrichment analysis aids researchers in understanding lengthy gene lists by highlighting biological themes and functional annotations connected to the genes. It helps users to comprehend the underlying biological processes, pathways, and molecular activities represented within those gene sets and ac-

[4] https://david.ncifcrf.gov

quire insights into the biological importance of those gene sets. Following discusses the key features of DAVID enrichment analysis tool.

(a) Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), pathways, protein domains, disease associations, and many more biological and functional annotation databases are integrated by DAVID. Functional annotations are assigned to genes, and enrichment analysis is performed utilizing these annotations.

(b) DAVID users can enter lists of interesting genes, such as genes with differential expression or genes connected to a particular trait or experimental situation. Both the statistical significance of the enrichment and the over-representation of functional words or annotations in these gene lists are evaluated.

(c) DAVID offers a number of statistical methods for enrichment analysis, including the gene set enrichment analysis (GSEA) [651]. By comparing the input gene list to the background set, these approaches identify functional categories that are disproportionately over-represented.

(d) Based on how semantically similar related phrases or annotations are, DAVID employs a clustering approach to group them together. The clustering, which helps to arrange and summarize the enhanced functional words, makes it easier to comprehend and visualize the results.

(e) DAVID provides interactive visualizations such as bar plots, pie charts, and heat maps to represent richer functional concepts and their relationships. These visualizations, which also aid users in understanding the enrichment findings, can be used to identify the most pertinent and significant functional categories.

(f) DAVID offers further tools and functionalities so that users can analyze the functional annotations in more detail. Users can categorize genes based on their functional properties, browse gene lists connected to particular terms, connect enriched terms to external databases for extra data, and even view functional annotation networks.

(g) It is simple for DAVID users to combine their results with those from other bioinformatics tools and resources. It provides options for exporting the results in various formats for additional analysis or external software-based visualization.

The DAVID enrichment analysis tool is a helpful tool that helps scientists working with gene lists understand vast volumes of biological data. It aids users in understanding the biological meaning and functional context of their gene sets by providing users with insights on the underlying biological processes and pathways linked to the genes of interest.

2.2.4 GENIE3

A computational approach called GENIE3 (Gene Network Inference with Ensemble of Trees) [264] [5] is used to infer gene regulatory networks from data on gene expression. By utilizing gene expression data, GENIE3 tries to identify the regulatory connections between genes. It seeks to reconstruct the underlying gene regulatory network, which depicts how genes interact and regulate the levels of their expression.

To determine the regulatory relationships, GENIE3 uses a collection of regression trees. The two stages of the algorithm's operation are feature selection and network creation. a) GENIE3 evaluates each gene's importance as a possible regulator for each target gene (TG) in this stage. Using tree-based techniques, it evaluates how well a gene's expression profile predicts the expression of the TG. b) GENIE3 builds the gene regulatory network by giving regulatory weights to the edges linking the regulators and their TG after it has determined the pertinent regulators for each TG. The weights are determined by how crucial the regulators are for foretelling the expression of the TG.

GENIE3 is suited for analyzing intricate regulatory interactions because it can capture both linear and nonlinear links between genes.The technique can handle massive gene expression datasets with thousands of genes and samples and is scalable. Each TG's possible regulators are ranked by GENIE3, which aids in determining the network's most significant regulators. To increase the precision and resilience of the estimated network, GENIE3 can combine diverse datasets, such as gene expression profiles under various experimental circumstances or perturbations. In benchmark experiments, GENIE3 was found to perform better than other available techniques for inferring gene regulatory networks. It captures known regulatory linkages with reasonable accuracy and offers insightful information about the regulatory structure of biological systems.

2.3 Knowledge Repositories

Biological knowledge bases, often referred to as biological databases, are collections of structured information and data pertaining to diverse biological study topics, such as

[5] https://bioconductor.org/packages/GENIE3/

genes, proteins, pathways, diseases, and species. These databases are often made to be searchable and interoperable, enabling researchers to access and integrate data from various sources for a variety of purposes, including hypothesis creation, experimental design, and data analysis. Based on the kind of information they include and their breadth, biological knowledge bases can be divided into a number of categories.

(a) Genomic and sequence databases: In addition to information regarding sequence variation, such as single nucleotide polymorphisms (SNPs), these databases also contain information about DNA and RNA sequences, including the sequence itself, annotations of genes, and functional elements.

(b) Protein databases: These databases include data on proteins' sequence, structure, interactions, and post-translational modifications and variants, as well as information on their function and relationships.

(c) Pathway and network databases: These databases include details on gene regulatory networks, including transcriptional and post-transcriptional networks, as well as biological processes, including metabolic and signalling pathways.

(d) Disease and phenotype databases: These databases include data on phenotypes, including traits, relationships, and genetic underpinnings, as well as data on diseases, including clinical characteristics, genetics, and treatments.

(e) Taxonomy and organism databases: In addition to information about biodiversity, such as species distributions and genetic diversity, these databases also contain information about organisms, such as their taxonomy, evolutionary relationships, and features.

The most important sources for biological research are biological knowledge bases, which offer a plethora of information and data that can be utilized to develop new hypotheses, test old ones, and acquire understanding of biological systems and processes. Many of these databases are freely accessible and available online, and the scientific community maintains and updates them to guarantee their relevance and accuracy.

2.3.1 Entity Identifiers

Entity identifiers are specific labels or codes that are given to various biological entities in genomics and transcriptomics in order to aid in their identification, organisation, and retrieval. These identities are essential for the integration, annotation, and exchange of data among different databases and research projects. Here are a few entity identifiers that are frequently used in genomics and transcriptomics.

- Gene Identifiers: Gene identifiers are special labels that have been given to certain genes in a genome. Entrez, Ensembl, RefSeq, and UniProtKB Gene IDs are a few examples. Through the use of gene identifiers, genes can be precisely and consistently referenced in various databases and analysis.
- Transcript Identifiers: Generated from genes, RNA transcripts are given distinctive labels called transcript IDs. Transcript IDs from Ensembl, RefSeq, and UniProtKB are a few examples. With the help of transcript identifiers, individual transcript isoforms or variations can be accurately identified and analyzed.
- Protein Identifiers: Protein products that have been encoded by genes are given distinctive names called protein IDs. UniProtKB Accession numbers, RefSeq Protein IDs, and Ensembl Protein IDs are a few examples. The identification and study of proteins as well as their properties are made easier by protein IDs.
- Variant Identifiers: The identification of particular genetic variants, such as SNPs or structural differences, is done using variant identifiers. Examples include ClinVar Variation ID, COSMIC ID, and dbSNP ID. The tracking and study of genetic variations in populations and illnesses is made possible by variant identifiers.
- Probe Identifiers: The distinct labels given to each probe on a microarray are known as probe identifiers. Affymetrix Probe ID, Agilent Probe ID, and Illumina Probe ID are a few examples. The identification and interpretation of gene expression data from microarray experiments is made possible by probe identifiers.
- Sequence Identifiers: Unique labels known as sequence IDs are applied to certain DNA or RNA sequences. GenBank accession numbers and FASTA sequence numbers are two examples. The retrieval and analysis of particular sequences in databases is made easier by sequence identifiers.

These entity identifiers provide standardized and unique labels for genes, transcripts, proteins, variants, probes, and sequences in genomics and transcriptomics research. They play a vital role in data integration, interoperability, and communication among researchers and databases, enabling accurate identification, comparison, and annotation of biological entities across different studies and resources.

Gene names and symbols, usually referred to as gene symbols or gene names, are frequently used names for genes that are readable by humans. Gene symbols, which are frequently taken from the official gene nomenclature criteria, give genes a clear and recognizable representation. Gene symbols, which are frequently used to refer to genes in research publications, databases, and annotations, are typically made up of uppercase letters and digits (for example, TP53, BRCA1, and EGFR). Gene symbols may be shared by several genes in various species, therefore they are not always unique. To ensure accurate gene identification in these circumstances, additional identifiers are needed.

Entity Identifiers are an integral part of interesting gene expression analysis results and their biological relevance against knowledge repositories. In the process of validating the results presented by each proposed framework we employ the following knowledge repositories.

2.3.2 Gene Ontology (GO)

The widely used Gene Ontology (GO) [30] standard vocabulary offers a common language for describing the biological processes, functions, and cellular components of genes and proteins. GO classifies concepts and their links into three basic groups: biological process, molecular function, and cellular component. Access to GO keywords and annotations can be found in the following databases.

- Gene Ontology Consortium: The GO database is created and maintained by the Gene Ontology Consortium (GOC) [105] [6], which is the primary source for GO annotations. The GOC compiles and maintains GO annotations from a variety of sources, including direct inputs from researchers, computational techniques, and literature curation.
- UniProt:A comprehensive database of protein sequences and functional details is called UniProt [1] [7]. Based on a review of the literature and computer projections, UniProt offers GO annotations for proteins.
- Ensembl: Ensembl [489] [8] is a database of genomic annotations that offers thorough annotations for a variety of species. Based on data gathered through the curation of literature, homology, and computational techniques, Ensembl gives GO annotations for genes and proteins.
- DAVID: DAVID (Database for Annotation, Visualization and Integrated Discovery) [628, 253] [9] offers functional annotation and enrichment analysis for genes and proteins. GO annotations from different sources are one of the gene annotation resources

[6] http://geneontology.org/

[7] https://www.uniprot.org/

[8] http://asia.ensembl.org/

[9] https://david.ncifcrf.gov

that DAVID includes.

- PANTHER: PANTHER (Protein ANalysis THrough Evolutionary Relationships) [682, 681] [10] is an extensive tool for classifying and analyzing proteins. PANTHER gives GO annotations for proteins based on evidence from homology, phylogenetic analysis, and the curation of relevant literature.

2.3.3 Kyoto Encyclopedia of Genes and Genomesm(KEGG)

A comprehensive database of biological pathways, networks, and functional annotations is called Kyoto Encyclopedia of Genes and Genomes (KEGG) [297]. The molecular interaction and reaction networks of cells, as well as metabolic pathways and disease-related pathways, are all covered in great detail by KEGG. There are three primary parts of KEGG.

- KEGG Pathway: In-depth details on metabolic, signaling, and other biological pathways in many organisms are provided in this component. Maps and diagrams are just a couple of the visual representations of pathways offered by KEGG Pathway.
- KEGG BRITE: This component includes functional annotations for genes and proteins, which may include details on cellular elements, molecular processes, and biological processes. Additionally, KEGG BRITE offers hierarchical categories for biological processes and pathways.
- KEGG Orthology: The evolutionary connections between genes and proteins in various animals are discussed in this component. According to their orthologous relationships, KEGG Orthology assigns K numbers to genes and proteins, enabling comparisons of function and regulation among other animals.

In order to understand how genes and proteins are regulated and function in many animals and biological processes, researchers in the fields of bioinformatics, genomics, and systems biology frequently consult the KEGG database.

2.3.4 Hgu133plus2.db

A database file called hgu133plus2.db [235] [11] is connected to the GeneChip Human Genome U133 Plus 2.0 Array, a microarray device for gene expression profiling. The database file includes details about the many probesets, or short DNA sequences, that make up the array and are specific to particular genes or gene areas. The probesets' an-

[10] http://www.pantherdb.org/

[11] https://bioconductor.org/packages/hgu133plus2.db/

notation data, including the gene symbols, gene descriptions, and other pertinent details, are provided in the hgu133plus2.db file. It enables the interpretation of the outcomes of gene expression tests carried out using the U133 Plus 2.0 Array and the mapping of the probe sets to specific genes. This database file is frequently used in bioinformatics analysis and is compatible with a number of software programmes and coding languages, including R and Bioconductor, allowing researchers to carry out gene expression analysis, find differentially expressed genes, and gain understanding of biological pathways and processes.

2.3.5 Org.Hs.eg.db

Bioconductor, a popular software platform for the analysis of genomic data, has a database package called org.Hs.eg.db [66] [12]. The annotation of the human genome is especially mentioned. An extensive database of data on human genes and the characteristics linked to them is available as part of the "org.Hs.eg.db" package. Data on gene names, genomic locations, functional annotations, gene identifiers, gene symbols, and other pertinent information for human genes are included in this database package. Researchers in the fields of genomics and bioinformatics can use it to integrate experimental data with gene annotations and carry out a variety of downstream analysis, making it a valuable resource. For gene-centric analyses including pathway analysis, gene ontology analysis, and gene set enrichment analysis, the org.Hs.eg.db package is frequently used in conjunction with other Bioconductor packages and the R programming language. It helps in the analysis of genomic data and the elucidation of interesting genes by providing researchers with insights into the biological processes, pathways, and regulatory mechanisms connected to human genes.

2.4 Biological Analysis

Genes identified by each of our four proposed frameworks as candidates for potential biomarkers for a disease of interest are validated biologically. We achieve this through functional enrichment analysis and regulatory behavior analysis. These methods are disscused in brief.

[12] https://bioconductor.org/packages/org.Hs.eg.db/

2.4.1 Functional Enrichment analysis

A bioinformatics method known as functional enrichment analysis is employed to identify functional categories, such as pathways or Gene Ontology (GO) terms, that are over-represented in a set of genes or proteins relative to what would be predicted by chance. A popular technique to learn more about the biological processes and pathways that a group of genes or proteins affect is functional enrichment analysis. Functional enrichment analysis typically involves the following steps.

1. The background set refers to the group of genes or proteins that are being studied; typically, they are the ones that exhibit differential expression or are otherwise relevant to a given experimental setting.
2. To establish which functional categories are deemed highly enriched, a statistical significance criterion is chosen. The desired level of significance, which is commonly expressed as a *p-value* or FDR, can be taken into account while adjusting this threshold.
3. By comparing the frequency of functional category members in the background set to the frequency anticipated by chance, enriched functional categories are identified.
4. Multiple testing correction techniques can be used to alter the significance level in order to take into account the numerous functional categories being tested concurrently. The Bonferroni adjustment [49], FDR correction, and Benjamini-Hochberg [43, 764] correction are most widely used techniques.

Various functional databases, including Gene Ontology (GO) [30] , KEGG pathways [297], Reactome [110], and WikiPathways [551], can be used to do functional enrichment analysis. Functional enrichment analysis can be carried out using a variety of software packages and tools, including R and Python libraries as well as online platforms like DAVID [628, 253], and Enrichr [74].

2.4.1.1 GO Enrichment

GO enrichment analysis [30, 17] is a computational method that is used to examine whether a group of genes or proteins is noticeably over-represented for a given collection of GO terms in comparison to a chance expression. A certain group of genes or proteins may be enriched in certain biological processes (BP), molecular functions (MF), or cellular components (CC). By identifying these processes, molecular functions, and cellular components, this type of analysis can shed light on their functional roles and relationships. The functional analysis of genes and proteins can be aided by the strong tool of GO enrichment analysis, which also aids in the identification of important biological pathways and processes linked to specific phenotypes, illnesses, or experimental situations.

2.4.1.2 Pathway Enrichment

Biological pathways that are considerably enriched in a group of genes or proteins when compared to what would be predicted by chance are found using the bioinformatics technique known as pathway enrichment analysis. Finding out which biological processes are influenced by the target genes or proteins can be done with the use of a pathway enrichment study. Pathway enrichment analysis can be performed using various pathway databases, such as KEGG [297], Reactome [110], and WikiPathways [551]. Pathway enrichment analysis is an effective method for determining functional relationships between genes and proteins, and it can aid researchers in developing ideas about the underlying biological processes that underlie phenotypic variations or disease states.

2.4.2 Regulatory Network Behavior Analysis

Proteins called transcription factors (TFs) [278] attach to the DNA [758] with the aim to control the transcription of genes into mRNA as described in the central dogma of molecular biology (Section 1.1. TFs play a significant role in the process of transcription by controlling when and how much gene is translated into RNA thus having an impact on the quantity and activity of the protein that is generated. RNA polymerase is activated or repressed by the action of TFs. TFs play a key role in a variety of crucial biological processes, including differentiation, development, and reaction to environmental cues. Cancer and developmental abnormalities are just two examples of the diseases that might occur as a result of TF mutations or dysregulation. The human genome has hundreds of different TFs, each with a distinct DNA binding specificity and regulatory function. Changes in the activity of one TF can have a cascade impact on the expression of numerous downstream genes. Many TFs collaborate in intricate networks to govern the coordinated expression of genes. The understanding of the regulatory networks that govern gene expression is crucial for creating novel treatments for diseases and for expanding our knowledge of fundamental biological processes, making research into TFs an important topic of study in both basic and practical biomedical research.

A gene regulatory network (GRN) [278, 122] is an intricate network of genes, transcription factors, and other molecular components that collaborate to regulate cellular functions and gene expression. Gene regulatory interactions, which can either be activating or inhibitory, connect genes in a GRN and work together to control the coordinated expression of genes. GRNs are crucial for cells to operate normally and are important for growth, differentiation, and disease. They can be applied to simulate the intricate regulatory mechanisms that underlie biological systems and comprehend the processes that give rise to various cellular phenotypes. A GRN can be developed and then examined to locate important regulatory hubs and subnetworks involved in particular cellular processes.

2.5 Biomarker Criterion

For each proposed framework, we identify a set of genes that are candidates for potential biomarkers . We term these genes as Biomarker Candidate Genes (BCGs). Through biological analysis of the BCGs identified by each framework we establish their relevance to their respective datasets. We achieve validation through a) GO enrichment analysis, b) KEGG pathway enrichment analysis, c) gene regulatory network (GRN) analysis , and d) tracing literature evidence that associate the BCG with corresponding disease and other diseases associated with it. The biological relevance of a BCG to its respective dataset is considered based on the following three criteria.

(a) Annotated to at least one GO term in two out of three GO databases (BP: Biological Process, CC: Cellular Process, and MF: Molecular Function) with *p-value* ≤ 0.05, i.e. significant with 5%,

(b) Annotated to at least one KEGG pathway with *p-value* ≤ 0.05, i.e. significant with 5%, and

(c) It's a TF and thus exhibits regulatory behavior towards other genes in the network.

For a BCG to be considered a potential biomarker for the corresponding disease (Here we consider, ESCC), we consider following four cases.

Case 1: Strong literature evidence of association of the BCG with the disease (here, ESCC) and biologically relevant to the corresponding dataset where the BCG was detected based on all three criteria a,b and c (i.e., $a \cap b \cap c$),

Case 2: Strong literature evidence of association of the BCG with the disease (here, ESCC) and biologically relevant to the corresponding dataset where the BCG was detected based on criteria a and b but not c (i.e, $a \cap b \not{c}$),

Case 3: Strong literature evidence of association of the BCG with the disease (here, ESCC) and biologically relevant to the corresponding dataset where the BCG was detected based on criteria a or b (and/or c) (i.e, $a \cup b(\cap/\cup)c$),

Case 4: Biologically relevant to the corresponding dataset where the BCG was detected based on all three criteria a,b and c, (i.e., $a \cap b \cap c$) and has literature evidence of association of the BCG with a few prominent diseases strongly related to the disease but has no literature evidence of association of the BCG with the disease (here, ESCC).

We employ this biomarker criteria across all four of our proposed frameworks to identify BCGs and potential biomarkers for a critical disease of interest.

2.6 Datasets Used

Esophageal Squamous Cell Carcinoma (ESCC) is known to originate in the cell lining of the esophagus. The esophagus is a muscular tube that connects the throat to the stomach. ESCC is the most prevalent type of esophagus cancer throughout the world, particularly in the developing nations, an notably in the North-East of India. Development of ESCC is frequently attributed to risk factors such as the use of tobacco or betel nut chewing, specific dietary factors, such as consuming hot beverages and foods with high nitrosamine levels., and chronic inflammation brought on by some disease.

2.6.1 Microarray Data

We have chosen two ESCC microarray datasets , GSE20347 and GSE23400 to analyze three proposed frameworks. The datasets are discussed in brief below.

Tab. 2.1: Summary of the microarray datasets, GSE20347 and GSE23400, and the bulk RNA-Seq dataset, GSE130078 for ESCC

Dataset	No. Of Genes	Normal Samples	Tumor Samples	Data Type	Public On
GSE20347	22,277	17	17	Expression Data	Mar 15, 2011
GSE23400	22,283	53	53	Expression Data	Sep 1, 2010
GSE130078	57,783	23	23	Count Data	Oct 28, 2019

Dataset GSE20347 [13] titled "Analysis of gene expression in esophageal squamous cell carcinoma (ESCC)" proposed by Hu et al. [249] characterize gene expression in ESCC.

[13] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20347

They achieved this through the examination of gene expression in tumor and matched normal adjacent tissue from 17 ESCC patients from a high-risk region of China. The rows of the dataset represent the genes. For each gene, the first 17 columns represent the gene expression for that gene in samples for normal while the last 17 columns represent the expression for that gene in samples for disease. Probe IDs are used as unique gene (row) identifiers.

Dataset GSE23400 [14] titled "Global gene expression profiling and validation in esophageal squamous cell carcinoma (ESCC)" was proposed by Su et al. [648] with the aim towards understanding molecular changes in ESCC. For each gene, each pair of columns represent the normal and disease sample pairs for each of the 53 patients which results in 106 columns. Probe IDs are used as unique gene (row) identifiers.

2.6.2 BulkRNAseq data

Alongside the two microarray datasets previously discussed, we have chosen a ESCC bulk RNA-Seq dataset, GSE130078 to analyze three of the proposed frameworks. Dataset GSE130078 [15] titled "A Novel LincRNA HERES Epigenetically Regulates Wnt Signaling Pathway via Interaction with EZH2 in Esophageal Squamous Cell Carcinoma" was proposed by You et al. [832] with the aim to identify ESCC-driving lncRNAs in the transcriptome level and exploit a therapeutic target with understanding their modes of action. For each gene, each pair of columns represent the normal and disease sample pairs for each of the 23 patients which results in 46 columns. Ensembl IDs are used as unique gene (row) identifiers.

2.6.3 ScRNAseq data

To analyze our single cell RNA-Seq (scRNA-Seq) analysis framework, we have chosen an ESCC scRNA-Seq dataset, GSE160269. Dataset GSE160269 [16] titled "Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis" was proposed by Zhang et al. [877]. Zhang et al.[877] performed scRNA-seq on 60 ESCC tumors and 4 adjacent normal tissue samples obtained from 60 individuals using the 10X Genomics platform. Single-cell suspension was stained with CD45-FITC and sorted into immune (CD45+) or non-immune (CD45-) cells.

[14] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23400

[15] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130078

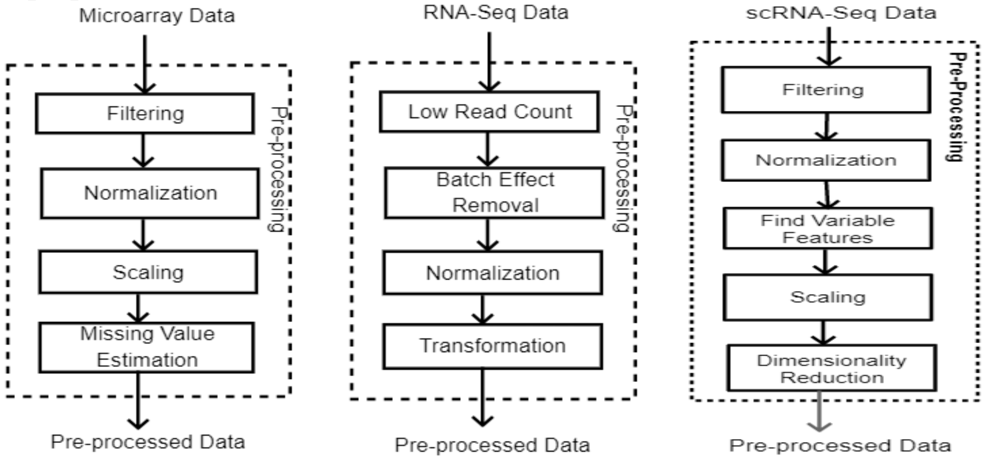
[16] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160269

Tab. 2.2: Summary of the scRNA-Seq dataset, GSE160269[877] for ESCC

Subset dataset	Cell Types	No. of Cells (Samples)
CD45+ (No. of Genes: 15,175)	Tcells	69,278
	Bcell	22,477
	Myeloid	19, 273
CD45+ (No. of Genes: 17,012)	Epithelial	44,730
	Endothelial	11,267
	Fibroblasts	37,213
	Pericytes	3,102
	Fibroblastic Reticular Cells	1,319

2.7 Pre-processing Of Gene Expression Data

In this section we discuss in brief the pipelines we have employed to pre-process all three types of datasets, namely microarray, bulk RNA-Seq , and scRNA-Seq across all the proposed frameworks.



(a) Microarray *(b)* Bulk RNA-Seq *(c)* ScRNA-Seq

Fig. 2.1: Pre-processing pipeline employed by our proposed frameworks for the three types of gene expression data, microarray, RNA-Seq, and scRNA-Seq data.

2.7.1 Pre-processing of Microarray Data

The general pipeline involved in pre-processing of microarray data is illustrated in Fig 2.1a and discussed in brief below.

1. Filtering helps in removing irrelevant or noisy data from the dataset, reducing noise and increasing the signal-to-noise ratio.
2. Normalization aims to remove technical variations that may arise during the experimental process, such as differences in labeling efficiency, hybridization efficiency, and scanning intensity, so that the biological variations can be accurately detected.

Rescaling the data so that it falls within a specific range, usually between 0 and 1.

3. Scaling is used to adjust the range of expression values so that they are comparable across different genes and samples. This step is necessary because the expression levels of different genes may vary over several orders of magnitude, and the scale of measurement may also differ between samples.Transforming the data so that it has a mean of 0 and a standard deviation of 1.
4. Microarray data may contain missing values due to various reasons, such as low signal intensity, poor probe quality, or experimental failure. We can estimate missing values before performing some of the common methods such as K Nearest Neighbor (KNN) [169], PCA [546, 292], etc.

2.7.2 Pre-processing of bulk RNA-Sequencing (Bulk RNA-Seq) Data

The general pipeline involved in pre-processing of bulk RNA-Seq data is illustrated in Fig 2.1b and discussed in brief below.

1. Low read count RNA sequencing data refers to RNA sequencing data that has a low number of reads per sample. Low read count can arise due to various reasons, such as low RNA input, poor sequencing quality, or technical variation during library preparation and sequencing. DESeq normalization and edgeR can inherently handle low read counts.
2. Batch effects refer to systematic variations in gene expression data that arise due to technical factors, such as differences in sample processing, sequencing, or labeling. PCA [546, 292], Limma [637, 638] empirical Bayes framework are effective in batch effect removal.

2.7.3 Pre-processing of Single Cell RNA-Sequencing (scRNA-Seq) Data

The general pipeline involved in pre-processing of scRNA-Seq data is illustrated in Fig 2.1c and discussed in brief below.

1. Quality control identifies and removes low-quality cells and genes. Quality control metrics can include gene count, mitochondrial gene content, and ribosomal gene content.
2. Genes that exhibit high levels of expression variation across cells. Variable features characterize cell types and can identify genes that drive cell differentiation.
3. Reducing the high-dimensional gene expression data into a lower-dimensional space

while preserving the major sources of variation in the data. It can help to identify cell types and states, visualize the data, and reduce the noise in downstream analyses.

2.8 Discussion

In section 2.1 of this chapter, we discuss all statistical methods and measures employed across various stages of all our proposed frameworks. Section 2.2 gives a summary of the R programming platform where all our experiments were conducted as well as various other tools used for implementation of our proposed frameworks and validation of the results of our analysis. In section 2.3 we give a brief summary of all biological knowledge repositories utilized through various stages of all our proposed frameworks. In section 2.4 we give a summary of the methods employed for biological analysis of the results of each framework. We have proposed a criteria employed by all our frameworks that identifies a BCG as potential biomarker for the corresponding critical disease which is discussed in detail in section 2.5. Section 2.6 gives a summary of the two microarray datasets and one bulk RNA-Seq datset for ESCC used by the first three contributions as well the scRNA-Seq dataset for ESCC used by the final contribution. Finally, we discuss the pre-processing employed by our frameworks for all three types of gene expression datasets, namely, microarray, bulk RNA-Seq and scRNA-Seq.

Next chapter explores the biclustering approach for gene expression data analysis. We review the pros and cons of various biclustering methods presented throughout the years. We have chosen eight biclustering methods that span across multiple approaches and incorporate them into our proposed bicluster analysis framework. We employ the chosen eight biclustering algorithms to generate biclusters and using a selection criteria filter relevant biclusters for subsequent analysis and validation.