Study of Cancer Drug Resistance: Insights from Differentially Expressed Genes and Computational Approaches

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Shwetang Mupade

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

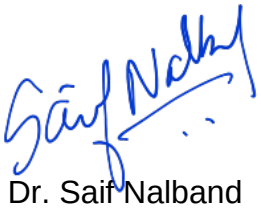
Dec 2023

Supervisor: Dr. Saif Nalband

Shwetang Mupade

Certificate

This is to certify that this dissertation entitled study of cancer drug resistance: insights from differentially expressed genes and computational approaches towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study carried out by Shwetang Mupade at Thapar Institute of Engineering & Technology, Patiala Punjab. Under the supervision of Dr. Saif Nalband, Department of Computer Science, during the academic year 2022-2023.



Dr. Saif Nalband

Committee:

Dr. Saif Nalband

Dr. Krishnapal Karmodiya

This thesis is dedicated to my Parents

Declaration

I hereby declare that the matter embodied in the report entitled study of cancer drug resistance: insights from differentially expressed genes and computational approaches are the results of the work carried out by me at the Department of Computer Science, Thapar Institute of Engineering and Technology, Patiala Punjab, under the supervision of Dr. Saif Nalband and the same has not been submitted elsewhere for any other degree

Shwetang Mupade

Date: 20 Oct 2023

Contents

List of Tables

List of Figures

Abstract

Breast cancer remains a global health challenge, with over 1.5 million cancer patient and half a million deaths reported annually. The Indian subcontinent has witnessed a high prevalence rate, emphasizing the urgent need for scientific research. Drug resistance poses a significant hurdle in chemotherapy, with over 90% of cancer-related deaths attributed to this phenomenon. Understanding the genetic underpinnings of drug resistance is crucial for improving treatment efficacy and reducing tumor relapse. This study leveraged RNA-Seq data to identify differentially expressed genes of drug-resistant versus sensitive cancer cell lines. Utilizing the Galaxy platform, a comprehensive analysis was conducted, encompassing data collection, alignment, quantification, and differential expression analysis. Furthermore, various unsupervised machine learning techniques, including K-means clustering, Gaussian Mixture Model, and Hierarchical Clustering, were applied to discern underlying structures and patterns within the data. The evaluation of clustering methods was performed using the Davies-Bouldin Index (DBI), a metric assessing the separation and compactness of clusters. The results were employed to benchmark the performance of different clustering techniques across various cancer types and drug treatments. The study focuses on potential of computational models and omics data in predicting drug responses, thereby advancing personalized cancer therapy. This research contributes to the burgeoning field of personalized medicine, providing valuable insights into the genomic basis of drug resistance. By integrating bioinformatics tools, machine learning, and comprehensive data analysis, this study paves the way for more effective cancer treatment strategies, ultimately improving patient outcomes.

Acknowledgments

I wish to convey my profound appreciation to Dr. Saif Nalband, my supervisor, for his priceless guidance, steadfast backing, and enlightening mentorship during the progress of this research. His knowledge, commitment, and motivation played a pivotal role in moulding this thesis. I also owe a debt of gratitude to Dr. Krishanpal Karmodiya for his expertise and invaluable input into this project. His profound knowledge and critical insights significantly enriched the depth and quality of the research. I extend my appreciation to the faculty and staff of IISER Pune, whose continuous support created an enriching academic environment. Their collective efforts played a crucial role in the successful completion of this thesis. I am grateful to my colleagues and fellow researchers for their camaraderie, stimulating discussions, and shared passion for scientific inquiry. Their contributions were instrumental in refining the ideas and methodologies presented in this work. I want to thank my dear friend Pallavi Kiratkar stimulating discussions, and shared passion for scientific inquiry. Their contributions were instrumental in refining the ideas and methodologies presented in this work. In conclusion, I'd like to express my gratitude to my family for their unwavering support, motivation, and comprehension during my academic pursuit.

Contributions

Contributor name	Contributor role
Dr. Saif Nalband and Shwetang Mupade	Conceptualization Ideas
Shwetang Mupade and Dr.Saif Nalband	Methodology
Shwetang Mupade Dr. Saif Nalband	Software
Dr. Saif Nalband	Validation
Shwetang Mupade and Dr.Saif Nalband	Formal analysis
Shwetang Mupade	Investigation
---	Resources
Shwetang Mupade and Dr.Saif Nalband	Data Curation
Shwetang Mupade	Writing - original draft preparation
Dr.Saif Nalband	Writing - review and editing
Shwetang Mupade	Visualization
Dr. Saif Nalband	Supervision
Dr. Saif Nalband and Dr. Krishanpal Karmodiya	Project administration
----	Funding acquisition

This contributor syntax is based on the Journal of Cell Science CRediT Taxonomy[1].

[1] https://journals.biologists.com/jcs/pages/author-contributions

Chapter 1 Introduction

New Breast cancer cases found globally near one million yearly and half million deaths, therefore it is focused for scientific research(Siegel *et al.*, 2018; DeSantis *et al.*, 2019). In Indian subcontinent it has found that in one lakh population around 25 are breast cancer patient (Ferlay *et al.*, 2015). There is surge in cancer related deaths in Indian subcontinent this is studied by global health organization and also locally (Porter, 2009; Babu *et al.*, 2013). It is found that cancer ranks second in death, around 1 in 6 deaths occur due to some form of cancer

Data indicate that there is need for the research on this field. There is personalised medicine which is widely used for the treatment of cancer according to patient history, how they response to therapy, what are genomic changes(Chen *et al.*, 2017). In last decade researchers are inclined towards genomic changed in patient which give information about related molecular changes. Now a day's chemotherapy is mainstream treatment for the treatment for breast cancer. There are many drugs which are used for the chemotherapy for example Cytoxan , Adriamycin, Fluorouracil, Tamoxifen, Endoxifen these are the medications used for treatment of cancer. But there is rising drug resistance in this chemotherapy drugs which is causing problem for cure in cancer(Nounou *et al.*, 2015; Jayaraj *et al.*, 2019). According to data it is found that more than 90% deaths in cancer patient are related to drug resistance. Drug resistance changes to person to person(Vulsteke *et al.*, 2013; Liu *et al.*, 2020). There is need to know the mechanism behind the chemo resistance which will help to increase the effectiveness chemotherapy that will result in less relapse of tumor in patient. There may be case in which there are multiple drug resistance of cancer cells while getting chemotherapy. There are many mechanisms of drug resistance which are increase in outflow of drugs, genetic alteration, gene mutation, increase in DNA repair capacity, increase metabolic activity of xenobiotics(Wu *et al.*, 2014; Wang *et al.*, 2017, 2019). because of this mechanism there is failure of treatment with chemotherapy drugs. Therefore, drug resistance is the main reason for failure in treatment of patient. The difficulty in drug resistance

and anti-microbial resistance are identical to each other. Both of them are having very high amount of aggressors which may be intrinsic or extrinsic. In early stage of anti-microbial treatment and chemotherapy treatment both were successful but at later stages there is rise in antimicrobial resistance and tumor relapse.

One of the reason behind cancer is genetic mutation, changes in genes. There is complication in tumor cell level which result in making cancer more complicated disease for treatment. In process of treatment of cancer those patient having same cancer type shows different response to the therapy they receive, therefore cancer is heterogeneous(Dagogo-Jack and Shaw, 2018). This kind of response of patients to the therapy is due to genetic changes in patients. While treating patients with individual specific treatment is very daunting task, therefore researchers are working on this. Researchers have performed drug screening in very high number for finding if there is any co-relation in genomic changes and response of drug in patient (Costello *et al.*, 2014; Azuaje, 2017). This will produce huge amount of data to study. Therefore, there is need work on computational model which can predict correct therapy for patient using genomic data of patient. There are a lot of cancer therapy drug available in market which can be used for treatment of cancer patient, but individual having identical cancer shows distinct response towards drugs this is due to heterogeneous characteristics of tumor. Researchers found that patient response to the drug is dependent on genetic changes in cancer patient. Therefore, Nowadays a lot of cancer drug response data to a different types of cancer is available. There were many in vitro research happened on drug response on different types of cancer(Basu *et al.*, 2013). As cancer is heterogeneities there is need of proper predication of drug response towards tumor for the development of new therapy. As everyone is trying to develop new cancer therapy this will only work in some patient for other it will not work. To address this issue there is need to know about biomarkers which will help to predict response of drug in person to person basis(Menden *et al.*, 2018). A there is rise in development of hightroughput technology and reduction in cost many researchers are working for developing molecular feature of multiple gene on response towards drugs from this they will get to know accurate biomarker eventually it will help in building prediction o drug response(Ding *et al.*, 2016). This response of drug prediction can be used for developing new drug, drug repurposing etc... Drug repurposing is used to get to know if drugs which are available can be used for other cancer. In this scenario drug is not discovered for specific cancer just it is not tested for that drug. For the

development of new drug there is need to predict how existing cancer patient response to the newly designed drug.

There are many genomic data projects which give response of patient towards drug for e.g TCGA (The Cancer Genome Atlas), and GDSC(Genomics of Drug Sensitivity in Cancer) projects(Barretina *et al.*, 2012; Yang *et al.*, 2013). Database of this project help in building new computational model which help in predicting response of drug using genomic data of cell line also this data help to find compound which will work in tumor which is resistant to drug(Gönen and Margolin, 2014; Choi *et al.*, 2017; Chaudhary *et al.*, 2018). As this data is huge and useful for making computational model which can use this data for building model of interest. This dataset contain Gene expression data, protein expression, copy number variation. Use of mutliomics data for the prediction of biomarkers in many cancers is evolved as a main approach. Building models for patient survival and how patient response to the drug is new way for treatment of individual's therapy. In computational model use of omics data helps in analysing genome of patient at various level and get some important insight on it(Zhang *et al.*, 2018a). As omics data having heterogeneity also very steep dimensionality because of this linear prediction model will not work. Therefore, there is need change the approach for handling this kind of data.

Recently for prediction of drug sensitivity there is increase in use of computational model in researchers because there is increase in individual specific therapy for the treatment of cancer patient. Also there is huge amount of pharmacogenomics data which is available for prediction. In treatment of individual specific drug therapy prediction of drug sensitivity is very daunting task. As individual specific treatment will not follow one size fits for all more readily it is very narrowed for individual specific drug. The main goal in individual specific drug treatment is to recommend drugs to patient depending on their genetic data (Costello *et al.*, 2014). Before the individual specific drug treatment many drug treatment are given on the basis anatomical rise of disease.

In last few years' researchers focused on understanding response of drug prediction while using genomic specification. Use of copy number variation also methylation was the main strategy for identifying biomarkers to which response of drug is found. In (Zhang *et al.*, 2018b)If we look into their study they showed a process to know about important biomarker and building classifier using genomic data for the prediction of response of drug. In (He *et al.*, 2011). review they showed how copy

number variation play major role drug effectiveness. Many studies conducted, using Bayesian factorization for response of drug prediction while taking consideration of pathway of drug. In Wang et al. study they used expression of gene data for response of drug prediction towards cancer. They considered pathway of gene expression based model for this. But all this method are focused on using single omics data only. For identification of biomarker and their detailed information use of multiomics data of genome will help to improve prediction of response of drug and also get to know about mechanism of it.

There are lot of methods which can be used for response of drug prediction and constructing model which can be classified in different ways. While predicting sensitivity of drug and combination of drugs for the cure of cancer many computational models available such as regression, kernel based, ensemble learning(Sharma and Rani, 2020). This models are dependent on genetic information for the sensitivity drugs prediction. This models are built in such a way that drugs having same target with same drugs. In literature it is found that this model uses same structure of drugs and their resemblance towards cell lines. While drug response prediction in cancer cell researchers used more often Random forest, Elastic net machine learning model. Many researchers also used support vector machine, Random forest and Neural network.In last few yearas ther is rise in use of deep learning in prediction of drug reponse towards cancer therapy

In many studies researchers did prediction on the basis of dosage of drug to the cell growth. Some researchers also build prediction model without depending on dosage of drug to the cell(Xia *et al.*, 2018). Others also looked in half dosage required for inhibition of cell growth. Many researchers build prediction model on single cancer while others used multiple cancers for response of drug prediction. In many studies transcriptomic data, proteomic data, genomic data and omics data is used for the prediction of response of drug(Rampášek *et al.*, 2019). In all this data transcriptomic data shows more accurate prediction towards response of drugs. In many prediction model researchers focused on using one drug, some of them also used multiple drugs for prediction of response of drug(Zhu *et al.*, 2020). While building prediction model many researchers focused on similarity between drug and tumour they used molecular specification of drugs and tumor , they used many machine learning methods such as Bayesian kernel learning(Sharma and Rani, 2018), weighted graph regularized collaborative matrix factorization, Collaborative Filtering this are used for predicting

response of drugs.

We used RNA-Seq data and did Deseq2 on it to know about differentially expressed gene in chemo resistance cancer drugs. Then we used unsupervised machine learning to study data and get some insights on it. We used different types of clustering model such as K-nearest neighbour, Gaussian mixture model, Mean shift clustering.

Chapter 2 Materials and Methods

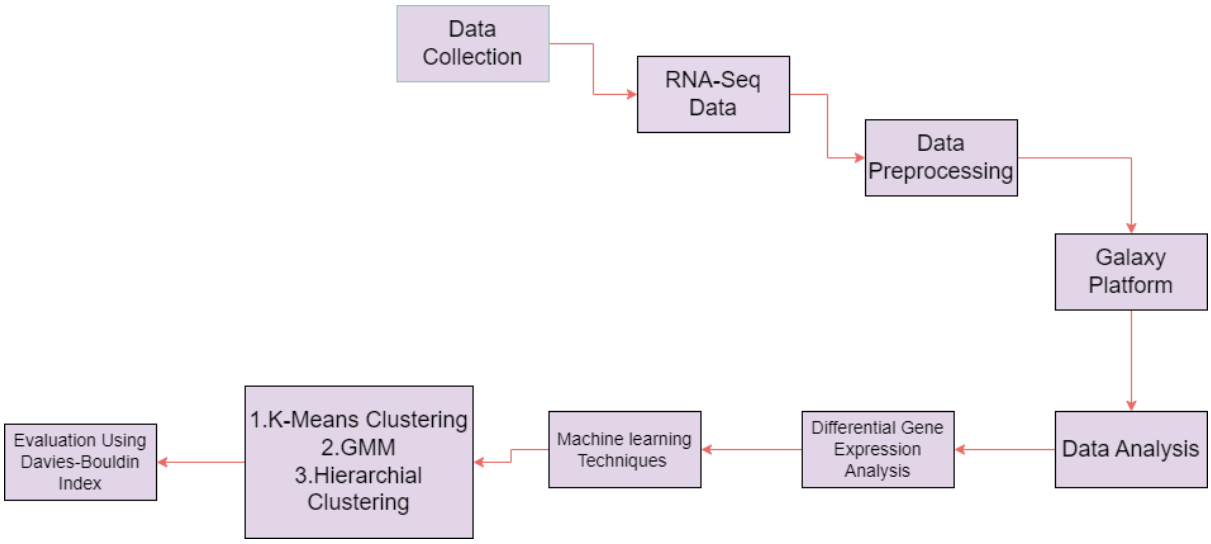


Figure 1 Overall workflow of thesis

Bioinformatics

1. Data Collection and Preparation

The RAN-Seq data has been downloaded from https://www.ncbi.nlm.nih.gov/sra. It contains transcriptomic data of drug resistance and sensitive cancer cell line (Table 1).

Table1 Dataset

Cancer Type	NCBI Database	Platoform	Samples (Drug resistance versus sensitive)	Drug Name
Breast cancer	SRR11123344	Illumina HiSeq 2000	3 versus 3	Doxorubicin
Breast cancer	SRR13398517	Illumina HiSeq 2500	3 versus 3	Endoxifen
Breast cancer	SRR7064965	Illumina HiSeq 2000	3 versus 3	Paclitaxel
Ovarian Cancer	SRR14785840	Illumina HiSeq 2500	3 versus 3	Doxorubicin
Prostate Cancer	SRR17196727	Illumina HiSeq 2000	3 versus 3	Docetaxel
Pancreatic cancer	SRR14271601	Illumina NovaSeq 6000	3 versus 3	Gemicitadine
Leukaemia	SRR13239311	Illumina HiSeq 4000	3 versus 3	Etoposide

In last few years, Constant change in transcriptome of cell i.e. group of RNA molecule in a cell. therefore, researchers are using RNA sequencing technology for the analysis. Many researchers use RNA-Seq data for differentially expressed gene in different condition such as sensitive vs resistance. There is need of computational instruments for finding differentially expressed gene while using RNA-sequencing data because of this many users it will not easy. Therefore, Galaxy is a software which is based on web for the biological data analysis. It has very user-friendly web interface. In Galaxy for finding differentially expressed gene from RNA sequencing data, there is easy steps and tools. Galaxy software has great tools and developers has made this very interactive because of this it becomes very easy and fast for analysis. In background Galaxy uses lot of computational environment and cloud technology for data analysis which user can do by using only web browser to do any task on data. More than 100

open servers can be used on Galaxy for free and it can be installed in system. We used galaxy software for finding differentially expressed gene in RNA seq data of drug resistance vs sensitive in different types of cancer cell lines.

2.Step for finding differentially expressed gene from RNA-Seq Data while using Galaxy

1. Searching required data in NCBI in our case chemo resistance in breast cancer and searched for suitable dataset
2. RNA Seq data which has FASTQ files which is unprocessed data obtained fom experimental processes
3. Uploading FASTQ files in galaxy through NCBI database
4. Then In galaxy need to upload reference genome which should be compatible for Galaxy software is FASTA for the genomic sequence and also gene annotations files which is in GTF format. This reference genome of an organism should be related to the study you are going to conduct which will give accurate alignment and also annotation.
5. Bowtie(Read alignment) tool which is present in Galaxy software can be used for mapping uploaded RNA –seq file with reference genome corresponding of it. We need to be very careful while choosing correct reference genome file.
6. Next step is to gene expression quantification and feature counting. For feature count there is tool in Galaxy software known as Feature counts in which it count the number of reads assigned to each gene this results in a count table
7. After getting count table from feature counting which contains information about read count associated with each gene. We gave input as resistance count table and sensitive count table which has three replicates each in Deseq2 tool which is present in Galaxy platform for differentially gene expression analysis. Data provided for this analysis included:
8. Append annotation from GTF to differentially expressed tool format
9. Deseq2 result file contain advanced statistical analysis and gene expression changes during drug resistance and sensitive.
10. For visualization of Deseq2 results there are heatmap and volcano plots generated.

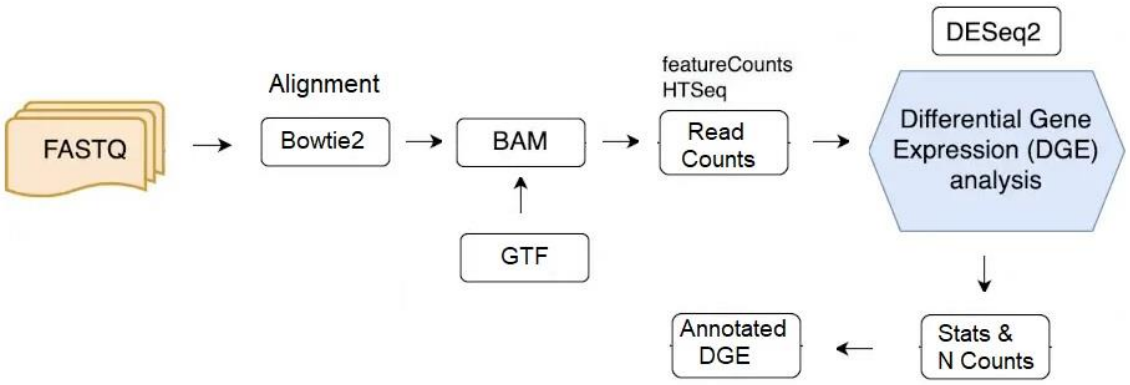


Figure 2 Workflow of Deseq2 analysis

Machine Learning

In machine learning there is subcategory of it known as unsupervised machine leaning. It is based on training algorithm without target variable which help to understand structure of data, pattern of it and relationship within data without any previous knowledge. There are many methods such as dimensionality reduction, density reduction generative model and density reduction in unsupervised machine learning but most commonly used is clustering.

Clustering in Unsupervised Learning

In clustering there is grouping of data points into the cluster according to similarities or dissimilarities in data. Main aim of clustering is to create dataset within the data in such a way that data points in sample cluster show similarity to other cluster. Cluster is collection of data point which has similarities in them. There are many clustering model which is found in unsupervised machine learning. Every clustering model has their own way of froming data points. Majorly clustering models used are K-means clustering, Hierarchical clustering, Mean shift clustering, DBSCAN(Density based spatial clustering of application with noise) and Gausisan mixture model. In our study we used K-means clustering, Gaussian mixture model and Hierarachial Clustering.

1. K-means Clustering

K-means, which is also known as Lloyd's algorithm, can be broken down into three main steps. In the first step, you start by picking initial centroids, usually by randomly selecting some data points from your dataset. Then, the algorithm goes through a repetitive process that involves two key actions: assigning each data point to the nearest centroid and updating the centroids by computing the average of the data points assigned to each previous centroid.This repetitive process continues until there's only a minimal difference between the old centroids and the new ones. This small difference indicates that the algorithm has converged to a stable solution. You can think of K-means as having similarities to the expectation-maximization algorithm, with the added constraint of using a small, equal, diagonal covariance matrix for each cluster. Another way to grasp K-means is by considering Voronoi diagrams. Initially, you compute the Voronoi diagram using the current centroids, where each segment in the diagram corresponds to a distinct cluster. After that, you update the centroids to be the mean of the data points in each segment, and this whole process repeats until a certain stopping condition is met.Typically, the algorithm stops when the change in the objective function between iterations becomes very small, less than a specified tolerance value. Alternatively, it can also stop if the centroids move by a very small amount, less than the specified tolerance. This signifies that K-means has successfully partitioned your data into clusters that are well-defined and stable.

It's important to note that K-means may converge to a local minimum, which is heavily influenced by the initial centroid selection. To mitigate this, multiple runs with different centroid initializations are often performed. The k-means++ initialization method is a technique for enhancing the initial setup of centroids. It ensures that the centroids are placed at relatively distant positions from each other, which frequently results in improved outcomes compared to random initialization. Additionally, K-means enables the incorporation of sample weights, which allows you to give greater significance to specific data points when determining cluster centers and inertia values. For instance, if you assign a weight of 2 to a data point, it's akin to having that data point appear twice in the dataset. Lastly, K-means can also be utilized for vector quantization by utilizing the 'transform' function of a trained KMeans model.

2. **Gaussian Mixture model**

This model is also known as probablilistic model. Which is mixture of unknown parameters with finitite number of gausian mixure which results in generation of data points. In simple way Mixture models are like an upgraded version of k-means clustering. They not only find the central points of groups in your data but also consider how spread out or squished together the points in each group are . The Gaussian Mixture object utilizes the expectation-maximization (EM) algorithm to build models that involve mixtures of Gaussian distributions. Moreover, it has the ability to generate confidence ellipsoids for models that deal with multiple variables, providing a way to visually represent the uncertainty in the data. A Gaussian mixture model can be understood as a combination of M Gaussian densities, each weighted differently, as described by the following equation

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \, g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

Here, we're dealing with a D-dimensional data vector, where D represents continuous values, such as measurements or features. The mixture includes various components, each with its own weight represented as wi for i ranging from 1 to M. These components are characterized by Gaussian densities, denoted as g(x| μi,∑i), with i also ranging from 1 to M.Furthermore, it utilizes the Bayesian Information Criterion to assess and decide on the best number of clusters within your dataset. You can employ the Gaussian Mixture fit method to train a Gaussian Mixture Model using your training data. Once this model is trained, you can use the Gaussian Mixture predict method to assign each sample in your test data to the Gaussian group it is most likely a part of based on the model's learning. The Gaussian Mixture offers various choices to control how it handles the covariance of the different classes it figures out. Furthermore, it utilizes the Bayesian Information Criterion (BIC) to assess and decide on the best number of clusters within your dataset. You can employ the Gaussian Mixture fit method to train a Gaussian Mixture Model using your training data. However, it's important to note that BIC accurately identifies the true number of components only under certain conditions, primarily when ample data is available and the data conforms to the assumptions of being independently and identically distributed from a mixture of Gaussian distributions. The Expectation-Maximization algorithm provides a robust statistical approach to address this challenge. The procedure begins by making an

initial assumption of random components and associating probabilities with each data point, indicating the likelihood that they belong to each component in the model. Afterward, the algorithm fine-tunes the model's parameters to optimize the likelihood of the data, based on these assignments. Iterating through this process ensures convergence to a local optimum, making it a reliable way to uncover latent structures in your data.

3. Hierarchical clustering

Hierarchical clustering is a versatile family of clustering methods that create a hierarchy of clusters by gradually combining or dividing them. This hierarchy is depicted as a tree structure, often called a dendrogram. At the top of the tree is a single cluster encompassing all samples, while the branches, or leaves, represent clusters containing only one sample. The representation of the hierarchical merging of clusters can be visualized as a dendrogram . This graphical depiction provides a visual means to comprehend the structural relationships within the data. While it can be particularly informative for small sample sizes, visual inspection remains a valuable tool for gaining insights into the data's underlying hierarchy. The Agglomerative Clustering object employs a bottom-up approach for hierarchical clustering. Initially, each data point starts as its own cluster, and these clusters are gradually merged. The specific method used for merging, referred to as the linkage criteria, determines the metric used for this merging strategy:

1. Ward's linkage aims to minimize the sum of squared differences within clusters, similar to reducing variance, and it resembles the objective function seen in k-means clustering but applied hierarchically.
2. Maximum (or complete) linkage works to minimize the maximum distance between any pair of data points from different clusters.
3. Average linkage seeks to minimize the average distances between all pairs of data points from different clusters.
4. Single linkage aims to minimize the distance between the closest data points from different clusters.

When combined with a connectivity matrix, Agglomerative Clustering can effectively manage extensive datasets. However, without such constraints, it becomes computationally demanding as it considers all possible merging combinations at each step.Additionally, there's a tool called Feature Agglomeration, which employs agglomerative clustering to group similar features together, reducing the number of features. This technique is particularly useful for dimensionality reduction, helping simplify complex datasets. Agglomerative Clustering offers several linkage strategies, including Ward, single, average, and complete. However, when it comes to cluster sizes, Agglomerative Clustering exhibits a "rich get richer" behavior, meaning some clusters may end up much larger than others. Among these strategies, single linkage tends to result in the most uneven cluster sizes, with Ward leading to the most uniform cluster sizes.While Ward is effective with Euclidean distances, it's not adaptable to non-Euclidean metrics. For such cases, average linkage serves as a good alternative. Single linkage, on the other hand, is not particularly robust when dealing with noisy data, but it is exceptionally efficient and is well-suited for hierarchical clustering of larger datasets. Single linkage can also handle non-globular data structures effectively.

Evaluation of clustering model

Davies-Bouldin Index (DBI) Evaluation

The evaluation of clustering quality is a fundamental aspect of unsupervised machine learning and data analysis. In clustering, the primary objective is to group data points into clusters, where data points within a cluster are similar to each other, and data points in different clusters are dissimilar. Clustering methods can vary in their effectiveness in achieving this goal, and one crucial tool for assessing the quality of clustering is the Davies-Bouldin Index (DBI).The DBI is a widely used metric for quantifying the separation and compactness of clusters generated by clustering algorithms. It provides a means to evaluate the goodness of a clustering solution, allowing researchers and data analysts to make informed decisions regarding the choice of clustering methods.

The DBI is defined based on the following key principles:

1. Cluster Separation: The DBI measures the separation between different clusters. It evaluates how far apart the cluster centers are from each other. A smaller distance between cluster centers suggests better separation.
2. Cluster Compactness: The DBI also evaluates the compactness of individual clusters. It measures how close data points within the same cluster are to each other. Smaller distances within clusters indicate greater compactness.
3. Comparison of Clusters: The DBI compares each cluster to all other clusters. It calculates a similarity measure based on the distances between the cluster centers and the average cluster size. A lower DBI score indicates better clustering quality.

In this study, we employed the DBI to assess the quality of clustering results for various cancer types and drug treatments. The lower the DBI score, the better the clustering quality. Higher DBI scores may suggest that the clusters are less separated and compact.In the subsequent sections, we will present the clustering results for specific cancer types and drugs, along with their corresponding DBI scores. These scores will serve as an essential benchmark for evaluating the performance of different clustering methods and their effectiveness in uncovering meaningful patterns within the datasets.

The DBI, along with visual representations of clustering results, offers a comprehensive understanding of the suitability of clustering methods in the context of cancer research. It enables us to make informed decisions about the most appropriate clustering techniques for different datasets, ultimately contributing to the advancement of personalized medicine and cancer treatment strategies.

Chapter 3 Results

In this section, we present the results of our unsupervised machine learning clustering analysis using K-means, Hierarchical Clustering, and Gaussian Mixture Model (GMM) Clustering on gene expression data for different cancer types and their response to specific drugs. The evaluation of these models is based on the Davies-Bouldin Index (DBI), with lower values indicating superior cluster quality.

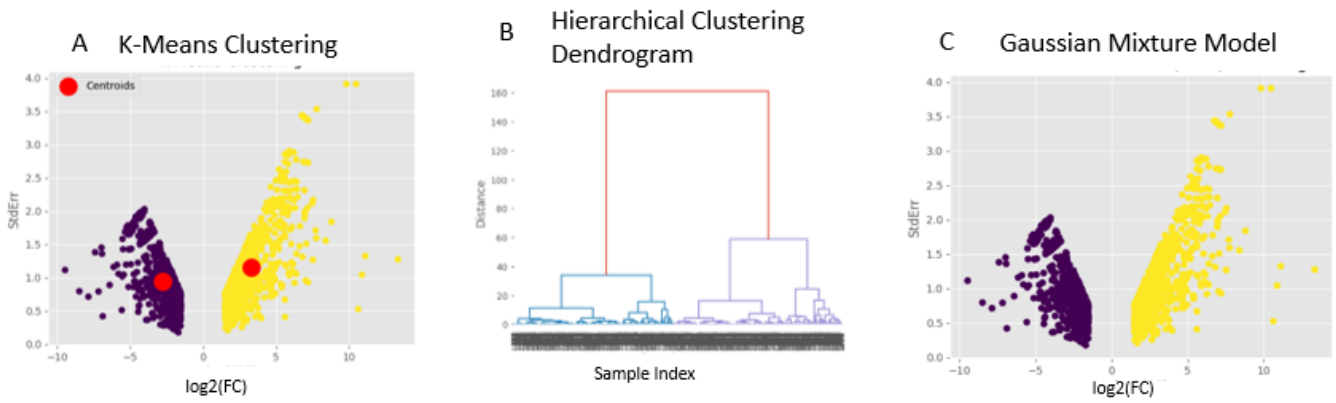


Figure 3 Breast Cancer data associated with the drug Paclitaxel

The clustering results for Breast Cancer data associated with the drug Paclitaxel are depicted in the plot above Figure 3. Three clustering methods, K-means(A), Hierarchical(B) Clustering, and Gaussian Mixture Model Clustering(C), were applied to this dataset.

- K-means Clustering: K-means and Hierarchical Clustering both achieved effective separation of data points with DBI scores of 0.47 and 0.48, respectively. These scores indicate that both methods successfully grouped data points into distinct clusters.
- Hierarchical Clustering: The DBI score for Hierarchical Clustering was also low, suggesting good clustering quality.

- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering resulted in a higher DBI score of 0.53, implying that it may not have separated the clusters as effectively as the other methods.

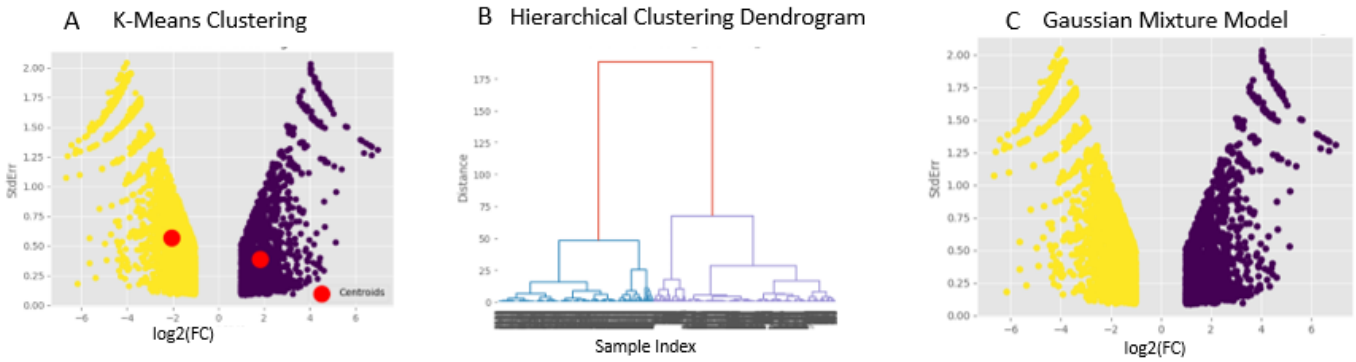


Figure 4 Pancreatic Cancer data associated with the drug Gemcitabine

The clustering results for Pancreatic Cancer data associated with the drug Gemcitabine are depicted in the plot above. Three clustering methods, K-means, Hierarchical Clustering, and Gaussian Mixture Model Clustering, were applied to this dataset.

- K-means Clustering: Both K-means and Hierarchical Clustering produced low DBI scores of 0.48 and 0.46, respectively, indicating effective separation of data points into distinct clusters with good compactness.
- Hierarchical Clustering: The DBI score for Hierarchical Clustering was also low, suggesting good cluster separation and compactness.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering resulted in a slightly higher DBI score of 0.57, indicating some complexity in cluster separation. This method may be appropriate for datasets with moderately distinct clusters.

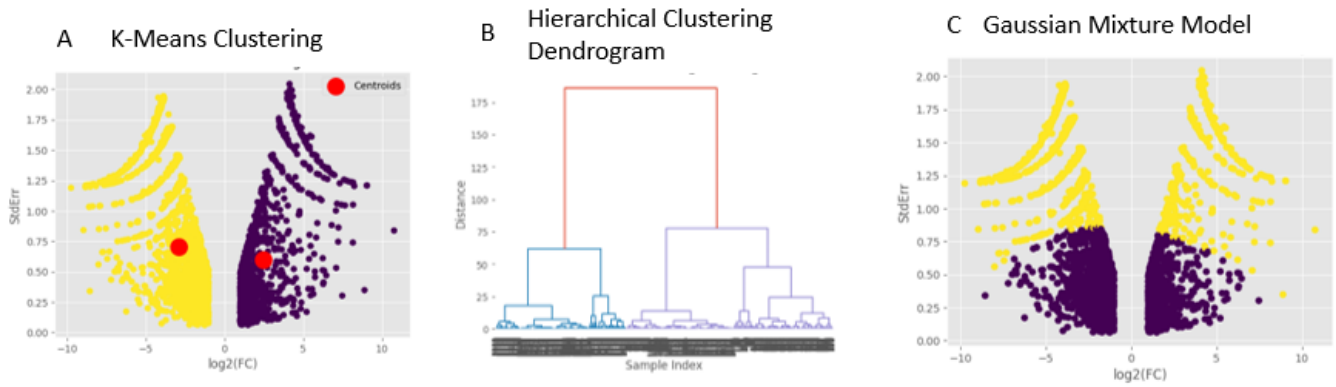


Figure 5 Leukemia data associated with the drug Etoposide

The clustering results for Leukemia data associated with the drug Etoposide are presented in the plot above. Three clustering methods, K-means, Hierarchical Clustering, and Gaussian Mixture Model Clustering, were applied to this dataset.

- K-means Clustering: K-means clustering resulted in a DBI score of 0.55, indicating effective separation of data points into clusters. The clusters are relatively well-defined, suggesting good cluster separation.
- Hierarchical Clustering: Hierarchical Clustering produced a slightly lower DBI score of 0.56, suggesting effective cluster separation with good compactness. The clusters are reasonably well-separated.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering yielded a DBI score of 0.57, indicating some complexity in cluster separation. This method may be suitable for datasets with moderately distinct clusters.

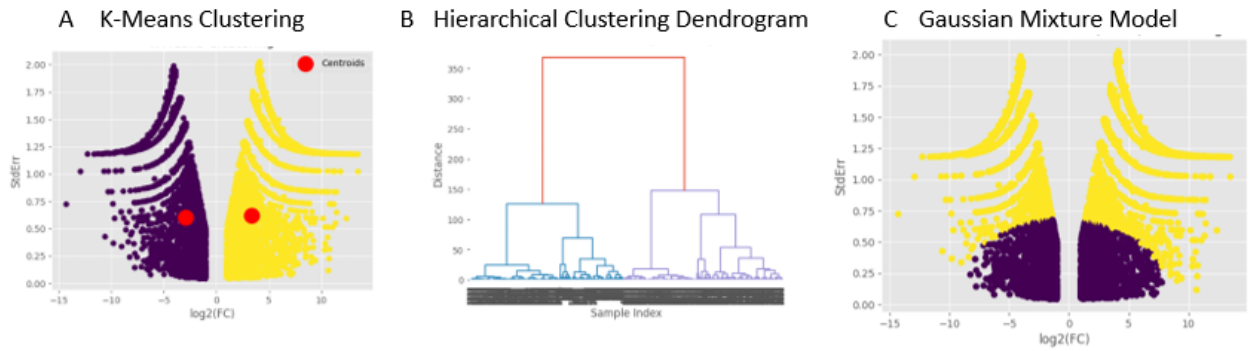


Figure 6 Ovarian Cancer data associated with the drug Doxorubicin

The clustering results for Ovarian Cancer data associated with the drug Doxorubicin are presented in the plot above figure 6. Three clustering methods, K-means(A), Hierarchical Clustering(B), and Gaussian Mixture Model Clustering(C), were applied to this dataset.

- K-means Clustering: K-means clustering produced a DBI score of 0.55, indicating reasonable separation of data points into clusters. The clusters are somewhat well-defined, although there may be some overlap or complexity.
- Hierarchical Clustering: Hierarchical Clustering resulted in a slightly higher DBI score of 0.57, suggesting that it may not have separated the clusters as effectively as K-means. The clusters may exhibit more overlap or intricacy.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering yielded a DBI score of 0.58, indicating a higher level of complexity in cluster separation. This method may be less suitable for datasets where clusters are closely interrelated.

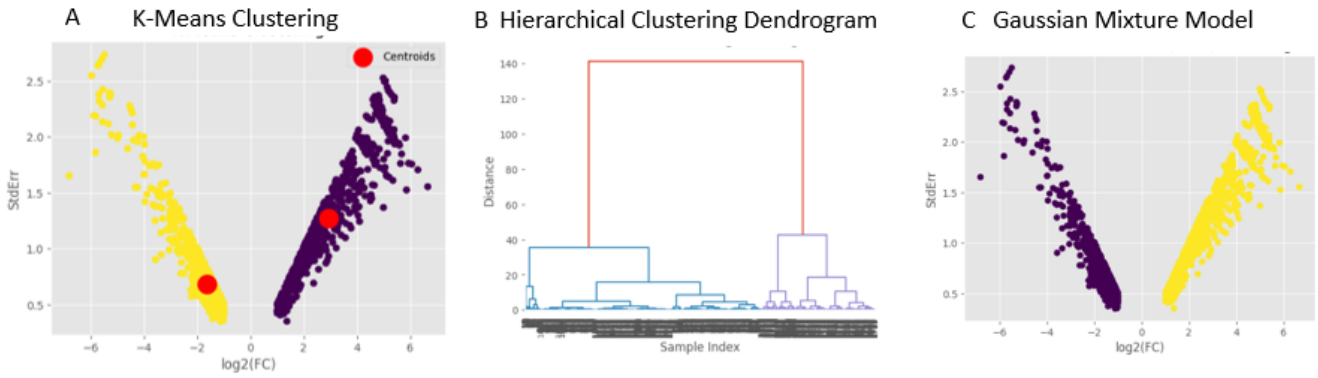


Figure 7 Prostate Cancer data related to the drug Docetaxel

The clustering results for Prostate Cancer data related to the drug Docetaxel are displayed in the plot above. Three clustering methods, K-means, Hierarchical Clustering, and Gaussian Mixture Model Clustering, were applied to this dataset.

- K-means Clustering: K-means achieved a DBI score of 0.43, suggesting relatively good separation of data points into distinct clusters. The clusters are reasonably well-defined.
- Hierarchical Clustering: Hierarchical Clustering produced a slightly lower DBI score of 0.42, indicating effective cluster separation with good compactness.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering yielded a DBI score of 0.54, indicating some complexity in cluster separation. This method may be suitable for datasets with moderately distinct clusters.

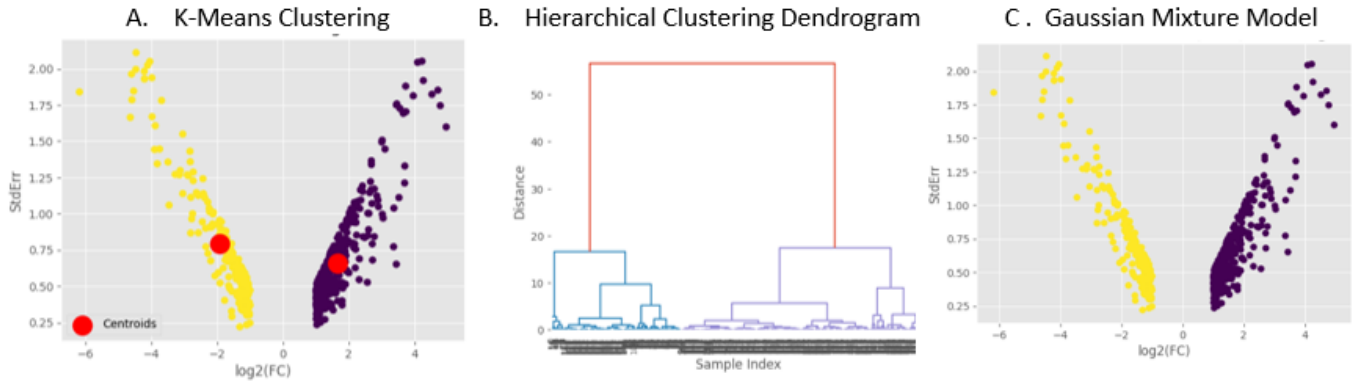


Figure 8 Breast Cancer data associated with the drug Doxorubicin

Doxorubicin The clustering results for Breast Cancer data associated with the drug Doxorubicin are shown in the above plot figure 8. Three different clustering methods, K-means(A), Hierarchical Clustering(B), and Gaussian Mixture Model Clustering(C), were applied to this dataset.

- K-means Clustering: The K-means clustering method achieved the best separation of data points, as indicated by the lowest Davies-Bouldin Index (DBI) score of 0.43. This suggests that K-means effectively grouped data points into distinct clusters in this specific dataset.
- Hierarchical Clustering: Hierarchical Clustering, while effective, produced a slightly higher DBI score of 0.52, indicating that it may not have separated the clusters as optimally as K-means. However, the clusters are still reasonably well-defined.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering yielded a higher DBI score of 0.57, suggesting that it may not have effectively separated the clusters. This method may be better suited for datasets with more complex or overlapping clusters.

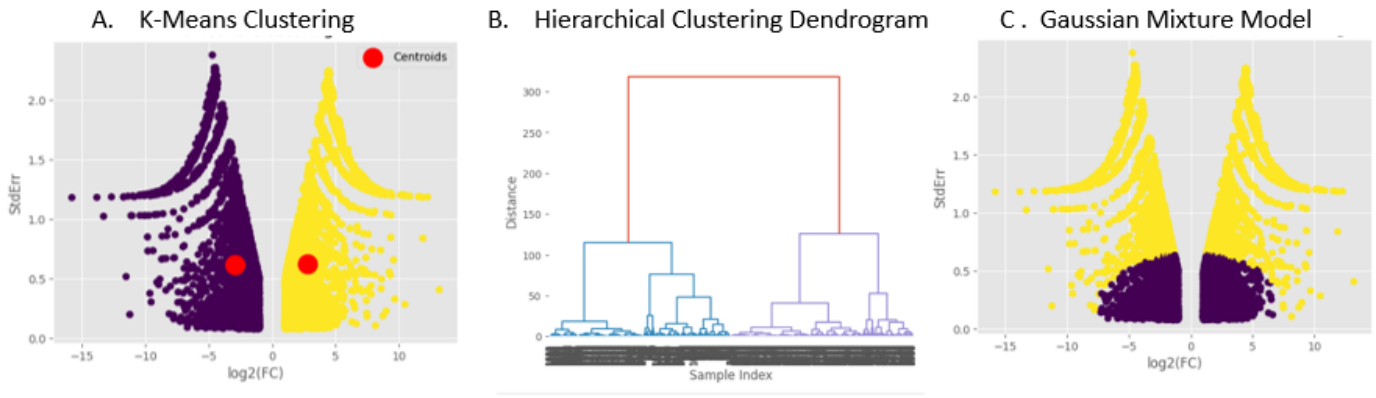


Figure 9 Breast Cancer data related to the drug Endoxifen

The clustering results for Breast Cancer data related to the drug Endoxifen are displayed in the plot above. Three clustering methods, K-means(A), Hierarchical Clustering(B), and Gaussian Mixture Model Clustering(C), were applied to this dataset.

- K-means Clustering: K-means achieved the best separation of data points with a DBI score of 0.56. This indicates that K-means effectively grouped data points into distinct clusters in this dataset.
- Hierarchical Clustering: Hierarchical Clustering produced a DBI score of 0.53, suggesting effective clustering, though slightly less optimal than K-means.
- Gaussian Mixture Model Clustering: Gaussian Mixture Model Clustering resulted in a higher DBI score of 0.62, implying that it may not have separated the clusters as effectively as the other methods.

Table 2 Performance of clustering model using Davies-Bouldin Index

Cancer Type	Drug Name	K-means Clustering	Hierarchial Clustering	Gaussian Mixture model Clustering
Breast Cancer	Doxorubicin	0.43	0.52	0.57
Breast Cancer	Endoxifen	0.56	0.53	0.62
Breast Cancer	Paclitaxel	0.47	0.48	0.53
Ovarian Cancer	Doxorubicin	0.55	0.57	0.58
Prostate Cancer	Docetaxel	0.43	0.42	0.54
Pancreatic Cancer	Gemicitadine	0.48	0.46	0.57
Leukemia	Etoposide	0.55	0.56	0.57

Chapter 4 Discussion

The major patterns observed in the study are the clustering results for various cancer types and drug treatments. These results show how the data points, representing genes or gene expression profiles, are grouped into distinct clusters using different clustering methods. The Davies-Bouldin Index (DBI) scores provide a quantitative measure of the quality of these clusters.

The relationships and trends among the results indicate that the effectiveness of clustering methods varies depending on the specific cancer type and drug treatment. In general, K-means clustering and Hierarchical Clustering consistently achieved lower DBI scores, suggesting better separation of data points. Gaussian Mixture Model Clustering, on the other hand, tended to result in higher DBI scores, indicating more complex or overlapping clusters. Exceptions to these patterns can be seen in the cases where Gaussian Mixture Model Clustering outperformed the other methods for certain datasets. These exceptions highlight the importance of selecting clustering methods based on the characteristics of the dataset. The likely causes underlying these patterns may relate to the nature of the data and the suitability of different clustering methods.

The study contributes to our understanding of the performance of clustering methods in the context of RNA-seq data analysis. It provides insights into the strengths and weaknesses of different clustering techniques for cancer-related gene expression data. Some datasets may have well-defined clusters that are better suited for K-means and Hierarchical Clustering, while others may have more complex structures that Gaussian Mixture Model Clustering can capture. By understanding the inherent patterns and structures within gene expression data, these methods can be harnessed to make predictions related to chemo-resistance or other critical outcomes. For instance, the clusters identified through these techniques can serve as features in

predictive models, enabling the development of models that forecast patient response to specific cancer treatments. This has the potential to revolutionize personalized medicine by aiding in treatment selection and customization. The integration of clinical data, such as patient demographics, treatment history, and genetic variations, with the clustering results holds promise for creating comprehensive predictive models. These models can offer valuable insights into the likely efficacy of chemo-resistant cancer drugs for individual patients, leading to more informed clinical decisions

Chapter 5 References

1. Ammad-ud-din, M, Khan, SA, Malani, D, Murumägi, A, Kallioniemi, O, Aittokallio, T, and Kaski, S (2016). Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics 32, i455–i463.

2. Azuaje, F (2017). Computational models for predicting drug responses in cancer research. Briefings in Bioinformatics 18, 820–829.

3. Babu, GR, Lakshmi, S, and Jotheeswaran, A (2013). Epidemiological Correlates of Breast Cancer in South India. Asian Pacific Journal of Cancer Prevention : APJCP 14, 5077–5083.

4. Barretina, J et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.

5. Basu, A et al. (2013). An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. Cell 154, 1151–1161.

6. Chaudhary, K, Poirion, OB, Lu, L, and Garmire, LX (2018). Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clinical Cancer Research 24, 1248–1259.

7. Chen, X, Shachter, RD, Kurian, AW, and Rubin, DL (2017). Dynamic strategy for personalized medicine: An application to metastatic breast cancer. Journal of Biomedical Informatics 68, 50–57.

8. Choi, M, Shi, J, Zhu, Y, Yang, R, and Cho, K-H (2017). Network dynamics-based cancer panel stratification for systemic prediction of anticancer drug response. Nat Commun 8, 1940.

9. Costello, JC et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol 32, 1202–1212.

10. Dagogo-Jack, I, and Shaw, AT (2018). Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15, 81–94.

11. DeSantis, CE, Ma, J, Gaudet, MM, Newman, LA, Miller, KD, Goding Sauer, A, Jemal, A, and Siegel, RL (2019). Breast cancer statistics, 2019. CA: A Cancer Journal for Clinicians 69, 438–451.

12. Ding, Z, Zu, S, and Gu, J (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics 32, 2891–2895.

13. Ferlay, J, Soerjomataram, I, Dikshit, R, Eser, S, Mathers, C, Rebelo, M, Parkin, DM, Forman, D, and Bray, F (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer 136, E359–E386.

14. Garnett, MJ et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575.

15. Gönen, M, and Margolin, AA (2014). Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics 30, i556–i563.

16. He, Y, Hoskins, JM, and McLeod, HL (2011). Copy number variants in pharmacogenetic genes. Trends in Molecular Medicine 17, 244–251.

17. Jayaraj, R et al. (2019). Clinical Theragnostic Relationship between Drug-Resistance Specific miRNA Expressions, Chemotherapeutic Resistance, and Sensitivity in Breast Cancer: A Systematic Review and Meta-Analysis. Cells 8, 1250.

18. Liu, Q, Borcherding, NC, Shao, P, Maina, PK, Zhang, W, and Qi, HH (2020). Contribution of synergism between PHF8 and HER2 signalling to breast cancer development and drug resistance. eBioMedicine 51.

19. Menden, MP, Casale, FP, Stephan, J, Bignell, GR, Iorio, F, McDermott, U, Garnett, MJ, Saez-Rodriguez, J, and Stegle, O (2018). The germline genetic component of drug sensitivity in cancer cell lines. Nat Commun 9, 3385.

20. Menden, MP, Iorio, F, Garnett, M, McDermott, U, Benes, CH, Ballester, PJ, and Saez-Rodriguez, J (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLOS ONE 8, e61318.

21. Nounou, MI, ElAmrawy, F, Ahmed, N, Abdelraouf, K, Goda, S, and Syed-Sha-Qhattal, H (2015). Breast Cancer: Conventional Diagnosis and Treatment

Modalities and Recent Patents and Technologies. Breast Cancer﷯(Auckl) 9s2, BCBCR.S29420.

22. Porter, PL (2009). Global trends in breast cancer incidence and mortality. Salud Pública de México 51, s141–s146.

23. Rampášek, L, Hidru, D, Smirnov, P, Haibe-Kains, B, and Goldenberg, A (2019). Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. Bioinformatics 35, 3743–3751.

24. Sharma, A, and Rani, R (2018). KSRMF: Kernelized similarity based regularized matrix factorization framework for predicting anti-cancer drug responses. Journal of Intelligent & Fuzzy Systems 35, 1779–1790.

25. Sharma, A, and Rani, R (2020). Ensembled machine learning framework for drug sensitivity prediction. IET Systems Biology 14, 39–46.

26. Siegel, RL, Miller, KD, and Jemal, A (2018). Cancer statistics, 2018. CA: A Cancer Journal for Clinicians 68, 7–30.

27. Vulsteke, C et al. (2013). Genetic variability in the multidrug resistance associated protein-1 (ABCC1/MRP1) predicts hematological toxicity in breast cancer patients receiving (neo-)adjuvant chemotherapy with 5-fluorouracil, epirubicin and cyclophosphamide (FEC). Annals of Oncology 24, 1513–1525.

28. Wang, C, Machiraju, R, and Huang, K (2014). Breast cancer patient stratification using a molecular regularized consensus clustering method. Methods 67, 304–312.

29. Wang, J, Seebacher, N, Shi, H, Kan, Q, and Duan, Z (2017). Novel strategies to prevent the development of multidrug resistance (MDR) in cancer. Oncotarget 8, 84559–84571.

30. Wang, X, Zhang, H, and Chen, X (2019). Drug resistance and combating drug resistance in cancer. Cancer Drug Resist 2, 141–160.

31. Wu, Q, Yang, Z, Nie, Y, Shi, Y, and Fan, D (2014). Multi-drug resistance in cancer chemotherapeutics: Mechanisms and lab approaches. Cancer Letters 347, 159–166.

32. Xia, F et al. (2018). Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinformatics 19, 486.

33. Yang, W et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research 41, D955–D961.

34. Zhang, F, Wang, M, Xi, J, Yang, J, and Li, A (2018a). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep 8, 3355.

35. Zhang, X, Li, B, Han, H, Song, S, Xu, H, Hong, Y, Yi, N, and Zhuang, W (2018b). Predicting multi-level drug response with gene expression profile in multiple myeloma using hierarchical ordinal regression. BMC Cancer 18, 551.

36. Zhu, Y, Brettin, T, Evrard, YA, Xia, F, Partin, A, Shukla, M, Yoo, H, Doroshow, JH, and Stevens, RL (2020). Enhanced Co-Expression Extrapolation (COXEN) Gene Selection Method for Building Anti-Cancer Drug Response Prediction Models. Genes 11, 1070.