International Journal for Research in Biology and Pharmacy

GREEN PUBLICATION

ISSN (Online): 2208-2093
Volume 10 Issue 03 September 2024
Available on-line at: https://ijrbp.com/index.php/bp/issue/view/559

DOI:*https://doi.org/10.69980/bp.v10i3.2492* *Received:12-Jun-2024 Accepted: 28-Aug-2024*

COMPUTATIONAL ANALYSIS OF GENOMIC CHARACTERIZATION AND DRUG RESPONSE AVAILABILITY IN CANCER CELL LINES FOR PERSONALIZED CANCER THERAPY

Dr. Habiba Alsafar[1], Dr. Basma AlBlooshi[2], Dr. Ayesha Salem AlDhaheri[3], Dr. Salem Y. Al Kaabi[4]

[1] Center for Biotechnology, Khalifa University, Abu Dhabi, United Arab Emirates
[2] Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates
[3] College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates
[4] Department of Oncology, Tawam Hospital, Al Ain, United Arab Emirates

Corresponding Author
Email: habiba.alsafar@ku.ac.ae

Abstract
Computational analysis of genomic characterization and drug response availability in cancer cell lines for personalized cancer therapy was conducted to evaluate the suitability of a cancer cell line dataset for descriptive pharmacogenomic research. The study analyzed 1001 cancer cell lines using variables related to cell line identity, genomic profiling, drug response availability, cancer classification, microsatellite instability status, screen medium, and growth properties. Genomic characterization was assessed through whole-exome sequencing, copy number alterations, gene expression, methylation, and MSI status, while drug response was examined as an availability-based variable. Descriptive statistical methods were used to calculate frequencies and percentages for genomic data availability, drug response availability, cancer type distribution, tissue descriptors, MSI status, and culture characteristics. The results showed strong genomic coverage across the dataset, with most cell lines containing major molecular characterization variables. Drug response availability was also high, supporting the usefulness of the dataset for pharmacogenomic model selection. Integrated analysis identified a large subset of cancer cell lines with both complete genomic characterization and drug response availability, indicating their relevance for future personalized therapy studies. The dataset also showed broad representation across cancer types and tissue groups. The main limitation was the absence of quantitative drug sensitivity measures such as IC50, AUC, or dose-response values. Overall, the dataset provides a useful foundation for computational oncology and personalized cancer therapy research.

Keywords: *Cancer cell lines, Genomic characterization, Drug response availability, Pharmacogenomics, Personalized cancer therapy*

Introduction

However, the threat of cancer continues to be among the foremost public health concerns, considering the rising incidence and death rate in many parts of the world. The monitoring of cancer trends worldwide has demonstrated that the disease burden is still rising owing to population increase, aging, changes in lifestyle and environmental exposure, and inequities in accessing preventive, diagnostic, and therapeutic measures (Sung et al., 2021). Moreover, current global figures indicate that cancer is not only among the top causes of mortality but also an escalating medical and economic problem (Bray et al., 2024). Historically, the conventional strategies employed in cancer treatment included location, histology, and chemotherapy. While these methods have continued to play their roles, they fail to address the diverse nature of the tumors among the different patients and types of cancers. The gaps created here have raised the level of interest in personalized cancer treatment based on genomics, molecular structure, and pharmacological features of the tumors.

Precision oncology has ushered in an era of revolutionary change in cancer treatment by taking oncological treatment from a generalized approach to a molecular-based approach. This approach includes the identification of genetic mutations, classification of patients based on molecular characteristics, and molecular-based treatment approaches for each unique type of cancer (Tsimberidou et al., 2020). The field of molecular diagnostics currently occupies a very significant position in oncology through its contribution to the identification of mutations that could be targeted, the presence of resistant genes, and various biomarkers useful for treating cancers (Riedl et al., 2024). Understanding the biological nature of cancer, as well as making decisions on how to treat it, requires genome sequencing.

The use of cancer cell lines is a common practice in oncological research since such cell lines provide an experimental model to understand the nature of the cancer, variations in its genetics, and drug sensitivities. Pharmacogenomics has made extensive use of cancer cell lines to make connections between molecular markers and drug sensitivity profiles (Feng et al., 2021). The significance of these models lies in the fact that they enable scientists to study the impact that certain genetic or molecular features have on the susceptibility and resistance to anti-cancer drugs. Datasets based on cell lines are particularly relevant in computational oncology due to the presence of structured data about the tissue source, genomic testing, and drug screening capabilities. The use of predictive modeling to predict patient drug sensitivity is one of the critical research domains within cancer pharmacogenomics. In the recent past, various computational approaches have been used to predict drug sensitivity based on genomic data and other omics data types. Additionally, machine learning techniques have been used to model correlations in cancer cell lines and xenografts datasets for drug sensitivities (Kurilov et al., 2020). In the process of designing predictive models for drug sensitivities, some guidelines recommend that special attention should be placed on data preprocessing and model validation, as well as using relevant biological predictors (Sharifi-Noghabi et al., 2021). Drug sensitivity prediction has also been found to be dataset-, platform-, and model-dependent (Xia et al., 2022). The use of multi-omics integration for computational drug response prediction has extended the scope of the field. While genomic changes are not sufficient to define the treatment effects, the presence of gene expression, epigenetics, methylation levels, and pathways is critical for determining the drug response. The effectiveness of drugs based on the application of omics technology was shown using the methods of computational analysis, where drug response-related methylation sites were predicted (Yuan et al., 2020). It was also suggested that the deep learning approach, as well as a multi-omics model, can be used for improving the performance of drug response prediction using complicated nonlinear relations between genome, transcriptome, and epigenome features (Wang et al., 2021; Sharma et al., 2023). MSI can also be regarded as one more important genomic feature in terms of cancer profiling. MSI has been associated with DNA mismatch repair gene mutations, implying its substantial significance in biological, immune, and clinical contexts (Greco et al., 2023). The incorporation of MSI information alongside sequencing, copy number alterations, gene expression data, and methylation profiling can contribute to the molecular characterization of cancer cell lines. About individual cancer treatment, MSI information might prove useful in distinguishing separate biological categories to be analyzed individually.

Objectives of the Study

1. To analyze the availability of genomic characterization data, including whole exome sequencing, copy number alterations, gene expression, methylation, and microsatellite instability status, across cancer cell lines.
2. To evaluate drug response availability in cancer cell lines and identify the subset of cell lines suitable for integrated pharmacogenomic analysis.
3. To examine cancer type distribution, tissue classification, and culture-related characteristics to assess the dataset's relevance for personalized cancer therapy research.

Methodology

Study Design

The genomic characterization and drug response data used for analysis in this study were obtained using computational descriptive analysis methods. This data was collected from the Cell_Lines_Details(1).xlsx file, containing a total of 1002 cancer cell lines having 13 variables. The variables considered in this dataset include information about cell lines, genomics, drug responses, cancer types, MSI, and culture conditions (Alipour, n.d.).

Dataset Description

The first dataset had attributes such as sample ID, COSMIC ID, whole exome sequencing, copy number change, gene expression, DNA methylation, response to drugs, tissue description, TCGA cancer category, microsatellite instability

condition, screen media, and growth properties. Such attributes have been classified into five major attribute groups: cell line, genomic characterization, drug response, cancer type, and growth properties.

Data Preprocessing
Before performing an analysis on the data set, missing values, duplicates, and inconsistent column names were addressed. Any column names that had a line break or any extra space were normalized. Within the genome and drug response variables, the presence of Y meant the value was available; anything else was missing.

Genomic Characterization Analysis
Genomic profiling analysis was performed through WES, CNA, gene expression, DNA methylation, and MSI. The presence of each genomic characteristic was determined by the number and percentage of cell lines possessing that characteristic.

Drug Response Availability Analysis
Drug response availability was analyzed using the drug response variable. The value **Y** was interpreted as the presence of drug response data, while blank entries were treated as unavailable.

Integrated Genomic and Drug Response Analysis
A comprehensive study was conducted to detect cancer cell lines that have genomic information and drug sensitivity data. Cancer cell lines that possess WES, CNA, gene expression, methylation, and drug sensitivity data were deemed appropriate for personalized cancer treatment studies.

Cancer Type and Tissue Classification Analysis
Cancer type distribution was analyzed using GDSC tissue descriptors and TCGA cancer type labels. This helped identify the major cancer types represented in the dataset.

Microsatellite Instability Analysis
MSI status was examined because it is an important genomic instability marker associated with cancer classification and therapy response. The distribution of MSI categories was summarized across the cancer cell lines.

Statistical Analysis
Descriptive statistics were used to summarize the dataset. Frequencies and percentages were calculated for genomic data availability, drug response availability, cancer type distribution, MSI status, screen medium, and growth properties.

Data Visualization
Presentation of the findings may be done through tables, bar charts, pie charts, and heat maps. This will present genomic data availability, availability of drug response, types of cancer, MSI status, and genomics-drug response patterns integration.

Results
Genomic Characterization Availability
The measures to characterize the genome used were whole-exome sequencing, copy number alteration, gene expression, and methylation measurements. The availability of these genomics characterizations was found to be very high among the cancer cell lines. Of the four genomic characterization measures, whole exome sequencing showed the highest availability, followed by copy number alteration, gene expression, and methylation data, respectively. This means that the majority of the cancer cell lines were well-characterized and ready for genomics analysis. As shown in Figure 1, genomic characterization data were highly available, especially whole-exome sequencing.

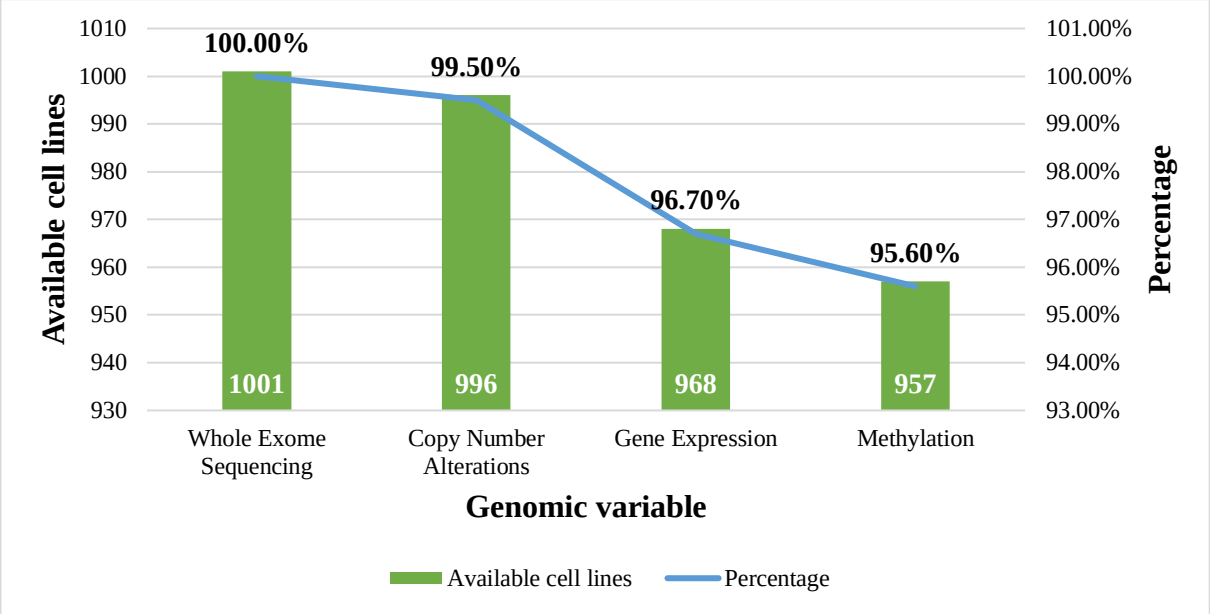

Figure 1. Availability of Genomic Characterization Data Across Cancer Cell Lines

Drug Response Availability

Drug response availability was determined using the drug response variable. The results indicated that drug response data were available for the majority of cancer cell lines, whereas only a few lacked such data. Therefore, this dataset seems to have great potential for use in pharmacogenomics. Nonetheless, it should be noted that drug response is a measure of drug response data availability and not a quantitative measure of drug sensitivity. Table 2 indicates that drug response data were available for the majority of cancer cell lines.

Table 2. Drug Response Availability

Drug response status	Cell lines	Percentage
Available	990	98.9%
Not available	11	1.1%
Total	1001	100.0%

Integrated Genomic and Drug Response Availability

Integrated analyses were used to find the cancer cell lines with the availability of their full genomics profile, along with their drug responses. Cancer cell lines having data regarding whole exome sequencing, copy number variation, gene expression, methylation, and drug response were found to be the best candidates for conducting research in the domain of personalized cancer therapies. It is evident from the analysis that a majority of cell lines in the dataset had both complete genomic information and drug responses. Table 3 provides further evidence of this observation.

Table 3. Integrated Genomic and Drug Response Availability

Integrated data category	Cell lines	Percentage
Complete genomic characterization	934	93.3%
Complete genomic characterization with drug response availability	928	92.7%
Complete genomic characterization, MSI status, and drug response availability	927	92.6%

Tissue Descriptor Distribution

Cancer cell lines were categorized based on the GDSC tissue description. The dataset was seen to have a good representation from various tissues. Some of the groups that were well represented included lung cancer, cancers of the urogenital system, leukemia, cancers of the aerodigestive tract, lymphoma, skin cancer, nervous system tumors, breast cancer, cancers of the digestive system, and cancers of the large intestine. Table 4 shows good representation across tissues from major cancer groups.

Table 4. Major GDSC Tissue Descriptor Categories

GDSC tissue descriptor 1	Cell lines	Percentage
lung_NSCLC	111	11.1%
urogenital_system	105	10.5%

leukemia	85	8.5%
aero_dig_tract	79	7.9%
lymphoma	70	7.0%
lung_SCLC	66	6.6%
skin	58	5.8%
nervous_system	57	5.7%
breast	52	5.2%
digestive_system	52	5.2%
large_intestine	51	5.1%

TCGA Cancer Type Distribution

The classification of cancer type was also explored through the use of TCGA-matched cancer type labels. There were several cancer types represented in the data set, such as lung cancer, melanoma, breast cancer, colon cancer, head and neck cancer, glioblastoma, esophageal cancer, lymphoma, ovarian cancer, kidney cancer, neuroblastoma, pancreatic cancer, and leukemia. Some of the cell lines did not have TCGA cancer type labels attached to them, a factor that needs to be considered while making cancer-type-based inferences. Table 5 shows that multiple TCGA cancer types were represented, with some missing cancer-type labels.

Table 5. Major TCGA Cancer Type Categories

TCGA cancer type	Cell lines	Percentage
Missing TCGA label	175	17.5%
SCLC	66	6.6%
LUAD	64	6.4%
SKCM	55	5.5%
BRCA	51	5.1%
COAD/READ	51	5.1%
HNSC	42	4.2%
GBM	36	3.6%
ESCA	35	3.5%
DLBC	35	3.5%
OV	34	3.4%
KIRC	32	3.2%
NB	32	3.2%
PAAD	30	3.0%
LAML	28	2.8%

Microsatellite Instability Status

The status of microsatellite instability was considered a genotypic instability marker for purposes of therapy. Most cancer cell lines could be described as microsatellite stable or MSI-low, whereas a small proportion could be described as MSI-high. This is critical since MSI status may correlate with biological features of tumors and their treatment options. However, some data on MSI status were missing, although it is significant and must be considered. According to Figure 2, most cancer cell lines were MSS/MSI-L, MSI-H, and MSI-missing, being less numerous.

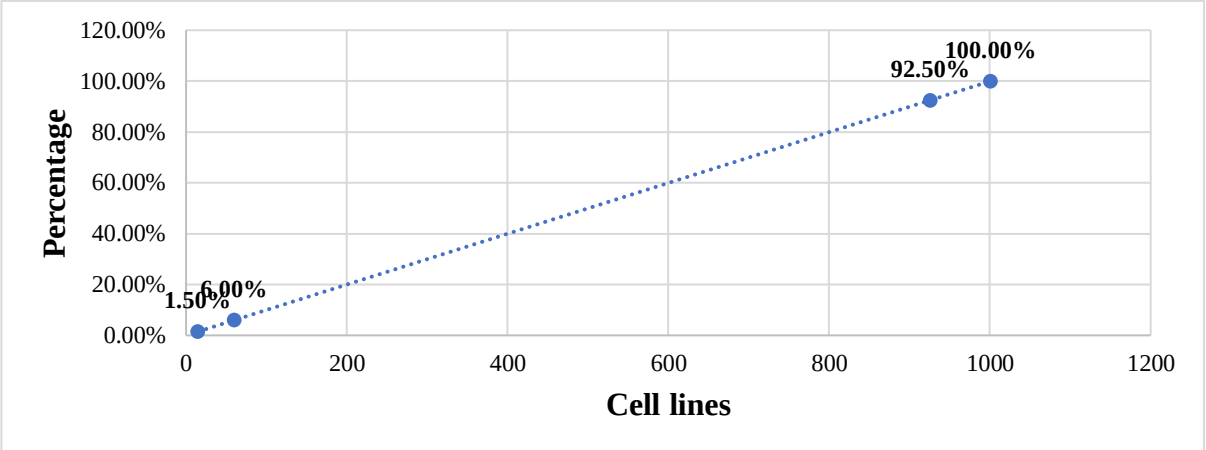

Figure 2. Distribution of Microsatellite Instability Status Among Cancer Cell Lines

Screen Medium Distribution

Screen medium was considered in order to establish experimental culture conditions related to the cancer cell lines. There are two types of screen media in the dataset, and both are well-represented. In this case, the screen medium is not considered a variable for interpreting personalized therapy but as experimental metadata. As can be seen from Table 7, R and D/F12 screen mediums have been used.

Table 7. Screen Medium Distribution

Screen medium	Cell lines	Percentage
R	557	55.6%
D/F12	444	44.4%
Total	1001	100.0%

Growth Properties Distribution

Cancer cell line growth properties were analyzed to describe the behavior of cancer cell lines. The majority of cell lines demonstrated adherent growth, while others demonstrated suspended or semi-adherent growth. Information on the growth properties of cell lines serves as important experimental background information, although not as a personalized medicine parameter. Based on Table 8, adherent growth was the dominant growth property of the cell lines.

Table 8. Growth Properties Distribution

Growth property	Cell lines	Percentage
Adherent	725	72.4%
Suspension	244	24.4%
Semi-Adherent	30	3.0%
NA	2	0.2%
Total	1001	100.0%

Discussion

In this regard, the present analysis shows that the available dataset presents a good basis for using computational approaches in the characterization of genomics and drug response. This is because of the high availability of data on whole-exome sequencing, copy number alterations, gene expression, methylation, and drug response measures. These features make the current dataset good for pharmacogenomic description, as well as the selection of cancer cell lines that can be used in further studies in personalized cancer treatments. The present results support the growing use of combined molecular and pharmacological datasets in precision oncology (Adam et al., 2020; Mechahougui et al., 2024).

One of the key advantages of the provided dataset is the extensive availability of genomic characterization data. Whole exome sequencing helps in identifying mutations, while the copy number alteration data is indicative of structural genomics data that can play a part in activating the oncogenes and suppressing the tumor. Gene expression and methylation analysis further help in providing insight into the transcriptomic and epigenomic aspects, respectively. Multi-omics studies have gained popularity in recent times due to their potential to improve our understanding of cancer drug responses. A combination of genomic, transcriptomic, epigenomic, and image-based information has proven to enhance the prediction and analysis of cancer drug response through multi-omics and machine learning techniques (Li et al., 2023; Partin et al., 2023).

There was also drug response availability in the data set, implying that most of the cell lines can be associated with pharmacological screening results. This is significant since drug response prediction requires the joint analysis of molecular properties together with the sensitivity to treatments. There have been many applications of machine learning and deep learning algorithms to this kind of issue using methods such as graph convolutional networks (Liu et al., 2020; Yu & Fan, 2025). The current dataset can therefore support the selection of well-characterized cell lines for future computational modeling, although the present analysis is limited to drug response availability rather than actual quantitative response values.

The comprehensive analysis revealed a significant proportion of cell lines with full genomics and drug sensitivity information available. This population is considered to be the most pertinent one for studies on personalized medicine for cancers since it provides the integration of biological and pharmacological information necessary for precision oncology, where the goal is to associate molecular changes with vulnerabilities. The recent developments in personalized medicine in oncology stressed the importance of having biologically relevant models to enable biomarker discovery and translational medicine (Masina & Caldas, 2024; Mechahougui et al., 2024).

From the analysis of the types of cancers as well as the tissues described, it can be deduced that the database comprises different forms of cancers. For instance, there were instances where cancer types like lung cancer, leukemia, lymphoma, breast cancer, melanoma, gastrointestinal cancer, and genitourinary cancer were found. This will make the comparison of

several kinds of cancer possible. This is important as the drug responses by cells are dependent on their cell lineages as well as the molecular processes involved in cancer formation itself (Carli et al., 2025).

Another feature of the data set was the microsatellite instability (MSI). Most of the cell lines belonged to the MSS/MSI-L type, whereas there were only a few that belonged to the MSI-H category. This feature assumes importance because, as far as defective DNA mismatch repair is concerned, microsatellite instability plays an important role. Although these form a small part of the data set, they need to be taken into consideration because of their significance.

Culture-related variables, including the media used and the growth features, provide useful background information for experimental studies. Adherent cells are common among most of the cell lines, whereas fewer adherent cells have been found in suspended and semi-adherent cell types. Even though they do not influence the individuality of treatment, they are likely to influence experimental procedures and drug testing in vitro.

The major disadvantage of the data set is that the field indicating the drug response seems to have data about whether there are drugs or not, rather than drug sensitivity. Therefore, the data set does not have any information on the drug sensitivity or resistance, or even any other form of IC50 or AUC. Nevertheless, it has good data that could be useful for personalized cancer therapy. (Adam et al., 2020; Partin et al., 2023; Yu & Fan, 2025).

Conclusion

In this study, the association between genomic characterization and drug response availability was assessed through the analysis of the relevance of using cancer cell lines for personalized cancer therapy research. The results of the investigation proved that there is a presence of good coverage of genomics in the dataset in terms of whole exome sequencing, copy number changes, gene expression, methylation, and microsatellite instability status. Moreover, drug response availability is relatively good, which means that it is highly possible that almost all cell lines can be used to conduct pharmacogenomic studies. In addition, an integrated approach revealed a considerable set of cancer cell lines, characterized by a complete set of genomic information and data availability regarding drug responses, which makes them perfect for further computational research. Also, tissue descriptors, TCGA cancer types, culture information about the screen medium, and growth were included in the dataset. However, the major drawback is that information about drug responses is provided only as availability, while the quantitative measures are not provided (IC50, AUC).

References

1. Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, *4*(1), 19.
2. Alipour, S. (n.d.). *Genomics of Drug Sensitivity in Cancer (GDSC)* [Data set]. Kaggle.
3. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229-263.
4. Carli, F., Di Chiaro, P., Morelli, M., Arora, C., Bisceglia, L., De Oliveira Rosa, N., ... & Raimondi, F. (2025). Learning and applying general principles of cancer cell drug sensitivity. *Nature Communications*, *16*(1), 1654.
5. Feng, F., Shen, B., Mou, X., Li, Y., & Li, H. (2021). Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine. *Journal of Genetics and Genomics*, *48*(7), 540-551.
6. Greco, L., Rubbino, F., Dal Buono, A., & Laghi, L. (2023). Microsatellite instability and immune response: from microenvironment features to therapeutic actionability—lessons from colorectal cancer. *Genes*, *14*(6), 1169.
7. Kurilov, R., Haibe-Kains, B., & Brors, B. (2020). Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific reports*, *10*(1), 2849.
8. Li, Y., Guo, Z., Gao, X., & Wang, G. (2023). Mmcl-cdr: enhancing cancer drug response prediction with multi-omics and morphology images contrastive representation learning. *Bioinformatics*, *39*(12), btad734.
9. Liu, Q., Hu, Z., Jiang, R., & Zhou, M. (2020). DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, *36*(Supplement_2), i911-i918.
10. Masina, R., & Caldas, C. (2024). Precision Cancer Medicine 2.0—Oncology in the postgenomic era. *Molecular Oncology*, *18*(9), 2065-2069.
11. Mechahougui, H., Gutmans, J., Colarusso, G., Gouasmi, R., & Friedlaender, A. (2024). Advances in personalized oncology. *Cancers*, *16*(16), 2862.
12. Partin, A., Brettin, T. S., Zhu, Y., Narykov, O., Clyde, A., Overbeek, J., & Stevens, R. L. (2023). Deep learning methods for drug response prediction in cancer: predominant and emerging trends. *Frontiers in medicine*, *10*, 1086097.
13. Riedl, J. M., Moik, F., Esterl, T., Kostmann, S. M., Gerger, A., & Jost, P. J. (2024). Molecular diagnostics tailoring personalized cancer therapy—an oncologist's view. *Virchows Archiv*, *484*(2), 169-179.
14. Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., ... & Haibe-Kains, B. (2021). Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in bioinformatics*, *22*(6), bbab294.
15. Sharma, A., Lysenko, A., Boroevich, K. A., & Tsunoda, T. (2023). DeepInsight-3D architecture for anti-cancer drug response prediction with deep-learning on multi-omics. *Scientific reports*, *13*(1), 2483.
16. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

17. Tsimberidou, A. M., Fountzilas, E., Nikanjam, M., & Kurzrock, R. (2020). Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer treatment reviews*, *86*, 102019.
18. Wang, C., Lye, X., Kaalia, R., Kumar, P., & Rajapakse, J. C. (2021). Deep learning and multi-omics approach to predict drug responses in cancer. *BMC bioinformatics*, *22*(Suppl 10), 632.
19. Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., ... & Stevens, R. (2022). A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in bioinformatics*, *23*(1), bbab356.
20. Yu, G., & Fan, Q. (2025). Deep learning-driven drug response prediction and mechanistic insights in cancer genomics. *Scientific Reports*, *15*(1), 20824.
21. Yuan, R., Chen, S., & Wang, Y. (2020). Computational prediction of drug responses in cancer cell lines from cancer omics and detection of drug effectiveness related methylation sites. *Frontiers in Genetics*, *11*, 917.