ADVANCED SCIENCE Open Access
www.advancedscience.com

DrugFormer: Graph-Enhanced Language Model to Predict Drug Sensitivity

Xiaona Liu, Qing Wang, Minghao Zhou, Yanfei Wang, Xuefeng Wang, Xiaobo Zhou, and Qianqian Song**

Drug resistance poses a crucial challenge in healthcare, with response rates to chemotherapy and targeted therapy remaining low. Individual patient's resistance is exacerbated by the intricate heterogeneity of tumor cells, presenting significant obstacles to effective treatment. To address this challenge, DrugFormer, a novel graph-augmented large language model designed to predict drug resistance at single-cell level is proposed. DrugFormer integrates both serialized gene tokens and gene-based knowledge graphs for the accurate predictions of drug response. After training on comprehensive single-cell data with drug response information, DrugFormer model presents outperformance, with higher F1, precision, and recall in predicting drug response. Based on the scRNA-seq data from refractory multiple myeloma (MM) and acute myeloid leukemia (AML) patients, DrugFormer demonstrates high efficacy in identifying resistant cells and uncovering underlying molecular mechanisms. Through pseudotime trajectory analysisunique drug-resistant cellular states associated with poor patient outcomes are revealed. Furthermore, DrugFormer identifies potential therapeutic targets, such as COX8A, for overcoming drug resistance across different cancer types. In conclusion, DrugFormer represents a significant advancement in the field of drug resistance prediction, offering a powerful tool for unraveling the heterogeneity of cellular response to drugs and guiding personalized treatment strategies.

1. Introduction

Drug resistance stands as a significant challenge in healthcare,[1] underscored by alarming response rates observed in clinical studies. A meta-analysis of 570 phase II single-agent clinical trials revealed a median response rate to chemotherapy of merely 11.9%, and this rate is only 30% in the context of personalized targeted therapy.[2] Despite ongoing efforts addressing inter-patient heterogeneity of treatment effects, very limited attention is paid to the intra-patient heterogeneity resulting in drug resistance. For instance, intra-tumoral heterogeneity refers to the diversity of cancer cells within a tumor, which can exhibit different genetic, epigenetic, and phenotypic characteristics. This heterogeneity poses challenges for cancer therapy because different subpopulations of cancer cells may respond differently to drugs, leading to treatment tolerance and tumor recurrence. Therefore, it is crucial to interrogate the diversity of drug-resistant cancer cells and identify specific cellular subpopulations leading to patient-level resistance.

With the rapid development of technology, single-cell RNA sequencing (scRNA-seq) has become a revolutionary technique, providing high resolution for investigating tumor cell resistance at the cellular and cell type levels.[3–12] scRNA-seq has been applied to exploit drug resistance mechanisms, leading to the discovery of effective targets and the development of optimized therapeutic strategies. For example, Heo et.al. analyzed scRNA-seq data and identified the cytidine deaminase as the potential druggable target to eliminate resistant cells in lung cancer.[13] Li et al. revealed a subpopulation of quiescent stem-like cells that contributed to the chemoresistance and poor outcomes of AML.[14] Those advanced technologies and generated data

X. Liu, X. Zhou
Center for Computational Systems Medicine
McWilliams School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, TX 77030, USA
E-mail: Xiaobo.Zhou@uth.tmc.edu

Q. Wang, M. Zhou, Y. Wang, Q. Song
Department of Health Outcomes and Biomedical Informatics
College of Medicine
University of Florida
Gainesville, FL 32611, USA
E-mail: qsong1@ufl.edu

X. Wang
Biostatistics and Bioinformatics
H. Lee Moffitt Cancer Center and Research Institute
Tampa, FL, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202405861

DOI: 10.1002/advs.202405861

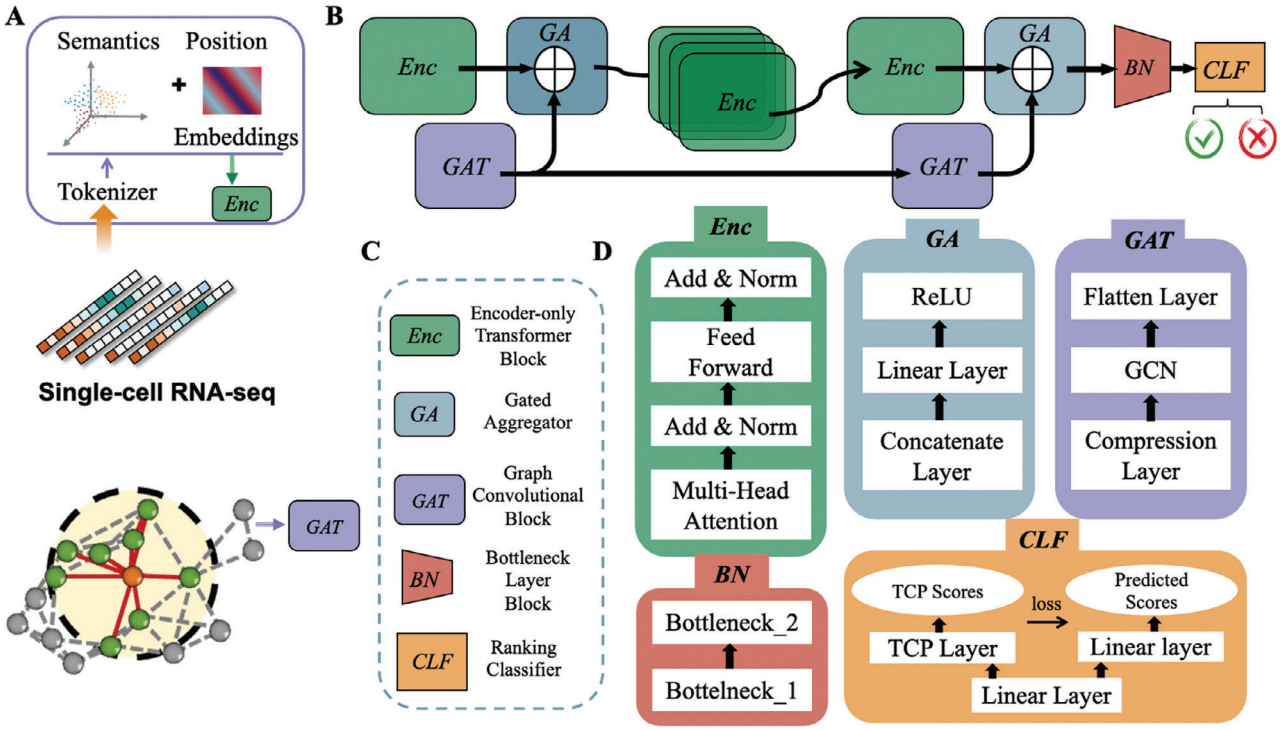

Figure 1. Overview of the DrugFormer model. DrugFormer integrates gene token representations and a gene-based knowledge graph for drug response prediction. A) Genes are serialized as inputs for the Transformer encoder, while a gene-based knowledge graph is constructed as input for the graph attention network. B) The overall framework of DrugFormer. C) Explanation of the four types of module components in DrugFormer. D) Detailed structure of each component.

resources have been collected in extensive datasets, such as the DRMref database,[15] which serves as a comprehensive reference map detailing drug resistance mechanisms in human cancer and provides a valuable resource for insightful analyses of drug resistance.

To delve into the drug resistance for individual patients, it is critical to develop tailored methods interrogating how cells in a complex tissue differentially respond to drugs. However, currently, there is a lack of advanced models designed for this purpose. Though several studies, such as DREEP,[16] scDEAL,[17] SCAD,[18] have leveraged existing drug screening databases[19,20] to investigate drug response at the single-cell level, these studies have significant limitations. First, they rely on cell line-based knowledge as the reference, lacking in vivo properties and failing to mimic real in vivo scenarios, which may result in poor prediction accuracy for cells in tissue samples. Second, their predictions are limited to certain drugs or compounds available in the drug screening databases,[19,20] lacking the generalization capability to predict cell responses to real-world applied drugs. To address those limitations, large language models (LLMs) have emerged as a promising solution for uncovering cellular responses to drugs. Given the existing scRNA-seq data collected for drug resistance research,[15] sophisticated large language models offer a tailored solution that not only aggregates information from extensive in vivo data resources, but also possesses the generalization capability to predict unknown cellular drug response in scRNA-seq data, facilitating a more comprehensive mechanistic understanding.

Since the BERT model,[21] the development of pre-trained large language models has become more and more successful in natural language processing, as well as in bioinformatics and biomedical areas. For example, OpenAI's GPT[22] series and Rostlab's ProtBert[23] have achieved great success. Meanwhile, as graphs can represent complex relationships and structures from biological data or external knowledge bases, through incorporating graph structures, graph-enhanced LLMs can leverage these intricate relationships, leading to a richer and more nuanced understanding of the data. This allows the LLMs to access and utilize graph information or external knowledge more effectively, thus improving their performance. Such graph-enhanced large language models can be refined to be more accurate, controllable, and adaptable across diverse tasks and application scenarios. In this work, we have proposed our DrugFormer model that leverages the large-scale drug resistance database[15] to achieve accurate predictions of cellular-level drug resistance. Such a cutting-edge model is designed with the primary objective of predicting drug resistance and unraveling novel therapeutic targets.

2. Results

2.1. Overview of DrugFormer

The overall framework of our proposed DrugFormer model is illustrated in **Figure 1**. DrugFormer integrates both gene representations and a knowledge graph through two processing paths. One path utilizes a Transformer encoder-based network to extract gene token representations. The other path leverages a knowledge graph utilizing the graph attention mechanism to extract

graph information. As shown in Figure 1A, we serialize genes as input for the Transformer encoder. Meanwhile, we construct a gene-based knowledge graph using haploinsufficiency (i.e., deletion intolerance) and triploidy sensitivity (i.e., duplication intolerance) information. Such gene-based knowledge graph serves as input for the graph attention network. Figure 1B illustrates the model framework, which consists of four types of modules. Specifically, *Enc* represents the Transformer encoder layer, *GAT* represents the graph attention network layer, *GA* represents the gated aggregation module, and *CLF* represents the output classifier. Figure 1C,D depict the name and specific structure of each network block. In this framework, gene tokens first pass through the Transformer encoder layer to obtain the gene-based latent representations. The gene-based graph is processed through the graph attention network to obtain the graph-based latent representations. Both representations are fused into a combined embedding, which serves as input of the gated aggregation module. Following this, four Transformer encoders are used for deep extraction of the fused embeddings. After an additional layer of graph attention, the fusion embeddings are further enriched with graph information and then serve as input for the output layer for drug response prediction. Through cross-validation on collected single-cells from DRMref,[15] DrugFormer outperforms other methods. Details are described in the Experimental Section.

2.2. Quantitative Evaluation of DrugFormer Performance

Since there were no generalized methods to predict cell responses to real-world applied drugs, here we included the "DrugFormer-" model (See Experimental Section) alongside three classic machine learning (ML) models (Random Forest: RF, Support Vector Machine: SVM, and Linear Regression: LR) for comparison. Given the collected single-cells with drug response labels from DRMref,[15] we applied a five-fold cross-validation approach, randomly dividing the dataset into five non-overlapping subsets. For each fold, the remaining four subsets were used to train the model. The performance comparison of these models is illustrated in **Figure 2**, with all results representing the average metrics from the five-fold cross-validation.

As shown in Figure 2A, DrugFormer outperformed other methods in AUC (DrugFormer: 0.975, DrugFormer-: 0.912, SVM: 0.718, RF: 0.703, LR: 0.602). Notably, in Figure 2B, DrugFormer presented higher accuracy score (acc: 0.932) than DrugFormer- (acc: 0.874), SVM (acc: 0.671), RF (acc: 0.659), and LR (acc: 0.573). For other metrics, including F1 score, Precision, and Recall, DrugFormer achieved values of 0.943, 0.958, 0.932, respectively. In contrast, DrugFormer- (F1: 0.886, Precision: 0.882, Recall: 0.880), SVM (F1: 0.683, Precision: 0.692, Recall: 0.688), RF (F1: 0.668, Precision: 0.678, Recall: 0.669), and LR (F1: 0.595, Precision: 0.578, Recall: 0.580) presented a poorer performance in these evaluation metrics. Given the lack of existing methods for comparison, we additionally identified two methods, scDEAL[17] and SCAD,[18] which used cell line data as reference data for predicting drug resistance. Through comparison on their provided datasets (GSE149383 and GSE117872), DrugFormer achieved a significantly higher AUC than both scDEAL and SCAD.

2.3. DrugFormer Unveils Resistant Cell State in Multiple Myeloma

To demonstrate the capability of DrugFormer, here we applied it to the scRNA-seq data profiled from refractory multiple myeloma (MM) patients who were subsequently treated with IMiD drugs.[24] Notably, DrugFormer identified the drug-resistant cells with F1 score as 0.939.

With DrugFormer recognizing the resistant cells, we next interrogated the cellular dynamics to understand the underlying heterogeneity resulted patient resistance. Herein, we identified the pseudotime trajectory to interpret the cellular states of malignant cells (**Figure 3**A). For each state on this trajectory, we calculated the stemness score and observed a progressive differentiation potential (Figure 3B). Moreover, based on the cell cycle-related markers, including 43 genes associated with the S phase and 54 genes associated with the G2M phase, each state is abundant with different cell cycle phases (Figure 3C). Of note, State5 had relatively low stemness scores, but was mostly in the S and G2M phases, representing a subpopulation with low differentiation potential, high proliferation, and increased malignancy. With the resistant cells predicted and confirmed by DrugFormer, next, we calculated the percentage of each state in resistant cells and identified that State5 had a significantly higher percentage in resistant cells (Figure 3D). These results indicated State5 as a unique resistant state in MM.

To further verify the resistant State5, we utilized the MMRF-COMMPASS patient samples and deconvoluted those samples by the five cell states. Though investigating the relationship between the abundance of each state and the clinical tumor stage, we found that clinical stage III had a significantly higher abundance of State5 (Figure 3E), indicating that State5 was associated with higher malignancy. Using the Shannon entropy, we calculated the intra-tumoral heterogeneity (ITH) of each patient based on the predicted abundance of each state. Higher ITH was shown to be associated with poorer overall survival and cancer-specific survival (Figure 3F,G). Importantly, MM patients with high enrichment of State5 tended to have higher ITH (Figure 3H), indicating that the resistant state, i.e., State5, plays a significant role in ITH. Meanwhile, a higher abundance of State5 was observed to be associated with lower overall survival (Figure 3I). These results further demonstrate that State5 is significantly related to patient outcomes and may serve as a potential drug-resistant subpopulation.

2.4. Molecular Mechanisms Underlie Resistant Cell State in Multiple Myeloma

To investigate the underlying genetic mechanisms of resistance in State5, next, we characterized the differentially expressed genes in each cell state especially State5 (**Figure 4**A). Compared with the other states, State5 highly expressed genes related to the dynamics of cell division (CDC42, YBX1, etc.), DNA repair (FEN1, RBX1, etc.), chromosomal stability (SMC4, etc.), and mitochondrial respiratory and oxidative phosphorylation (COX8A, etc.). Following GO-BP enrichment analysis, only the upregulated genes of State5 were significantly enriched in processes related to the Mitotic Cell Cycle (Figure 4B). Moreover, with the

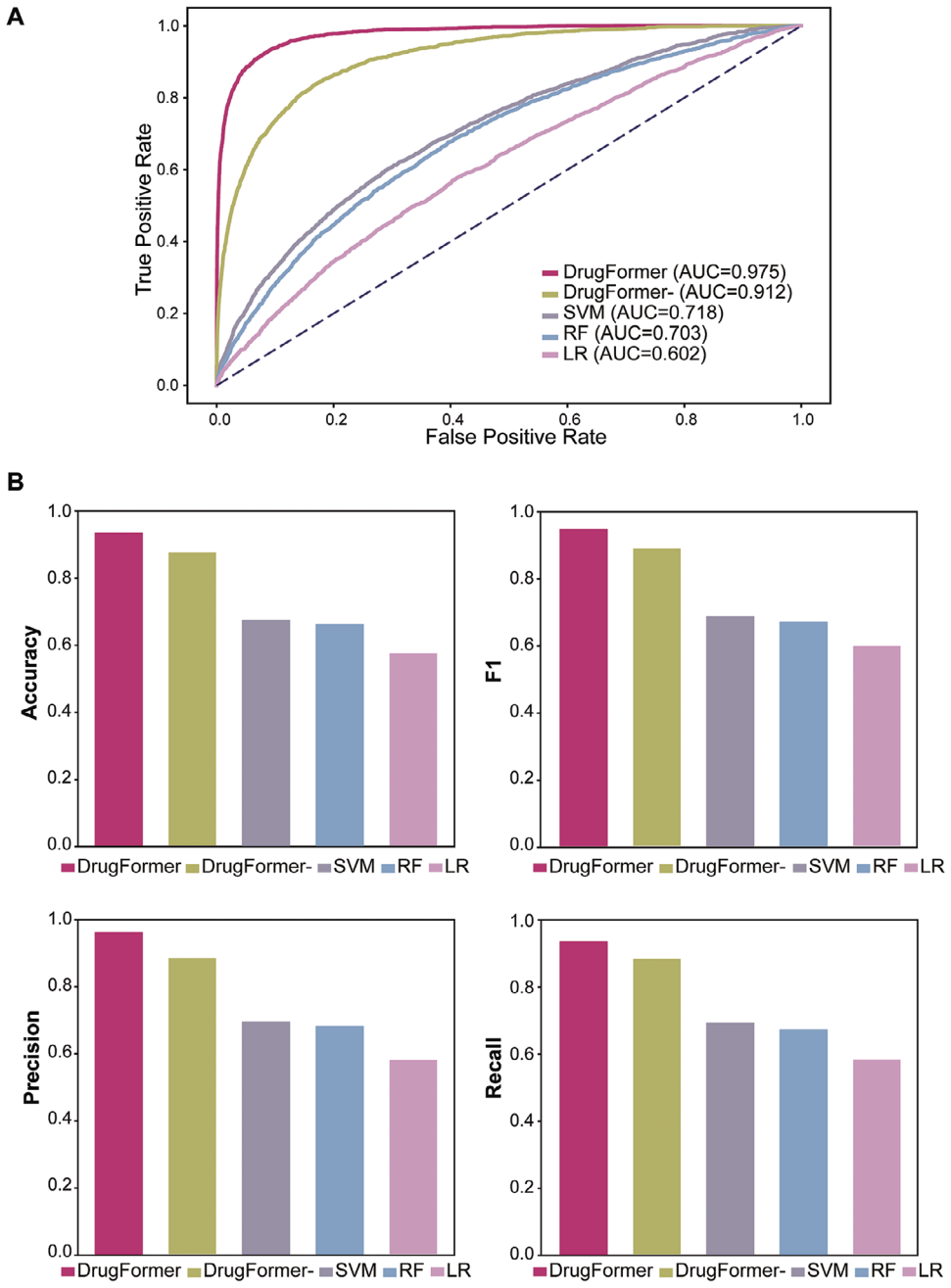

Figure 2. Performance evaluation of DrugFormer. A) AUC of DrugFormer in predicting single-cell response to drugs based on five-fold cross-validation. B) Average values of accuracy, F1, precision, and recall of DrugFormer in predicting cellular drug resistance, based on five-fold cross-validation.

copy number variations inferred by copycat,[25] State5 presented significant copy number amplifications in chromosomes 1, 11, and 19 (Figure 4C). Of note, some upregulated genes of State5, such as FEN1, RBX1, and COX8A,[26–28] showed copy number amplification. Interestingly, among the MMRF-COMMPASS patients, the expressions of FEN1, RBX1, and COX8A significantly increased along the patient treatment line (Figure 4D–F). Moreover, the expression of FEN1, RBX1, and COX8A was significantly increased in malignant-resistant cells compared with malignant-sensitive cells in the single-cell dataset (Figure S1, Supporting Information). These results suggest that these three genes were upregulated in drug-resistant patients. In addition, among the TCGA-MM patients, high expressions of either FEN1, RBX1, and COX8A were significantly associated with poor overall survival and cancer-specific survival (Figure 4G–I). Specifically, for COX8A, such association patterns were also observed in some other cancer types of TCGA patients, including ACC, LAML, LIHC, and LUAD (Figure 4J). These findings suggest that COX8A may serve as a potential new target for overcoming drug resistance for several types of cancers.

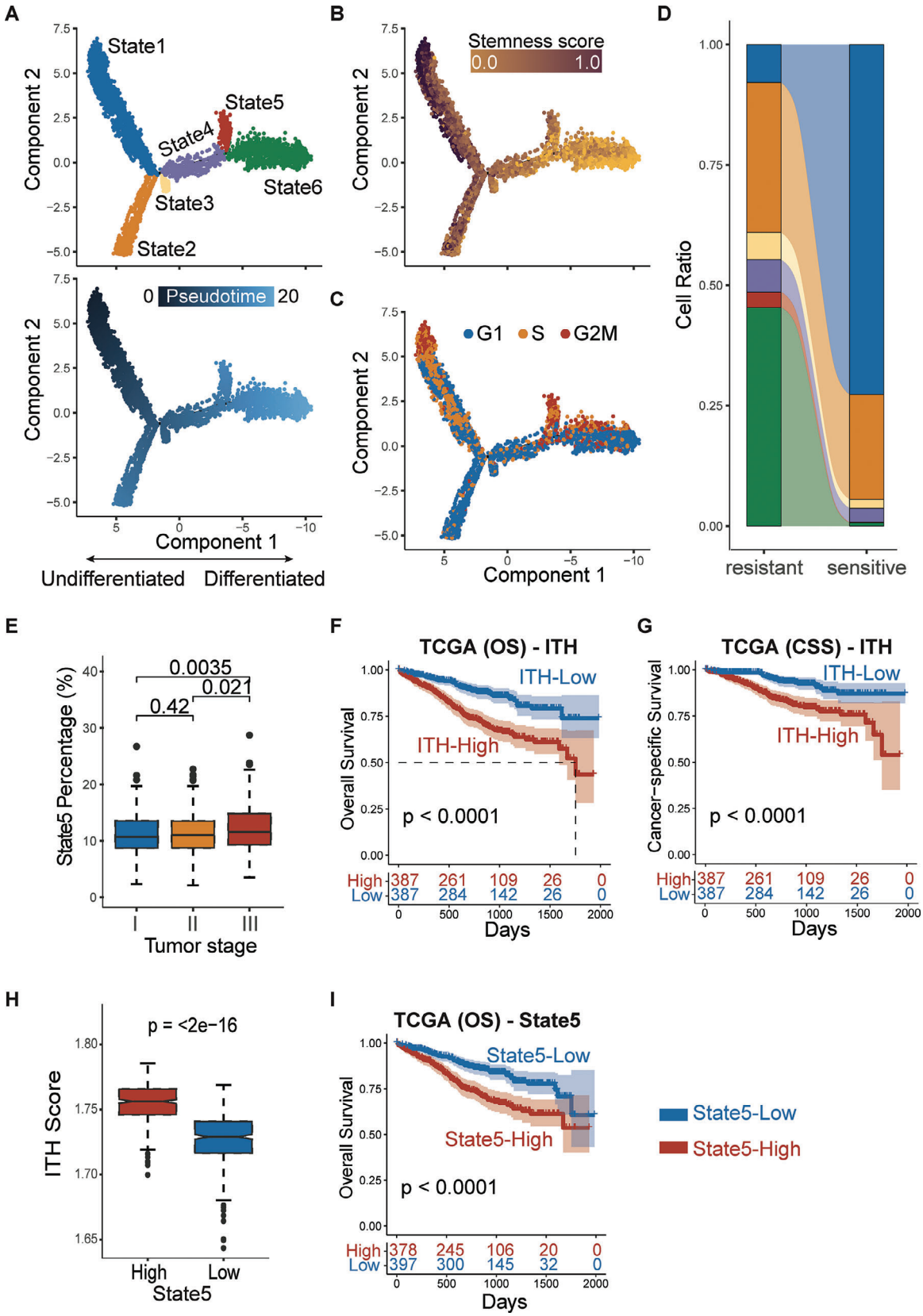

Figure 3. Identification of resistant cell state in multiple myeloma. A) The results of pseudotime analysis by Monocle2. B) The stemness score calculated by CytoTRACE. C) The cell cycle phase predicted by the "CellCycleScoring" function in Seurat. D) Flowchart illustrating the percentage of six states

Next, we further analyzed the cell-cell interactions and transcriptional regulation of malignant drug-resistant cells, drug-sensitive cells, and other cell types in the tumor microenvironment, to delve into the molecular characterization and mechanisms of drug resistance (Figure S2, Supporting Information). Through the cell-cell interaction analysis (Figure S2A,B, Supporting Information), the results showed that ligand-receptor pairs (MDK-NCL and IL16-CD4) were only present in cell-cell interactions where sensitive cells served as the source (Figure S2C, Supporting Information). The other ligand-receptor pair TNFSF13B-TNFRSF13B was only significantly present in cell-cell interactions where resistant cells served as the target (Figure S2D, Supporting Information). Previous studies have reported that IL16 induces TME cells migration, which improves cancer therapeutic efficacy.[29] An increase in TNFSF13B-TNFRSF13B could induce the proliferation of malignant cells in multiple myeloma.[30] Moreover, transcriptional regulation analysis showed that JUND transcription factors were significantly enriched and up-regulated in resistant cells (Figure S2E,F, Supporting Information). Previous studies have shown that high expression of JUND is associated with increased malignancy in certain tumors and may promote tumor cell proliferation and invasion.[31] These results shed further light on the characterization of drug resistance and the accuracy of our model.

2.5. DrugFormer Uncovers Resistant Cell Subpopulation in Different Cancers

To further demonstrate the capability of DrugFormer, we applied it to the scRNA-seq data profiled from refractory acute myeloid leukemia (AML) patients who were subsequently treated with ficlatuzumab.[32] Notably, DrugFormer identified the resistant cells with F1 score as 0.909. With the resistant cells confirmed by DrugFormer, we used similar analyses to interrogate the cellular dynamics to understand the underlying heterogeneity resulted drug resistance in AML.

First of all, we inferred the malignant cell trajectories and revealed the cell states from the AML scRNA-seq data (**Figure 5**A). Subsequently, we utilized those annotated cell states (State1–State3) as references to deconvolve those cell state proportions in TCGA-LAML patient samples. Then we analyzed the relationship between the proportion of each cell state and overall survival. The results showed that a higher proportion of State3 was associated with worse survival (Figure 5B). This specific State3 was more abundant in advanced tumors or clinical subtypes composed of mature cells (Figure 5C). Meanwhile, most of the State3 cells in the AML scRNA-seq dataset came from drug-resistant patients (Figure 5D), suggesting State3 as the unique drug-resistant state. Similarly, the upregulated differentially expressed genes (DEGs) of State3 were identified (Figure 5E). Consistent with the findings in the MM dataset, State3 of the Refractory AML also had significant copy number amplification in chromosome 1q and COX8A (Figure 5F). Interestingly, the State3 of AML shares the high enriched pathways including the "oxidative phosphorylation," "regulation of actin cytoskeleton," and "Proteasome" pathways with the State5 of the MM dataset (Figure 5G). These results suggest that the specific resistant states from different cancer types may share similar molecular characteristics, enabling the identification of a common potential drug resistance-specific subpopulation and potential targets, such as COX8A for overcoming drug resistance.

Next, we applied DrugFormer to the scRNA-seq data of solid tumors including melanoma (**Figure 6**), small cell lung cancer (Figure S3A, Supporting Information) and prostate cancer (Figure S3B, Supporting Information). The melanoma scRNA-seq dataset has the BRAFV600E mutation, which was subsequently treated with dasatinib.[33] The lung cancer dataset was from a small cell lung cancer (SCLC) CDX model, which was subsequently treated with chemotherapy.[34] The prostate cancer dataset was profiled before enzalutamide treatment.[35] Notably, DrugFormer identified the resistant cells with high F1 scores of 0.951 and 0.896, and 0.876 in these three cancers, respectively (Table S1, Supporting Information). These results demonstrate that our model can be applied to both liquid and solid tumors with strong generalization and reliability.

With the resistant cells identified by DrugFormer, we then interrogated the cellular dynamics to understand the underlying heterogeneity that resulted in drug resistance. For the melanoma scRNA-seq dataset, we then inferred the malignant cell trajectories and revealed the cell states (**Figure 6**A). Five cell states (State1–State5) were identified and used to decompose the cell state proportions in TCGA-SKCM patient samples. Then we analyzed the relationship between the proportion of each cell state and overall survival. The results showed that a higher proportion of State1 was associated with worse survival (Figure 6B). This specific State1 was more abundant in the advanced clinical stage (Figure 6C). Meanwhile, most of the State1 cells in the melanoma scRNA-seq dataset came from drug-resistant patients (Figure 6D), suggesting State1 as a unique drug-resistant state. Moreover, enrichment analysis revealed that State1 had highly enriched pathways including the "cell cycle," "regulation of actin cytoskeleton," and "Proteasome" pathways (Figure 6E), which were also observed in the State5 of MM dataset and the State3 of AML dataset (Figure 5G). For the lung cancer scRNA-seq data, through trajectory analysis, we also found a cell state (State2) in which most of the cells were from drug-resistant patients (Figure S3A, Supporting Information). For the prostate cancer scRNA-seq data, most cells in states 5, 6, and 7 of the trajectory were also abundant in drug-resistant patients (Figure S3B, Supporting Information). Those identified resistant cell states of different cancer types may share similar molecular characteristics, enabling the identification of a common potential drug

(State1–State5) between the resistant and sensitive cells. E) Boxplot of the percentage of State5 across clinical tumor stages. Data were analyzed with the Wilcoxon Test. There were 261 stage I patients, 270 stage II patients, and 224 stage III patients. The *p*-values were 0.42, 0.0035, and 0.021 for stage I and stage II, stage I and stage III, and stage II and stage III, respectively. F) Kaplan–Meier curve of overall survival based on ITH level. G) Kaplan–Meier curve of cancer-specific survival based on ITH-level. H) The boxplot of the ITH score in patient samples with high and low State5 abundance. Data were analyzed with the Wilcoxon Test. There were 379 State5-High patients and 397 State5-Low patients. The *p*-value was <2e−16. I) Kaplan–Meier curve of overall survival based on State5 abundance. For the Wilcoxon Test and survival analysis, $p < 0.05$ was considered significant.

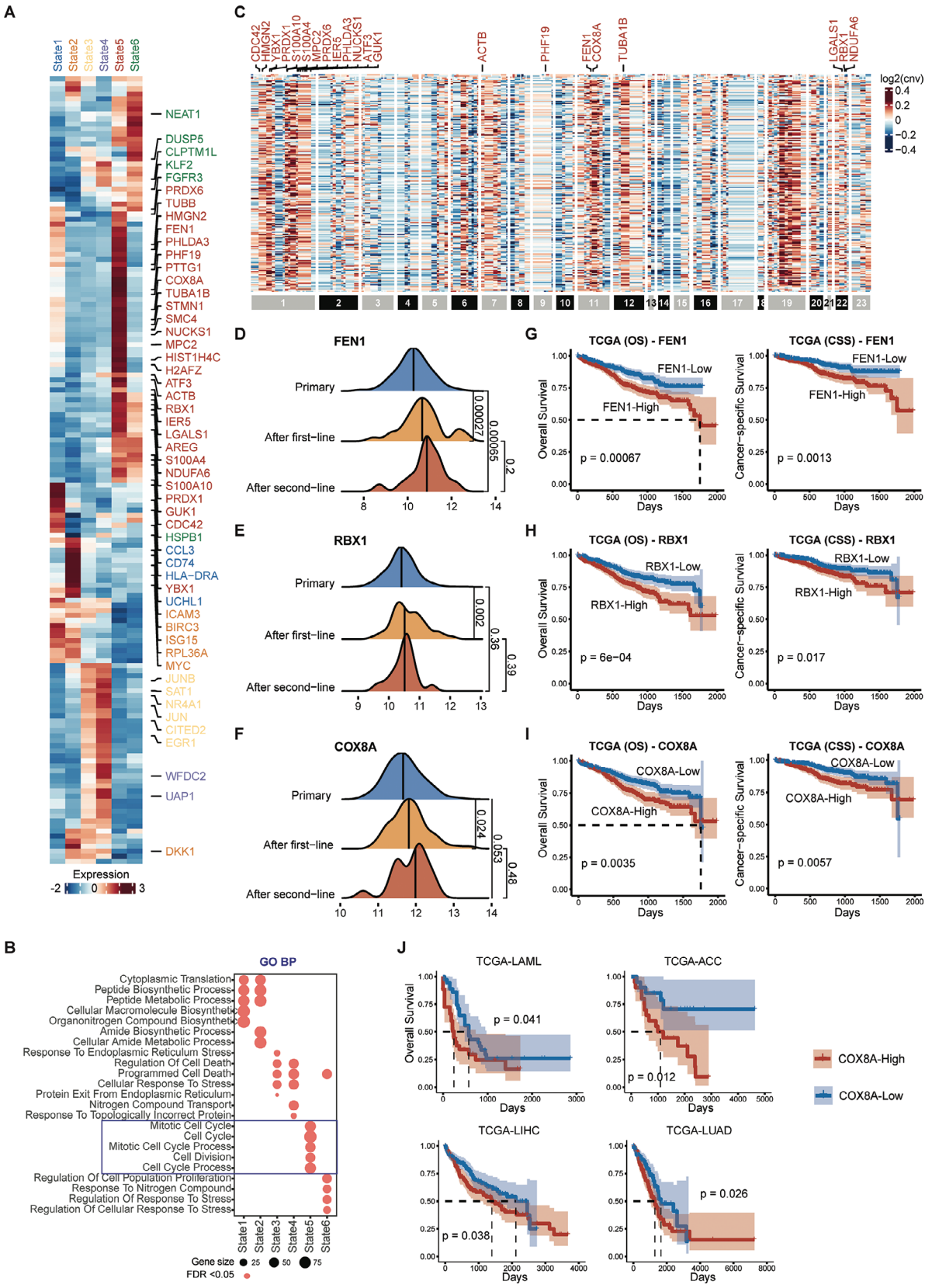

Figure 4. Underlying mechanisms of resistant cell state in MM. A) A portion of the differentially expressed genes in each state. B) GO-BP enrichment results of differentially expressed genes from each state. FDR < 0.05 was considered significant. C) Inferred copy number variation by copycat software.

resistance-specific subpopulation and potential targets for overcoming drug resistance.

3. Discussion

The intratumoral intricate heterogeneity exacerbates individual patient resistance, posing a primary hurdle to the effectiveness of targeted therapies. Therefore, it is critical to delve into how varying cell populations in different regions of tumor lesions manifest resistance or incomplete responses to treatment to enhance the overall efficacy of drug therapy. In this work, we have developed a tailored model, DrugFormer, to interrogate how cells and genes in a complex tissue respond to drugs differentially. DrugFormer harnesses advanced natural language processing capabilities to decipher extensive datasets from the DRMref database.[15] The performance of DrugFormer has been validated through comprehensive benchmarking. DrugFormer not only enhances the accuracy of predicting resistance but also sheds light on previously undiscovered targets for therapeutic intervention. Our model represents a significant advancement in the field, holding the potential to redefine how we approach drug resistance prediction and the identification of novel targets, ultimately contributing to more effective and personalized medical treatments.

In the application of DrugFormer to the scRNA-seq data of multiple myeloma, we unveiled a specific cancer cell state exhibiting stronger drug resistance while presenting elevated expressions of FEN1, RBX1, and COX8A. FEN1 and RBX1 are two genes related to DNA repair, which have been reported to be associated with drug resistance in several cancers.[26–28] COX8A, or Cytochrome C Oxidase Subunit 8A, is a nuclear-encoded subunit of cytochrome coxidase (COX), which is the terminal enzyme complex of the mitochondrial respiratory chain. It plays a crucial role in the regulation of mitochondrial oxidative phosphorylation and energy production. The discovery regarding COX8A suggests that it may serve as a potential new target for overcoming drug resistance in MM. Possible mechanisms of drug resistance may involve the increase in chromosome copy number leading to upregulation of COX8A gene expression, thereby enhancing energy support, promoting cancer cell proliferation, and contributing to drug resistance. Previous study shows that COX8A may lead to a number of inherited disorders such as Leigh-like syndrome and epilepsy. These disorders are often associated with mitochondrial dysfunction, which in turn is closely linked to the development of cancer.[36] Another recent work identifies four stemness-associated genes including COX8A in intrahepatic cholangiocarcinoma (ICC). Functional studies indicate that COX8A is associated with the self-renewal ability of ICC and transgenic expression of COX8A could enhance chemoresistance of cholangiocarcinoma cells.[37] These studies provide further evidence for COX8A as a new target to overcome drug resistance. Since our study focuses on method development, we anticipate validating this target in our future work.

Our findings in MM highlight the intra-tumoral heterogeneity related to drug tolerance. Such heterogeneity poses challenges for cancer therapy because different subpopulations of cancer cells may respond differently to drugs, leading to treatment tolerance and recurrence. Therefore, it is crucial to use DrugFormer to identify the drug-resistant cells, thus provide interpretable insights to study drug resistance mechanisms. Given the advantages of DrugFormer, we foresee several aspects for improvement. While the current DrugFormer is specifically designed for single-cell RNA-seq data, we envision the development of a spatial omics-based large language model utilizing emerging spatial datasets. Such an upgrade will unveil the spatial cellular response to drugs, enabling precise patient treatment by identifying drug-resistant tumor cells within tumor lesions and elucidating the cell-cell communications underlying such resistance. Furthermore, despite the promising performance of the DrugFormer model, its interpretability may pose a challenge. To address this, the employment of explainable AI approaches[38] will help understand the contributions of cells or genes to drug resistance. Additionally, DrugFormer faces the limitations in terms of generalization regarding data collection insufficiency, which may impact the applicability of DrugFormer. If data collection is inadequate, the dataset may not represent the broader population or various conditions. To address this limitation, we are working on collecting more drug treatment-related single-cell data and expanding the collection to include diverse populations and sufficient sample sizes, thereby enhancing the generalizability and reliability of DrugFormer.

4. Conclusion

DrugFormer represents a significant advancement in addressing the critical challenge of drug resistance in healthcare. By leveraging a novel graph-augmented language model, DrugFormer integrates serialized gene tokens and a gene-based knowledge graph to predict drug resistance at the single-cell level with high accuracy. Comprehensive single-cell data analysis from different cancer types highlights the efficacy of DrugFormer in identifying resistant cells and uncovering underlying molecular mechanisms. DrugFormer not only enhances the precision of predicting drug response but also reveals unique drug-resistant cellular states and potential therapeutic targets such as COX8A. This powerful tool offers valuable insights into the heterogeneity of cellular responses to drugs, ultimately guiding personalized

D) Expression of FEN1 along the treatment line. Data were analyzed with the Wilcoxon Test. There were 776 Primary patients, 64 After first-line patients, and 19 After second-line patients. The *p*-values of FEN1 gene were 0.00027, 0.00065, and 0.2 for Primary and After first-line, Primary and After second-line, and After first-line and After second-line, respectively. E) Expression of RBX1 along the treatment line. Data were analyzed with the Wilcoxon Test. There were 776 Primary patients, 64 After first-line patients, and 19 After second-line patients. The *p*-values of RBX1 gene were 0.002, 0.36, and 0.39, receptively. F) Expression of COX8A along the treatment line. Data were analyzed with the Wilcoxon Test. $p < 0.05$ was considered significant. There were 776 Primary patients, 64 After first-line patients, and 19 After second-line patients. The *p*-values of COX8A gene were 0.024, 0.053, and 0.48, respectively. G) Kaplan–Meier curve of overall survival and cancer-specific survival for gene FEN1. H) Kaplan–Meier curve of overall survival and cancer-specific survival for gene RBX1. I) Kaplan–Meier curve of overall survival and cancer-specific survival for gene COX8A. J) Kaplan–Meier curve of overall survival based on COX8A expression level in the TCGA-LAML, TCGA-ACC, TCGA-LIHC, and TCGA-LUAD datasets. For the Wilcoxon Test and survival analysis, $p < 0.05$ was considered significant.

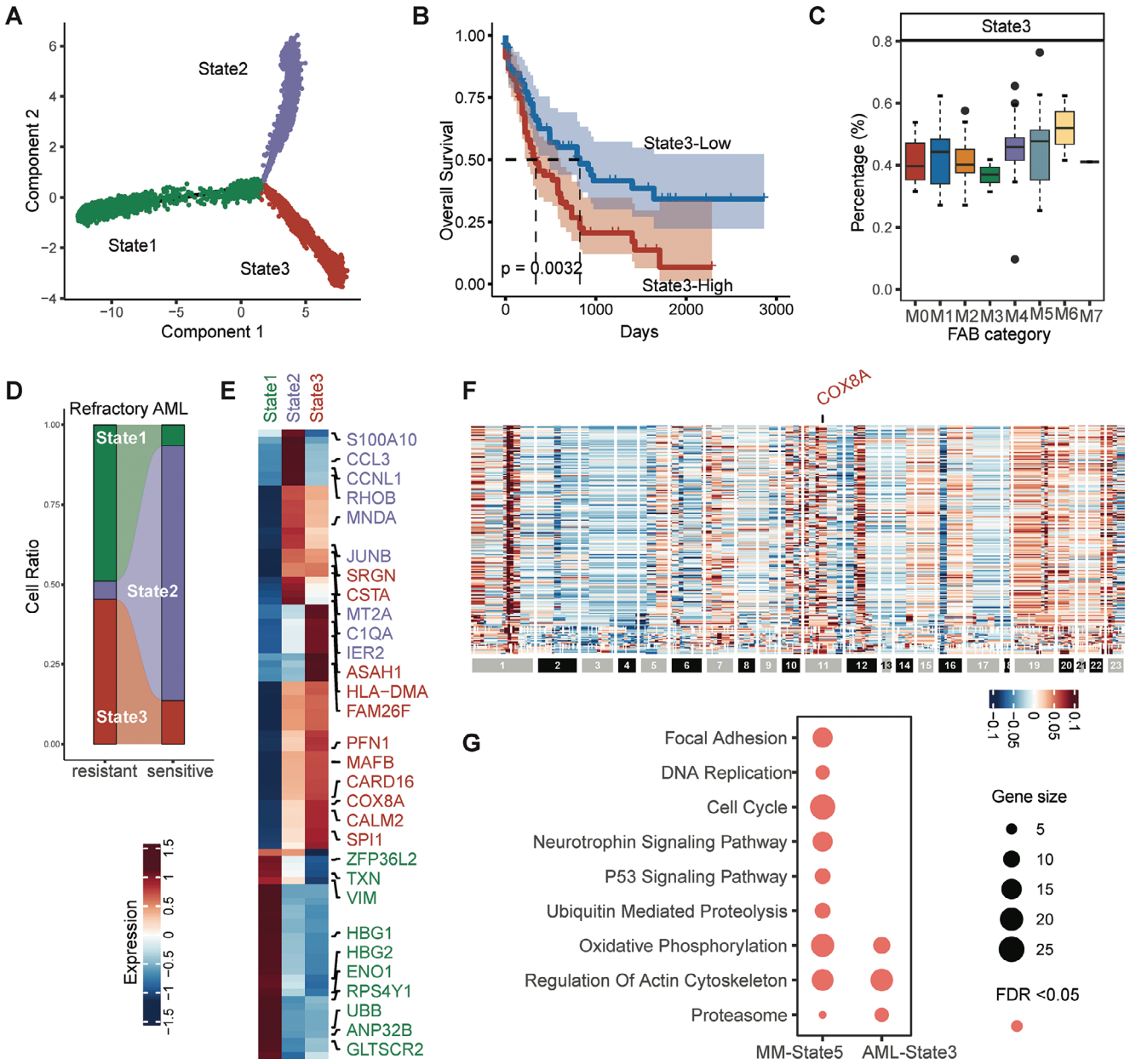

Figure 5. Resistant cell state of acute myeloid leukemia. A) Pseudotime analysis results using Monocle2 for the AML dataset. B) Kaplan–Meier curve of overall survival based on State3 abundance level in the TCGA-LAML dataset. $p < 0.05$ was considered significant. C) Boxplot of State3 proportions across clinical tumor subtypes in the AML dataset. There were 15 M0 patients, 35 M1 patients, 38 M2 patients, 14 M3 patients, 29 M4 patients, 15 M5 patients, 2 M6 patients, and 1 M7 patients. D) Flowchart illustrating the percentage of three states between the resistant and sensitive cells in the AML dataset. E) Subset of differentially expressed genes in each state of the refractory AML dataset. F) Inferred copy number variation of State3 in the AML dataset. G) KEGG enrichment results of differentially expressed genes from MM-State5 and AML-State3. FDR <0.05 was considered significant.

treatment strategies and paving the way for more effective cancer therapies.

5. Experimental Section

Here, a novel graph-enhanced language model termed DrugFormer was introduced, which harnessed the Transformer[39] and Graph Attention Network[40] architectures to predict cell-level drug response. Figure 1 illustrates the architecture of DrugFormer, which applied the Encoder-only Transformer Block to incorporate both gene sequence information but also graph information.

Encoder-Only Transformer Block: As provided in *Statistical Analysis*, for each gene symbol g_i in the scRNA-seq data,[41] the high-dimensional embedding of the token sequence was $t_i \in \mathbb{R}^{256}$ of gene g_i. For each cell in the scRNA-seq data, the top-expressed genes ($N_1 = 2048$) were selected, thus the input of Transformer encoder was $T = \{t_i\}_{i=1}^{N_1}$, $t_i \in \mathbb{R}^{256}$. The knowledge graph was denoted as $G = \{g_i\}_{i=1}^{N_2}$, $g_i \in \mathbb{R}^{k \times k}$, with adjacency matrix $A \in \mathbb{R}^{N_2 \times N_2}$. k represented the dimension of eigenvalue. N_2 represented the number of genes in the scRNA-seq data.

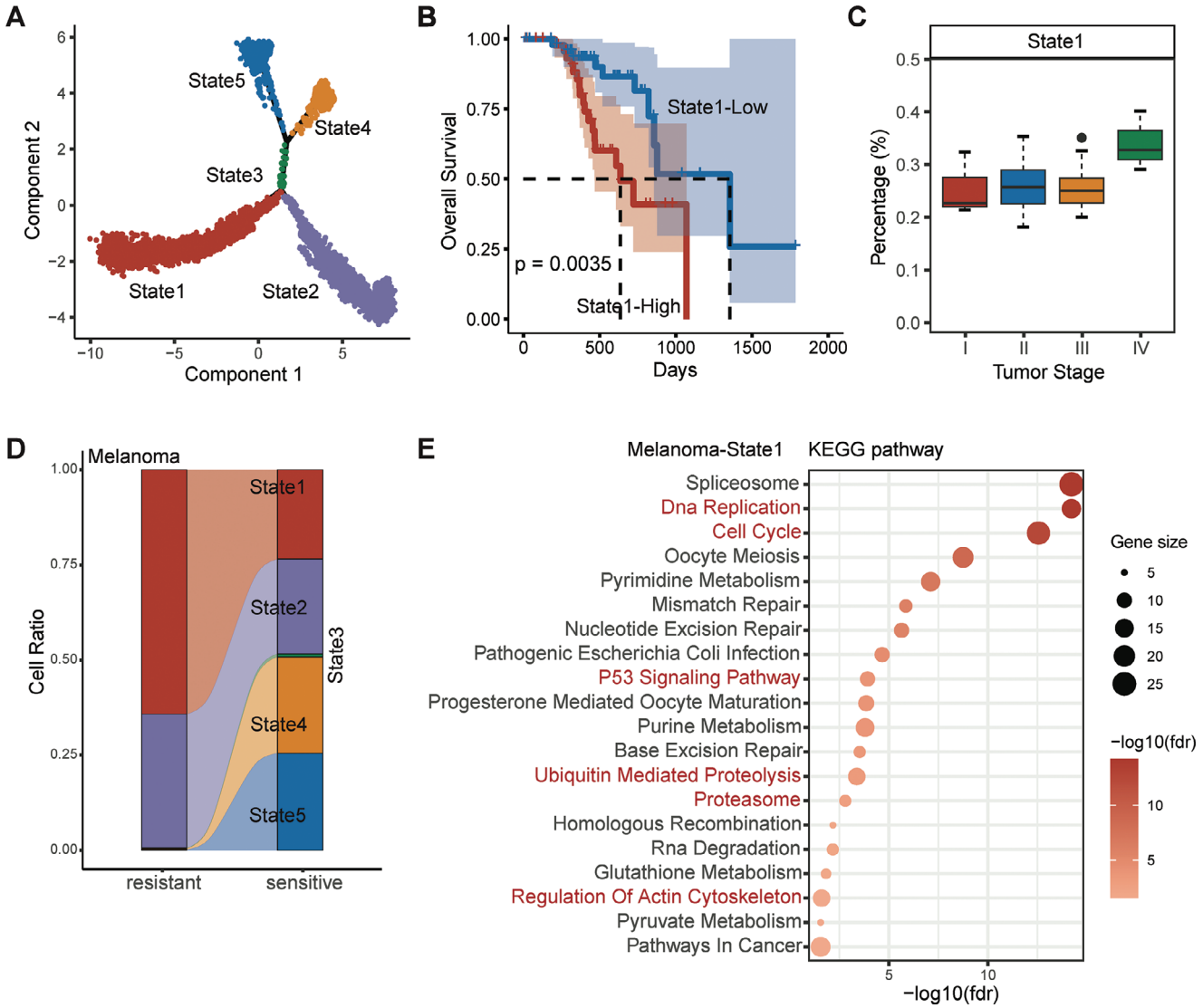

Figure 6. Resistant cell state of melanoma. A) Pseudotime analysis results using Monocle2 for the melanoma dataset. B) Kaplan–Meier curve of overall survival based on State1 abundance level in the TCGA-SKCM dataset. $p < 0.05$ was considered significant. C) Boxplot of State1 proportions across clinical tumor subtypes in the melanoma dataset. There were 3 stage I patients, 66 stage II patients, 27 stage III patients, and 3 stage IV patients. D) Flowchart illustrating the percentage of three states between the resistant and sensitive cells in the melanoma dataset. E) KEGG enrichment results of differentially expressed genes from melanoma-State1. FDR <0.05 was considered significant.

DrugFormer used the encoder-only transformer structure as the backbone. Six-layer encoder was used and each encoder block consists of an attention layer and a feedforward neural network layer, along with residual connections and normalization operations.[42] The attention layer had eight attention heads, with an embedding dimension as 256.

Given the multiple attention heads ($h = 8$), the input to each attention head was annotated as $T_h \in \mathbb{R}^{N\times 32}$, where $T = \{T_h\}_{h=1}^{8} = \{t_i\}_{i=1}^{N_1}$. The ouput of h-th attention head Z_h was:

$$Z_h = Attention_h = softmax\left(\frac{Q_h W_h^T}{C}\right) V_h \quad (1)$$

where $Q_h = Linear_q(T_h)$ was the query matrix, $K_h = Linear_k(T_h)$ was the key matrix, and $V_h = Linear_v(T_h)$ was the value matrix. C was a constant used for scaling. The output of each attention head (Z_h) concatenated as $Z = [Z_1, ., Z_8] \in \mathbb{R}^{N_1\times 256}$ and normalized by layer normalization, i.e.,

$$Z' = LayerNorm(T + Z) \quad (2)$$

where $LayerNorm(\alpha) = (\alpha - \mu)/\sqrt{\sigma^2 + \epsilon}$, $\mu = \frac{1}{n}\sum_{i=1}^{n}\alpha$, and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha - \mu)^2$.

Next, a feed-forward layer for nonlinear transformation and a normalization layer were used, with output as $T^{(1)} = \{t_i^{(1)}\}_{i=1}^{N_1}$, $t_i^{(1)} \in \mathbb{R}^{256}$, i.e.,

$$T^{(1)} = LayerNorm\left(max\left(0, Z'W_1 + b_1\right)W_2 + b_2\right) \quad (3)$$

Here $T^{(1)}$ was the output of the transformer encoder block. The entire model comprised six transformer encoders, each encoder block was

denoted as $Enc(\cdot)$. The output of *l-th* transformer encoder was $T^{(l)}$, $l \in [1, 6]$. Specifically, for the first transformer encoder, $T^{(1)} = Enc(T)$.

Graph Attention Block: To leverage the knowledge graph, the graph attention block with one-layer graph attention network was used. For two neighboring gene i and gene j in the graph G, this network learned the attention weight $e_{i,j}$ as:

$$e_{i,j} = LeakyReLU\left(a^{\top}\left[w_i^1 g_i \parallel w_j^1 g_j\right]\right) \tag{4}$$

where $a^{\top} \in \mathbb{R}^{1\times 512}$ was a learnable vector, $w_i^1 \in \mathbb{R}^{256\times k}$ and $w_j^1 \in \mathbb{R}^{256\times k}$ were learnable weight matrices. $g_i \in \mathbb{R}^{k\times 1}$ and $g_j \in \mathbb{R}^{k\times 1}$ were features of node i and node j. $\parallel$ represented the concatenation operation, $LeakyReLU(\cdot)$ was the activation function, i.e.,

$$LeakyReLU\ (\alpha) = \begin{cases} \alpha, & if\ \alpha > 0 \\ \gamma \cdot \alpha, & if\ \alpha \le 0 \end{cases} \tag{5}$$

where α was the input, $\gamma \in \mathbb{R}$ was a small constant used to control the negative slope. When $\alpha > 0$, $LeakyReLU(\cdot)$ was the same as $ReLU(\cdot)$; when $\alpha \le 0$, $LeakyReLU(\cdot)$ used a small negative slope γ to multiply input α.

The attention coefficient $\alpha_{i,j}$ was then obtained by normalizing the attention weight $e_{i,j}$ through the $softmax(\cdot)$ function, i.e.,

$$\alpha_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}} \exp\left(e_{i,k}\right)} \tag{6}$$

where $\mathcal{N}(i)$ represented the neighbor nodes of node i.

Herein, the aggregated representation g_i' considering weighted neighbor nodes, i.e., $g_i' = ReLU(\alpha_{i,i}\, w_i^2 g_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j} w_j^2 g_j)$, where $w_i^2 \in \mathbb{R}^{256\times k}$ and $w_j^2 \in \mathbb{R}^{256\times k}$ were learnable parameters. Therefore, the node embeddings of the entire graph w $G' = \{g_i'\}_{i=1}^{N_2}\, g_i' \in \mathbb{R}^{256}$. The Graph Attention Block was denoted as $GAT(\cdot)$, i.e., $G' = GAT(G)$. This node embedding was then aligned with the same list of genes within input T, thus we have $\tilde{G} = \{\tilde{g}_i\}_{i=1}^{N_1}$, where $\tilde{g}_i \in \mathbb{R}^{256}$.

Gated Aggregator: Following the graph attention block, the embeddings of knowledge graph ($\tilde{G}$) was fused with gene token embeddings ($T^{(1)}$) through the gated aggregator. The gated aggregator enabled the fusion of knowledge graph and gene token using a gating mechanism. In this module, the first layer was designated as a concatenated operation. For each gene i, the concatenated operation was shown as:

$$Z_c^{(1)} = Concat\left(t_1^{(1)} \parallel \tilde{g}_1, \ldots, t_{N_1}^{(1)} \parallel \tilde{g}_{N_1}\right) \tag{7}$$

where $t_i^{(1)} \in \mathbb{R}^{256}$ and $\tilde{g}_1 \in \mathbb{R}^{256}$ were denoted above. "||" referred to concatenating function. $Z_c^{(1)} \in \mathbb{R}^{N_1\times 512}$ was the concatenated output.

The concatenated features were then projected into a unified embedding space through linear mapping layer and the $ReLU(\cdot)$ function, to obtain a complementary embedding representation $Z_f^{(1)} \in \mathbb{R}^{N_1\times 256}$, i.e.,

$$Z_f^{(1)} = ReLU\left(Z_c^{(1)} W_3 + b_3\right) \tag{8}$$

where $W_3 \in \mathbb{R}^{512\times 256}$ and b_3 were the parameters of the linear mapping layer. To this end, the fused embeddings were obtained from the Gated Aggratator $GA(\cdot)$, i.e.,

$$Z_f^{(1)} = GA\left(T^{(1)}, \tilde{G}\right) \tag{9}$$

The fused embeddings $Z_f^{(1)}$ were processed by the second transformer encoder, i.e., $T^{(2)} = Enc(Z_f^{(1)})$. The output $T^{(2)}$ then went through three transformer encoders to obtain a deep representation $T^{(5)} = \{t_i^{(5)}\}_{i=1}^{N_1}\, t_i^{(5)} \in \mathbb{R}^{256}$, i.e., $T^{(l)} = Enc(T^{(l-1)})$, $l \in [3, 5]$. To further enhance the graph-represented information, the deep representation $T^{(5)}$ was fused with knowledge graph by the Gated Aggregator $GA(\cdot)$, i.e., $T^{(6)} = GA(T^{(5)}, GAT(\tilde{G}))$.

Output Layer: In the output layer, $T^{(6)}$ was first transformed through the shape function that flattens the elements of $T^{(6)}$ to obtain a 1D dense representation. Then a bottleneck layer was used to compress the flattened information through linear transformation and obtain a sparse representation $Z_s \in \mathbb{R}^{1\times 32}$,

$$Z_s = Bottleneck\left(Shape\left(T^{(6)}\right) W_4 + b_4\right) \tag{10}$$

Z_s further went through the classification layer to get the prediction result P_T of the model. The classification layer was shown as follows:

$$P_T = sigmoid\left(Z_s W_5 + b_5\right) \tag{11}$$

herein, the probability value P_T was obtained within [0,1], which referred to the drug response probability of cell.

Loss Function: To training the model, the batch size was set to 64 and trained on a single Nvidia A100X-40C GPU with 40GB memory. The training used five-folds cross-validation, and the number of training epochs was 5. The training objective was to minimize the cross-entropy loss L:

$$L = -\left(y \log\left(P_T\right) + (1-y)\log\left(1-P_T\right)\right) \tag{12}$$

where y was the ground truth label represented by a one-hot vector. Specifically,[10] represented one class (drug-resistant cell), and [0, 1] represented the other class (drug-sensitive cell). The first element of P_T referred to the predicted probability of cell resistance, and the second element was the predicted probability of cell sensitivity.

Statistical Analysis—Pre-Processing of Data: Each gene symbol g_i in the scRNA-seq data was first tokenzied[41] as token sequence. Each token sequence was performed with high-dimensional embedding (both semantic and position embedding). As genes with copy number changes often associate with drug response,[43–45] the probabilities of gene haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) were collected from Ryan et.al.[46] The input gene-level knowledge graph was constructed based on the similarity of these probability scores. If two gene nodes have similar probability scores, these two genes were connected in the knowledge graph. The node attributes were the gene eigenvalue from Ryan et.al,[46] complementing the graph information. For downstream analysis, the downloaded scRNA-seq dataset was preprocessed by the "SCTransform" function of the Seurat package, including filtering, removal of batch effects and normalization.

Statistical Analysis—Sample Size for Each Statistical Analysis: For the Wilcoxon-Test of MM-State5 percentage between tumor stages, there were 261 stage I patients, 270 stage II patients, and 224 stage III patients. For the Wilcoxon-Test of the ITH score between High- and Low- MM-State5 percentage, there were 379 State5-High patients and 397 State5-Low patients. For the Wilcoxon-Test of the expression of genes (FEN1, RBX1, and COX8A) between different treatment lines of MM, there were 776 Primary patients, 64 After first-line patients, and 19 After second-line patients. For the comparison of AML-State3 percentage between AML FAB subtypes, there were 15 M0 patients, 35 M1 patients, 38 M2 patients, 14 M3 patients, 29 M4 patients, 15 M5 patients, 2 M6 patients, and 1 M7 patients. For the comparison of melanoma-State1 percentage between tumor stages, there were 3 stage I patients, 66 stage II patients, 27 stage III patients, and 3 stage IV patients. In the Wilcoxon-Test test for FEN1, RBX1, and COX8A gene expression between malignancy-resistant and malignancy-sensitive cells, there were 354 malignant-resistant cells and 5134 malignant-sensitive cells.

Statistical Analysis—Data Presentation: To evaluate the performance of DrugFormer, it was compared with the "DrugFormer-" model and three classic machine learning (ML) models, i.e., Random Forest (RF), Support Vector Machine (SVM), Linear Regression (LR). Here the "DrugFormer-"

model referred to the ablated DrugFormer framework without leveraging gene-based knowledge graph, i.e., only six transformer encoder blocks. Quantitative metrics including accuracy, F1 score, AUC, precision, and recall were used to evaluate the benchmarking performance. In Figure 2B, the results of the comparison methods were shown by the average values of the evaluation metrics.

Statistical Analysis—Statistical Methods: For the *evaluation of proliferation and stemness* for each single-cell, the signature score of a set of cell cycle-related markers was computed to assess the cell cycle state of each individual cell. This cell cycle-related marker set included 43 genes associated with the S phase and 54 genes associated with the G2M phase. Additionally, the stemness score for each single-cell was computed to evaluate stemness based on transcriptional diversity. For the *Characterization of the differentiation trajectory of malignant cells*, the cellular states were identified based on the pseudotime analysis. For the *Identification of specifically expressed genes and enriched pathways in each cellular state*, genes specifically expressed in each cellular state were identified. Genes with log2 (fold-change) greater than 0.25 and expressed at least 25% of cells were identified as specific genes of each cellular state. Furthermore, enrichment analysis of KEGG and GO BP pathways was conducted. For the *Deconvolution of bulk transcriptomic data*, the cellular states abundances were estimated of bulk transcriptomic data from MMRF-COMMPASS, TCGA-LUAD, and TCGA-LAML datasets. Single-cells were used as a reference and labeled as different cellular states. Based on the estimated percentage of different cellular states, the relative abundances of each cellular state were compared in patients with various FAB subtypes and tumor stages. Additionally, Shannon entropy was used to calculate the intra-tumoral heterogeneity (ITH) of each patient from bulk RNA datasets. For the *Copy number variation analysis*, the copy number variations[25] were inferred to reveal underlying genetics mechanisms. This process was conducted for each patient to avoid batch effect. For the evaluation of drug-resistant genes, overexpressed genes in resistant cells (log2FC >0.25 & p.adjust <0.05) with copy number variation >0.02 were used. For *Survival Analysis*, Kaplan–Meier survival analysis was conducted, and the log-rank test was used to evaluate the survival differences between groups. For *cell–cell interaction analysis*, significant ligand–receptor interaction pairs were selected with a significant value of $p < 0.05$. The number of interactions between different cell types and the intercellular communication weights were represented by circle plots. Additionally, dot plots were used to show the significant ligand-receptor pairs in malignant-resistant cells and malignant-sensitive cells, serving as source cells and target cells respectively, communicate with other cell types.

Statistical Analysis—Software for Statistical Analysis: All statistical analyses were conducted in R (version 4.2.3). For the *evaluation of proliferation and stemness* for each single-cell, the "CellCycleScoring" function within the Seurat package was utilized.[47] Additionally, the stemness score was computed for each single-cell using CytoTRACE,[48] a well-established computational framework for evaluating stemness based on transcriptional diversity. For the *Characterization of the differentiation trajectory of malignant cells*, the R package Monocle2 (version 2.28.0)was used to characterize the differentiation trajectory.[49] For the *Identification of specifically expressed genes and enriched pathways in each cellular state*, the "FindMarkers" function was used in the Seurat package. Furthermore, enrichment analysis of KEGG and GO BP pathways was conducted using hypeR (version 1.14.0) package.[50] For *Deconvolution of bulk transcriptomic data*, the CIBERSORTx[51] tool was used. Due to the input size limitation of CIBERSORTx, single-cells were randomly selected as input. For each cellular state, half were randomly selected if the number of cells was less than 1500, and one-third for the others. For the *Copy number variation analysis*, copykat package in R was used to infer copy number variations.[25] For *Survival Analysis*, the "ggsurvplot" function in the R package survminer was used. For *cell–cell interaction analysis*, CellChat[52] (version 1.5.0) was used. For *transcriptional regulation analysis*, the pySCENIC package was used to identify key transcription factors (TFs) in the different cell types according to a "three-step" TF–target regulatory network construction[53] (version 0.12.1). All parameters were set as default values. Dot plots and heatmap were used to show enriched TFs with significantly upregulated expression in each cell type.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Q.S. was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089). X.Z is supported by the National Institutes of Health (R01GM123037, U01AR069395, and R01CA241930) and the National Science Foundation (2217515 and 2326879).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

X.L. and Q.W. are co-first authors and contributed equally to this work. X.L. performed data curation, formal analysis, and visualization. Q.W. performed methodology, validation, and software. X.L., Q.W., X.Z, and Q.S. wrote, reviewed, and edited the original draft. M.Z. and Y.W. reviewed and edited the original draft. X.Z. and Q.S. performed funding acquisition. X.Z. and Q.S. performed conceptualization, funding acquisition, project administration, resources, and supervision.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Keywords

drug resistance, knowledge graph, language model, single-cell RNA sequencing

Received: May 28, 2024
Revised: July 19, 2024
Published online: August 29, 2024

[1] R. A. Ward, S. Fawell, N. Floc'h, V. Flemington, D. McKerrecher, P. D. Smith, *Chem. Rev.* **2021**, *121*, 3297.
[2] M. Schwaederle, M. Zhao, J. Jack Lee, A. M. Eggermont, R. L. Schilsky, J. Mendelsohn, V. Lazar, R. Kurzrock, *J. Clin. Oncol.* **2015**, *33*, 3817.
[3] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, M. A. Surani, *Nat. Methods* **2009**, *6*, 377.
[4] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, Y. Luo, *Clin. Transl. Med.* **2022**, *12*, e694.
[5] D. T. Miyamoto, Y. Zheng, B. S. Wittner, R. J. Lee, H. Zhu, K. T. Broderick, R. Desai, D. B. Fox, B. W. Brannigan, J. Trautwein, K. S. Arora, N. Desai, D. M. Dahl, L. V. Sequist, M. R. Smith, R. Kapur, C.-L. Wu, T. Shioda, S. Ramaswamy, D. T. Ting, M. Toner, S. Maheswaran, D. A. Haber, *Science* **2015**, *349*, 1351.
[6] W. Li, B. Zhang, W. Cao, W. Zhang, T. Li, L. Liu, L. Xu, F. Gao, Y. Wang, F. Wang, H. Xing, Z. Jiang, J. Shi, Z. Bian, Y. Song, *Exp. Hematol. Oncol.* **2023**, *12*, 44.

[7] Y. C. Cohen, et al., *Nat. Med.* **2021**, *27*, 491.
[8] K. Hinohara, H.-J. Wu, Sébastien Vigneau, T. O. McDonald, K. J. Igarashi, K. N. Yamamoto, T. Madsen, A. Fassl, S. B. Egri, M. Papanastasiou, L. Ding, G. Peluffo, O. Cohen, S. C. Kales, M. Lal-Nag, G. Rai, D. J. Maloney, A. Jadhav, A. Simeonov, N. Wagle, M. Brown, A. Meissner, P. Sicinski, J. D. Jaffe, R. Jeselsohn, A. A. Gimelbrant, F. Michor, K. Polyak, *Cancer Cell* **2019**, *35*, 330.
[9] X. Ding, et al., *Sci. Rep.* **2022**, *12*, 12501.
[10] M. K. Samur, M. Fulciniti, A. Aktas Samur, A. H. Bazarbachi, Y.-T. Tai, R. Prabhala, A. Alonso, A. S. Sperling, T. Campbell, F. Petrocca, K. Hege, S. Kaiser, H. A. Loiseau, K. C. Anderson, N. C. Munshi, *Nat. Commun.* **2021**, *12*, 868.
[11] K. G. Paulson, V. Voillet, M. S. McAfee, D. S. Hunter, F. D. Wagener, M. Perdicchio, W. J. Valente, S. J. Koelle, C. D. Church, N. Vandeven, H. Thomas, A. G. Colunga, J. G. Iyer, C. Yee, R. Kulikauskas, D. M. Koelle, R. H. Pierce, J. H. Bielas, P. D. Greenberg, S. Bhatia, R. Gottardo, P. Nghiem, A. G. Chapuis, *Nat. Commun.* **2018**, *9*, 3868.
[12] A. Sharma, E. Y. Cao, V. Kumar, X. Zhang, H. S. Leong, A. M. L. Wong, N. Ramakrishnan, M. Hakimullah, H. M. V. Teo, F. T. Chong, S. Chia, M. T. Thangavelu, X. L. Kwang, R. Gupta, J. R. Clark, G. Periyasamy, N. G. Iyer, R. DasGupta, *Nat. Commun.* **2018**, *9*, 4931.
[13] H. Heo, J.-H. Kim, H. J. Lim, J.-H. Kim, M. Kim, J. Koh, J.-Y. Im, B.-K. Kim, M. Won, J.-H. Park, Y.-J. Shin, M. R. Yun, B. C. Cho, Y. S. Kim, S.-Y. Kim, M. Kim, *Exp. Mol. Med.* **2022**, *54*, 1236.
[14] K. Li, Y. Du, Y. Cai, W. Liu, Y. Lv, B. Huang, L. Zhang, Z. Wang, P. Liu, Q. Sun, N. Li, M. Zhu, B. Bosco, L. Li, W. Wu, L. Wu, J. Li, Q. Wang, M. Hong, S. Qian, *Leukemia* **2023**, *37*, 308.
[15] X. Liu, J. Yi, T. Li, J. Wen, K. Huang, J. Liu, G. Wang, P. Kim, Q. Song, X. Zhou, *Nucleic Acids Res.* **2023**, *52*, D1253.
[16] G. Gambardella, G. Viscido, B. Tumaini, A. Isacchi, R. Bosotti, D. di Bernardo, *Nat. Commun.* **2022**, *13*, 1714.
[17] J. Chen, X. Wang, A. Ma, Q.-E. Wang, B. Liu, L. Li, D. Xu, Q. Ma, *Nat. Commun.* **2022**, *13*, 6494.
[18] Z. Zheng, J. Chen, X. Chen, L. Huang, W. Xie, Q. Lin, X. Li, K.-C. Wong, *Adv. Sci.* **2023**, *10*, 2204113.
[19] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, M. J. Garnett, *Nucleic Acids Res.* **2012**, *41*, D955.
[20] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, et al., *Cell* **2016**, *166*, 740.
[21] J. Devlin, M.-W. Chang, K. Lee, K. B. Toutanova, *arXiv* **2018**, arXiv:1810.04805.
[22] S. Kublik, S. Saboo, *GPT-3*, O'Reilly Media, Incorporated, Sebastopol, CA, USA, **2022**.
[23] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, *Bioinformatics* **2022**, *38*, 2102.
[24] S. M. Tirier, J.-P. Mallm, S. Steiger, A. M. Poos, M. H. S. Awwad, N. Giesen, N. Casiraghi, H. Susak, K. Bauer, A. Baumann, L. John, A. Seckinger, D. Hose, C. Müller-Tidow, H. Goldschmidt, O. Stegle, M. Hundemer, N. Weinhold, M. S. Raab, K. Rippe, *Nat. Commun.* **2021**, *12*, 6960.
[25] R. Gao, S. Bai, Y. C. Henderson, Y. Lin, A. Schalck, Y. Yan, T. Kumar, M. Hu, E. Sei, A. Davis, F. Wang, S. F. Shaitelman, J. R. Wang, K. Chen, S. Moulder, S. Y. Lai, N. E. Navin, *Nat. Biotechnol.* **2021**, *39*, 599.
[26] F. Yang, Z. Hu, Z. Guo, *Biomolecules* **2022**, *12*, 1007.
[27] G. Zhong, et al., *Int. J. Gen. Med.* **2021**, *14*, 6477.
[28] E. Bao, et al., *Cancer Biol. Therapy* **2023**, *24*, 2231670.
[29] C. McFadden, R. Morgan, S. Rahangdale, D. Green, H. Yamasaki, D. Center, W. Cruikshank, *J. Immunol.* **2007**, *179*, 6439.
[30] R. Chen, X. Wang, Z. Dai, Z. Wang, W. Wu, Z. Hu, X. Zhang, Z. Liu, H. Zhang, Q. Cheng, *Front. Immunol.* **2021**, *12*, 713757.
[31] J. Zhou, et al., *Onco. Targets Ther.* **2023**, *16*, 347.
[32] V. E. Wang, B. W. Blaser, R. K. Patel, G. K. Behbehani, A. A. Rao, B. Durbin-Johnson, T. Jiang, A. C. Logan, M. Settles, G. N. Mannis, R. Olin, L. E. Damon, T. G. Martin, P. H. Sayre, K. M. Gaensler, E. McMahon, M. Flanders, V. Weinberg, C. J. Ye, D. P. Carbone, P. N. Munster, G. K. Fragiadakis, F. McCormick, C. Andreadis, *Blood Cancer Discovery* **2021**, *2*, 434.
[33] M. Schmidt, et al., *Cancer Biol. Med.* **2021**, *19*, 56.
[34] C. A. Stewart, C. M. Gay, Y. Xi, S. Sivajothi, V. Sivakamasundari, J. Fujimoto, M. Bolisetty, P. M. Hartsfield, V. Balasubramaniyan, M. D. Chalishazar, C. Moran, N. Kalhor, J. Stewart, H. Tran, S. G. Swisher, J. A. Roth, J. Zhang, J. de Groot, B. Glisson, T. G. Oliver, J. V. Heymach, I. Wistuba, P. Robson, J. Wang, L. A. Byers, *Nat. Cancer* **2020**, *1*, 423.
[35] S. Taavitsainen, N. Engedal, S. Cao, F. Handle, A. Erickson, S. Prekovic, D. Wetterskog, T. Tolonen, E. M. Vuorinen, A. Kiviaho, R. Nätkin, T. Häkkinen, W. Devlies, S. Henttinen, R. Kaarijärvi, M. Lahnalampi, H. Kaljunen, K. Nowakowska, H. Syvälä, M. Bläuer, P. Cremaschi, F. Claessens, T. Visakorpi, T. L. J. Tammela, T. Murtola, K. J. Granberg, A. D. Lamb, K. Ketola, I. G. Mills, G. Attard, et al., *Nat. Commun.* **2021**, *12*, 5307.
[36] K. Hallmann, A. P. Kudin, G. Zsurka, C. Kornblum, J. Reimann, B. Stüve, S. Waltz, E. Hattingen, H. Thiele, P. Nürnberg, C. Rüb, W. Voos, J. Kopatz, H. Neumann, W. S. Kunz, *Brain* **2016**, *139*, 338.
[37] L. Huang, et al., *Stem Cell Res. Ther.* **2022**, *13*, 292.
[38] S. M. Lundberg, S.-I. Lee, in Proc. 31st Int. Conf. NIPS, Curran Associates Inc, Long Beach, CA **2017**, 4768.
[39] V. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Advances in Neural Information Processing Systems*, **2017**, p. 30.
[40] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, arXiv preprint arXiv:1710.10903 **2017**.
[41] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, P. T. Ellinor, *Nature* **2023**, *618*, 616.
[42] K. He, X. Zhang, S. Ren, J. Sun, in *Proc. IEEE Conf. CVPR* IEEE, Las Vegas, NV, USA **2016**, 770.
[43] M. A. Wilson, F. Zhao, S. Khare, J. Roszik, S. E. Woodman, K. D'Andrea, B. Wubbenhorst, D. L. Rimm, J. M. Kirkwood, H. M. Kluger, L. M. Schuchter, S. J. Lee, K. T. Flaherty, K. L. Nathanson, *Clin. Cancer Res.* **2016**, *22*, 374.
[44] C. Willyard, *Nat. Med.* **2015**, *21*, 206.
[45] I.-W. Kim, N. Han, M. G. Kim, T. Kim, J. M. Oh, *Pharmacogenet. Genomics* **2015**, *25*, 1.
[46] R. L. Collins, J. T. Glessner, E. Porcu, M. Lepamets, R. Brandon, C. Lauricella, L. Han, T. Morley, L.-M. Niestroj, J. Ulirsch, S. Everett, D. P. Howrigan, P. M. Boone, J. Fu, K. J. Karczewski, G. Kellaris, C. Lowther, D. Lucente, K. Mohajeri, M. Nõukas, X. Nuttle, K. E. Samocha, M. Trinh, F. Ullah, U. Võsa, M. E. Hurles, S. Aradhya, E. E. Davis, H. Finucane, J. F. Gusella, et al., *Cell* **2022**, *185*, 3041.
[47] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, *Nat. Biotechnol.* **2015**, *33*, 495.
[48] G. S. Gulati, S. S. Sikandar, D. J. Wesche, A. Manjunath, A. Bharadwaj, M. J. Berger, F. Ilagan, A. H. Kuo, R. W. Hsieh, S. Cai, M. Zabala, F. A. Scheeren, N. A. Lobo, D. Qian, F. B. Yu, F. M. Dirbas, M. F. Clarke, A. M. Newman, *Science* **2020**, *367*, 405.

[49] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, C. Trapnell, *Nat. Methods* **2017**, *14*, 979.
[50] A. Federico, S. Monti, *Bioinformatics* **2020**, *36*, 1307.
[51] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, A. A. Alizadeh, *Nat. Biotechnol.* **2019**, *37*, 773.
[52] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, Q. Nie, *Nat. Commun.* **2021**, *12*, 1088.
[53] S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, S. Aerts, *Nat. Methods* **2017**, *14*, 1083.