Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Tumor2Line: population-aware tumor–cell line alignment for cancer cell line selection

Bin Duan

binduan@sjtu.edu.cn

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University https://orcid.org/0000-0002-2694-4427

Hua Zou

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University

Zhiye Zhao

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University

Xinyu Zhu

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University

Mingli Luo

Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University

Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11838470/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Tumor2Line: population-aware tumor–cell line alignment for cancer cell line selection

Hua Zou[1], Zhiye Zhao[1], Xinyu Zhu[1], Mingli Luo[1], Bin Duan[1],*

[1] Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China

*Corresponding author: Bin Duan, binduan@sjtu.edu.cn

Abstract

Cancer cell lines are widely used in cancer research and drug development, yet their representativeness of patient tumors varies across models and biological contexts. Here, we introduce Tumor2Line, a population-aware framework that reframes cancer cell-line selection as an alignment problem between tumor and cell-line populations using unbalanced optimal transport. Tumor2Line improved the recovery of documented tumor–cell line pairs over conventional similarity-based approaches. Systematic analysis across 22 cancer lineages revealed that tumor-cell-intrinsic programs were generally better preserved, whereas programs with greater tumor microenvironment dependence showed lower fidelity. Furthermore, Tumor2Line-selected cell lines showed drug-sensitivity patterns associated with clinical treatment outcomes and improved concordance in genetic dependency profiles with patient-derived organoids. These findings provide a basis for more informed selection of cancer cell-line models and help researchers interpret preclinical findings in the context of model-specific representational fidelity. Tumor2Line is available at https://github.com/Duan-Lab1/Tumor2Line

Keywords: Cancer cell lines; Tumor–cell line alignment; Population-aware modeling; Cancer model fidelity; Optimal transport

Introduction

Cancer cell lines remain among the most widely used experimental models for studying cancer biology, identifying therapeutic vulnerabilities, and evaluating pharmacological and genetic perturbations[1-3]. Several large-scale resources provide complementary transcriptomic, pharmacological, and genetic-dependency profiles across thousands of primary tumors and cancer cell lines[4]. These include the Treehouse Tumor Compendium, which integrates data from Treehouse[5], the Therapeutically Applicable Research to Generate Effective Treatments program (TARGET)[6], and The Cancer Genome Atlas (TCGA)[7], as well as the Cancer Cell Line Encyclopedia (CCLE)[2], Genomics of Drug Sensitivity in Cancer (GDSC)[8], and the Cancer Dependency Map (DepMap)[3]. These resources have greatly expanded the ability to investigate cancer mechanisms and therapeutic responses using cancer cell line models. However, the biological relevance of discoveries derived from these models depends critically on a fundamental question: which cancer cell line best represents a given human tumor, and to what extent? In practice, cancer cell line selection is often guided by tissue of origin, commonly used laboratory models, or molecular similarity to patient tumors[9-11]. Such strategies may overlook substantial heterogeneity among tumors within the same cancer lineage. Moreover, cancer cell lines can diverge from their tumors of origin through tumor evolution, model establishment, and long-term adaptation to in vitro conditions[12,13]. These differences may limit the ability of individual cell lines to faithfully capture the molecular and functional states of human tumors.

Recent efforts have sought to improve tumor–cell line selection by systematically comparing molecular profiles between cancer cell lines and primary tumors[14-18]. For example, Celligner addresses systematic molecular differences between tumors and cell lines through global alignment of expression landscapes[17], whereas TumorComparer incorporate biologically informative features to improve tumor–model similarity assessment[16]. Despite these advances, these studies also revealed substantial heterogeneity in the representativeness of cell lines, showing that cell lines from the same cancer type can differ markedly in how well they capture human tumors[16,17]. This heterogeneity highlights three unresolved challenges in cancer cell-line selection. (1) Most existing approaches evaluate tumor–cell line correspondence through pairwise similarity, treating each tumor–cell line comparison independently and overlooking the population context

provided by other tumors and candidate cell lines[10,14,16]. Specifically, a tumor represents one state within a heterogeneous population, and the suitability of a cell line may depend on both its intrinsic similarity and its position within the broader tumor landscape[12,14]. Likewise, the representational capacity of a cell-line collection depends on its coverage and distribution across diverse tumor states[17,18]. (2) Tumor and cell-line populations represent distinct and incompletely overlapping biological landscapes[13]. Patient tumors are shaped by evolutionary processes, microenvironmental interactions, and intratumoral heterogeneity, whereas cell lines represent selected subsets of tumor states that have undergone adaptation to in vitro conditions[12,13,19]. Thus, tumor-cell lines correspondence is inherently incomplete and need not be one-to-one. (3) Existing approaches primarily focus on prioritizing individual models but provide limited insight into which biological programs are faithfully represented—or poorly captured—by available cancer cell lines across cancer types[14,17].

Thus, we propose a novel perspective that cancer cell-line selection should not be viewed solely as a pairwise similarity matching problem, but rather as an alignment problem between heterogeneous biological populations[20,21]. In this formulation, the tumor population provides contextual information for interpreting an individual tumor within the broader spectrum of human cancer states, while the cell-line population defines the landscape of available experimental models. By jointly considering these two populations, model selection can leverage information beyond direct molecular similarity and account for the relationships between individual tumors and the broader model landscape. This perspective also enables systematic evaluation of how faithfully existing cell lines represent different aspects of human tumor biology.

Here, we developed Tumor2Line, a computational framework that implements this population-aware perspective by formulating cancer cell-line selection as a population-aware tumor–cell line alignment problem. Tumor2Line projects tumors and lineage-compatible candidate cell lines into a shared representation space and applies unbalanced optimal transport to infer population-level correspondence and tumor-specific model rankings. We first benchmarked Tumor2Line using experimentally tumor–cell line pairs to evaluate its matching performance and the contribution of population context. We then systematically characterized the representational fidelity of cancer cell lines across cancer lineages and biological programs, revealing systematic variation in their

ability to capture different aspects of tumor biology. Finally, we evaluated the functional relevance of selected cell lines by examining their concordance with clinical drug responses and patient-derived organoid genetic dependencies. Together, this work establishes a novel population-aware approach to cancer cell-line selection and provides a systematic assessment of the molecular and functional representativeness of existing models.

Results

Tumor2Line formulates cancer cell-line selection as a population-aware alignment problem

Tumors and cell lines represent heterogeneous and incompletely overlapping biological landscapes, making tumor–cell line matching fundamentally different from conventional nearest-neighbor similarity search (**Fig. 1A**). Existing approaches typically rank candidate models through independent pairwise comparisons, overlooking the population context provided by the broader tumor and cell-line landscapes. We therefore developed Tumor2Line, which reframes cancer cell-line selection as a population-aware alignment problem between tumor and cell line populations.

Tumor2Line consists of two key components: a shared representation and population-aware alignment (**Fig. 1B–D**). First, tumor and lineage-compatible cell-line transcriptomes are mapped into a shared low-dimensional space using principal component analysis (PCA). To establish a stable and reusable representation, PCA is constructed from large-scale reference tumor and cell-line datasets (Treehouse, TARGET, TCGA and CCLE), allowing new samples to be projected into a common coordinate system. This reference-based strategy is particularly advantageous in practical applications where only a limited number of patient tumor samples are available. Importantly, major transcriptional differences between tumors and cell lines can reflect tumor microenvironmental signals, which can contribute substantially to high-variance principal components (PCs) and obscure more subtle but potentially informative biological differences. To balance their contributions without discarding these signals, PC scores are standardized to unit variance before tumor–cell line alignment (**Fig. 1C**).

Second, Tumor2Line employs unbalanced optimal transport (uOT) to align tumor and cell-line populations within the shared representation space (**Fig. 1D**). Unlike conventional pairwise similarity approaches, which rank cell lines independently for each tumor, optimal transport considers the distributional relationships between the two populations when estimating tumor–cell line correspondence. The unbalanced formulation accommodates incomplete overlap between tumor and cell-line populations by relaxing, strict mass-conservation constraints. The resulting transport plan is used to generate ranked candidate cell lines for individual tumors.

To assess the utility of Tumor2Line, we designed a multi-level evaluation framework encompassing the recovery of documented tumor–cell line pairs, systematic characterization of tumor–cell line molecular state fidelity, and independent functional validation using clinical drug responses and patient-derived organoid genetic dependencies (**Fig. 1E**). Together, Tumor2Line provides a population-aware strategy for selecting cancer cell lines and evaluating their representativeness as experimental models of human tumors.

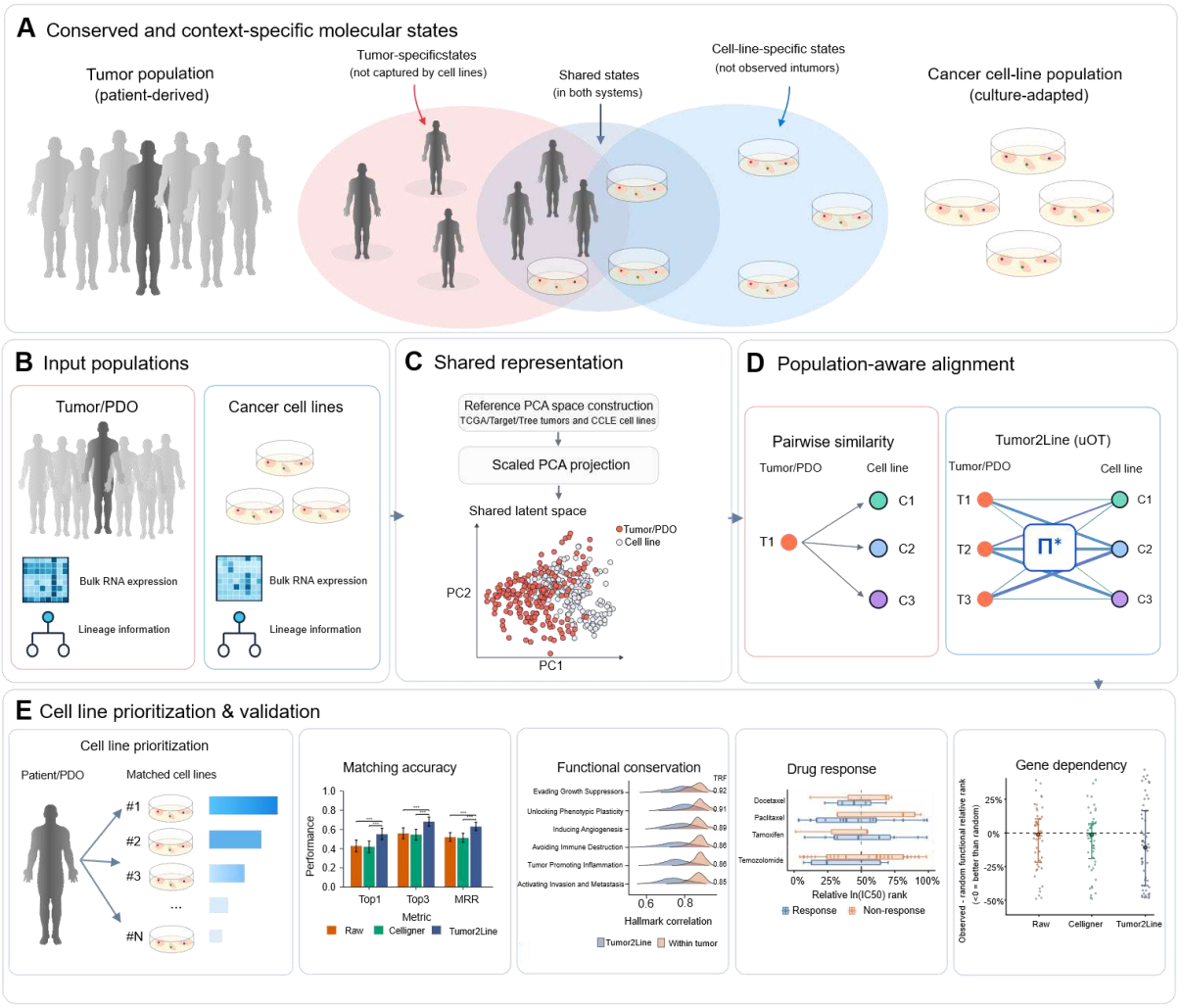

Figure 1 | Overview of the Tumor2Line framework for population-aware tumor–cell line matching. (A) Schematic illustration of shared and context-specific molecular states between human tumors and cancer cell lines. **(B)** Input tumor or patient-derived organoid (PDO) and cancer cell-line populations with transcriptomic profiles and lineage annotations. **(C)** Construction of a shared latent representation using scaled principal component analysis (PCA). For primary tumor analyses, an external reference PCA space was constructed using TARGET, TCGA and Treehouse tumor profiles together with CCLE cell-line profiles, followed by projection of study-specific samples into this space. **(D)** Comparison of conventional pairwise similarity matching and Tumor2Line population-aware alignment using unbalanced optimal transport (uOT). **(E)** Overview of Tumor2Line evaluation, including recovery of known tumor–cell line pairs, assessment of molecular state fidelity, and functional validation through drug-response and gene-dependency concordance.

Tumor2Line improves recovery of documented matched tumor–cell line pairs

To evaluate the ability of Tumor2Line to identify biologically related cell lines, we first benchmarked its performance using 177 tumor–cell line pairs with documented patient-tumor origins from the Patient-Derived Models Repository (PDMR)[22] (**Fig. 2A**). These tumor–cell line pairs provide a provenance-based reference because each cell line has a documented origin from a corresponding patient tumor, allowing an objective assessment of whether Tumor2Line can prioritize biologically related cell lines. All methods were evaluated using identical tumor samples and candidate cell line pools, with the documented tumor-derived cell line used as the reference ranking target (**Supplementary Table 1**).

We established a conventional baseline by ranking candidate cell lines according to their Pearson correlation with each tumor based on their raw gene expression profiles. We additionally compared Tumor2Line with Celligner, an established approach for aligning tumor and cell-line transcriptomic profiles[17]. As Celligner does not provide a dedicated tumor–cell line matching procedure, Pearson correlation was applied to its aligned expression profiles to generate cell-line rankings. Matching performance was assessed using Top1 and Top3 accuracy and mean reciprocal rank (MRR). Top1 and Top3 measure the proportions of tumors whose documented partners rank first or within the top three candidates, respectively. MRR accounts for the ranking positions of documented partners. We can see that Tumor2Line consistently outperformed both the conventional baseline and Celligner across all three metrics, achieving a Top1 recovery rate of 68.9%, compared with 50.3% and 49.7%, respectively (**Fig. 2B; Supplementary Table 2**). This improvement was observed across multiple cancer lineages and remained consistent when evaluating cumulative recovery across ranking thresholds (**Fig. 2C, D**).

To assess the contributions of Tumor2Line's two key components, scaled PCA representation and population-aware uOT alignment, we performed ablation analyses by removing each component in turn. Removing either component reduced matching performance across Top1, Top3, and MRR, demonstrating their complementary contributions to the recovery of documented tumor–cell line pairs (**Fig. 2E**).

We further evaluated the design choices underlying the shared representation. Specifically, we compared reference-based PCA, in which an external reference space was constructed using independent tumor and cell-line transcriptomic datasets, with PDMR-specific PCA, in which the representation was constructed directly from PDMR samples. The two approaches achieved comparable matching performance in the full benchmark, whereas reference-based PCA showed an advantage in smaller cohorts (**Supplementary Fig. 1**). To further investigate the biological basis underlying this design of scaled PCA, we examined the relationship between PC explained variance and tumor–cell line source effects. PC explained variance was positively correlated with source partial R^2 (**Supplementary Fig. 2C**). Moreover, PCs most strongly associated with tumor–cell line differences showed predominant enrichment for immune-related and tumor microenvironment-associated biological processes (**Supplementary Fig. 2D**). These findings support the use of a reference-based, scaled PCA representation, particularly for applications involving limited patient samples.

Furthermore, we investigated how scaled PCA and uOT contribute to improved matching performance. First, scaled PCA increased the similarity margin between documented tumor–cell line pairs and their most similar incorrect candidates compared with raw expression and Celligner-aligned representations (**Fig. 2F**), suggesting enhanced discrimination of documented matched models. Then, to assess the contribution of population context to uOT, we progressively incorporated additional same-lineage tumors while keeping the candidate cell-line pool fixed. We can see that increasing tumor population context was associated with improved rankings of documented partners in both colorectal adenocarcinoma (COAD) and pancreatic adenocarcinoma (PAAD) (**Fig. 2G**). Together, these findings suggest that scaled PCA enhances the discrimination of documented matched models, whereas uOT leverages population-level information beyond independent pairwise similarity.

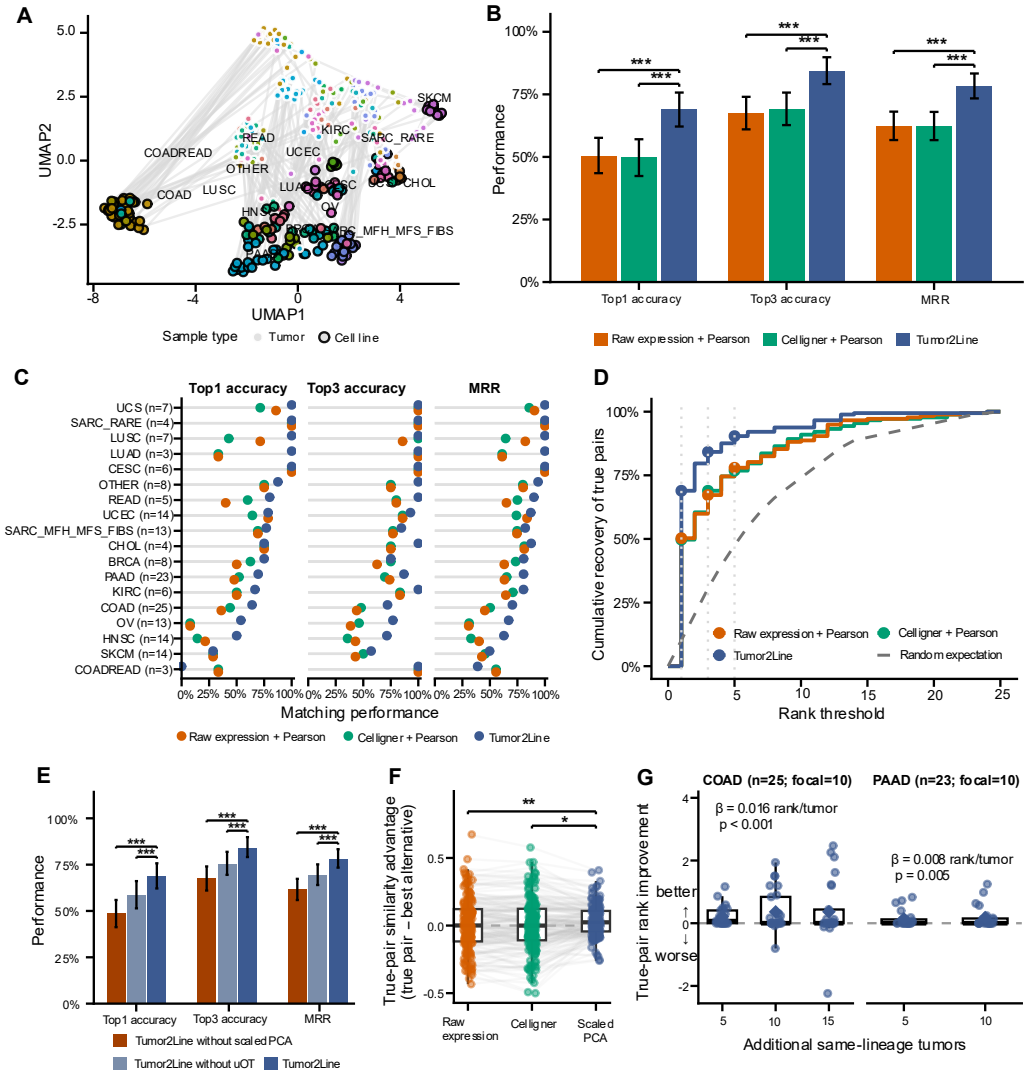

Figure 2 | Population-aware matching improves recovery of documented matched tumor–cell line pairs. (A) Uniform manifold approximation and projection (UMAP) of PDMR tumors and cell lines, colored by cancer lineage. Lines connect documented matched pairs. **(B)** Overall Top1 accuracy, Top3 accuracy and mean MRR across three matching methods. **(C)** Lineage-specific matching performance across the three methods. **(D)** Cumulative recovery of known tumor–cell line pairs across rank thresholds. **(E)** Ablation analysis of scaled PCA and uOT contributions to matching accuracy. **(F)** True-pair similarity advantage across transcriptomic representations, with lines connecting measurements from the same tumor. **(G)** Effects of additional same-lineage tumors on true-pair rank improvement during uOT matching in COAD and PAAD cancer. Positive values indicate improved ranking. Statistical comparisons in **B** and **E** used exact McNemar tests for Top1 and Top3 accuracy and paired Wilcoxon signed-rank tests for MRR. Comparisons in **F** used paired Wilcoxon signed-rank tests. Asterisks indicate statistical significance (*p < 0.05, **p < 0.01, ***p < 0.001). In **G**, β represents the estimated change in true-pair rank per additional tumor, with the corresponding linear regression p value.

Tumor2Line reveals heterogeneous representational fidelity of cancer cell lines

Having established the ability of Tumor2Line to recover documented matched tumor–cell line pairs, we next applied Tumor2Line to systematically assess how faithfully existing cancer cell lines represent human tumors across diverse biological contexts (**Fig. 3A**). We matched 11,087 TCGA tumors with 1,158 lineage-compatible CCLE cell lines across 22 cancer lineages (**Supplementary Table 3**), generating tumor-specific cell line rankings.

We first assessed representational fidelity at the whole-transcriptome level. Across cancer lineages, Tumor2Line-selected cell lines exhibited higher expression-pattern similarity to their matched tumors than randomly selected same-lineage cell lines (**Fig. 3B**). Nevertheless, within-lineage tumor–tumor similarity remained higher than tumor–cell line similarity, indicating a persistent molecular gap even among the most closely matched available cell lines.

We next examined whether this representational gap extended to cancer-associated biological programs. Using genes associated with cancer Hallmark programs[23,24] (**Supplementary Table 4**), we observed a similar pattern: selected cell lines showed greater similarity to matched tumors than random models, but remained less similar than tumors within the same lineage (**Fig. 3C**). These observations motivated a systematic assessment of fidelity across individual Hallmark programs.

To quantify program-specific fidelity, we defined a tumor representational fidelity index (TRF) as the ratio of median tumor–cell line similarity to median within-lineage tumor–tumor similarity for each lineage–Hallmark combination. Higher TRF values indicate greater preservation of tumor-associated expression patterns in selected cell lines relative to within-lineage tumor similarity. Across 13 cancer Hallmark programs, TRF values ranged from approximately 0.85 to 0.96 (**Fig. 3D; Supplementary Tables 4 and 6**). Programs associated with tumor-cell-intrinsic properties, including genome instability, epigenetic regulation, resistance to cell death and replicative immortality, generally exhibited higher TRF values. In contrast, programs related to invasion, tumor-promoting inflammation, immune evasion and angiogenesis showed lower fidelity. To exclude potential bias arising from the overlap between genes used for transcriptomic matching

and those used for Hallmark-level fidelity assessment, we repeated the analysis after excluding the union of genes from all 13 Hallmark programs from the matching representation. The resulting TRF patterns closely recapitulated the original results (Spearman's $\rho = 1.00$, $p < 0.001$; **Supplementary Fig. 4**). To investigate factors associated with heterogeneity in program-specific fidelity, we examined whether Hallmark programs exhibiting greater intertumoral variability within cancer lineages were more difficult to represent using cell-line models. Across the 13 Hallmark programs, greater within-lineage tumor variability was associated with lower TRF values (Spearman's $\rho = -0.75$, $p = 0.003$; **Fig. 3E**), suggesting that biological programs exhibiting greater diversity across patient tumors are less consistently captured by existing cell lines.

To further characterize differences in representational fidelity across cancer lineages, we performed hierarchical clustering of 22 cancer lineages based on their TRF profiles across 13 Hallmark programs (**Fig. 3F; Supplementary Table 6**). This analysis stratified cancer lineages into three groups with low, intermediate, and high overall representational fidelity, revealing substantial differences in how well existing cell lines capture tumor biology across cancer types. Blood cancers exhibited the highest fidelity, potentially reflecting their relatively lower dependence on the complex microenvironmental interactions characteristic of solid tumors[25]. In contrast, thyroid cancer showed the lowest fidelity, consistent with our existing understanding of the limited representativeness of available cell-line models for this lineage[26,27].

Given the reduced fidelity of several microenvironment-associated programs, we next investigated whether TME dependence was associated with representational fidelity. Across lineage–Hallmark combinations, greater TME dependence was associated with lower representational fidelity (Spearman's $\rho = -0.47$, $p < 0.001$; **Fig. 3G, H**). This association suggests that biological programs relying more heavily on interactions with the tumor microenvironment are less faithfully represented by conventional cell-line cultures, which lack the corresponding stromal and immune components. To assess whether the observed fidelity patterns extended beyond the 13 selected cancer-related Hallmark programs, we further evaluated all 50 MSigDB Hallmark gene sets, revealing similar patterns of representational fidelity (**Supplementary Fig. 5; Supplementary Table 5**). Together, these analyses reveal that the representational fidelity of cancer cell lines is highly context dependent, varying across cancer lineages and biological programs.

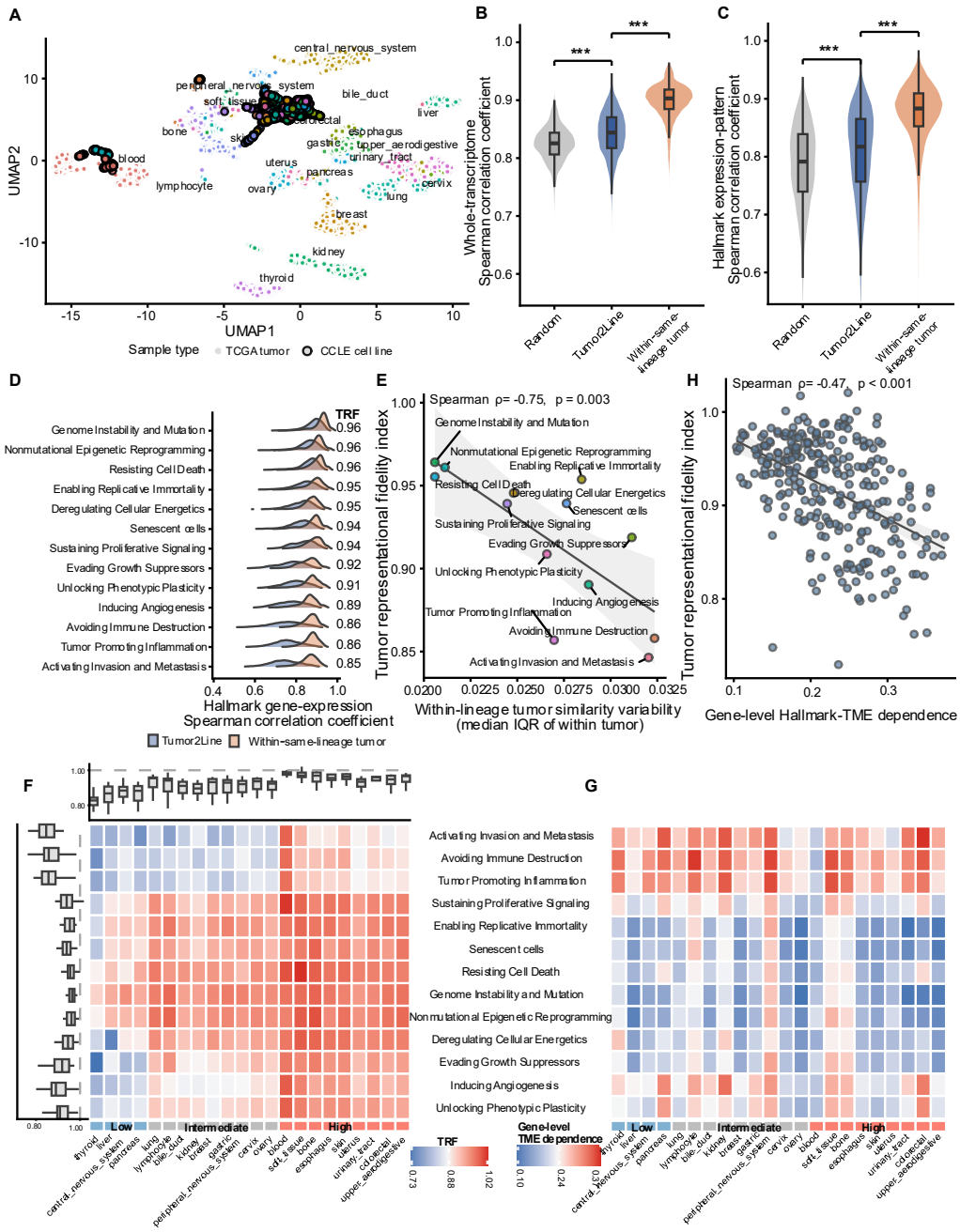

Figure 3 | Tumor2Line reveals heterogeneous representational fidelity across cancer Hallmark programs. **(A)** UMAP of 11,087 TCGA tumors and 1,158 CCLE cell lines across 22 cancer lineages. Samples are colored by cancer lineage, with outlined circles indicating cell lines. **(B)** Whole-transcriptome Spearman correlations between tumors and randomly selected cell lines, Tumor2Line-selected cell lines or within-same-lineage tumor. **(C)** Corresponding comparisons based on cancer Hallmark-associated genes. **(D)** Hallmark-specific expression similarity and TRF across 13 cancer Hallmark programs. **(E)** Association between within-same-lineage tumor similarity variability and TRF across Hallmark programs. **(F)** Lineage-by-Hallmark TRF landscape across 22 cancer lineages and 13 Hallmark programs. Low, Intermediate and High indicate cancer lineage groups stratified

by TRF. Marginal boxplots summarize TRF distributions across lineages and Hallmarks. **(G)** Corresponding landscape of TME dependence across lineage–Hallmark combinations. **(H)** Association between Hallmark–TME dependence and TRF across lineage–Hallmark combinations. Statistical comparisons in **B** and **C** were performed using unpaired Wilcoxon rank-sum test, with asterisks indicating significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Tumor2Line improves functional concordance in drug response and genetic dependency

While molecular similarity provides an important basis for selecting representative cancer models, whether it translates into functional relevance remains an open question. We therefore evaluated functional concordance in two complementary settings: the association between cell-line drug sensitivity and clinical treatment outcomes, and the concordance of genetic dependencies between cell lines and patient-derived organoids.

We first assessed pharmacological concordance by integrating TCGA treatment outcomes with GDSC drug-sensitivity measurements across matched lineage–drug contexts (**Fig. 4A; Supplementary Table 7**). For each tumor, the selected cell line was evaluated based on its relatice rank of ln(IC50) within the corresponding drug-sensitivity assay, with lower ranks indicating greater drug sensitivity. Patients with complete or partial responses (CR/PR) were classified as responders, whereas those with stable or progressive disease (SD/PD) were classified as non-responders. As shown in Fig. 4B, Tumor2Line-selected cell lines exhibited significantly lower relative ln(IC50) ranks for responders than for non-responders ($p = 0.034$; **Supplementary Table 8**). In contrast, no significant responder–non-responder differences were observed with random matching ($p = 0.574$), raw-expression Pearson matching ($p = 0.286$), or Celligner-based Pearson matching ($p = 0.995$). Across individual lineage–drug contexts, most showed the same direction of association, with greater drug sensitivity among cell lines assigned to clinically responding tumors, although this trend was not consistent across all contexts (**Supplementary Fig. 6**). Together, these findings suggest that Tumor2Line-based molecular matching can identify cell lines whose drug-sensitivity profiles are more closely associated with clinical treatment outcomes.

We next sought to extend functional validation to genetic dependencies, a more challenging dimension of cross-model concordance. Unlike pharmacological responses, which reflect the

effects of drug treatments, genetic dependencies capture the functional consequences of perturbing individual genes, which can be more subtle and context dependent. As patient-level genetic perturbation data are currently unavailable, we used patient-derived organoids (PDOs) as patient-derived functional models to provide a tumor-relevant reference. We integrated PDO transcriptomic profiles and MiniLib CRISPR perturbation data with corresponding cancer cell-line transcriptomic profiles and DepMap GeneEffect measurements across five cancer types (**Fig. 4C**). For this analysis, PDOs and lineage-matched cell lines were projected into jointly constructed, lineage-specific PCA spaces. Importantly, genetic dependency measurements were excluded from the matching process and used exclusively for downstream functional evaluation, enabling an independent assessment of genetic dependency concordance.

We quantified genetic dependency concordance using functional relative ranks (FRR), with lower values indicating better concordance. Across the pooled PDO–cell line cohorts, Tumor2Line achieved a lower mean FRR than expected under random matching (empirical $p = 0.024$; **Fig. 4D; Supplementary Table 9**). In contrast, neither raw-expression Pearson nor Celligner-based Pearson matching showed significant improvements over random matching ($p = 0.897$ and $p = 0.249$, respectively). Direct comparisons further showed lower FRR values for Tumor2Line than for raw-expression Pearson matching ($p < 0.001$) and Celligner-based Pearson matching ($p = 0.011$; **Fig. 4E**). Although the magnitude of improvement varied across cancer types (**Supplementary Fig. 7**), the pooled results support the ability of Tumor2Line to prioritize cell lines with genetic dependency profiles more concordant with those of PDOs.

Together, these analyses demonstrate that transcriptome-based matching can identify cell lines with greater functional concordance in both pharmacological responses and genetic dependencies, extending the evaluation of Tumor2Line beyond molecular representativeness.

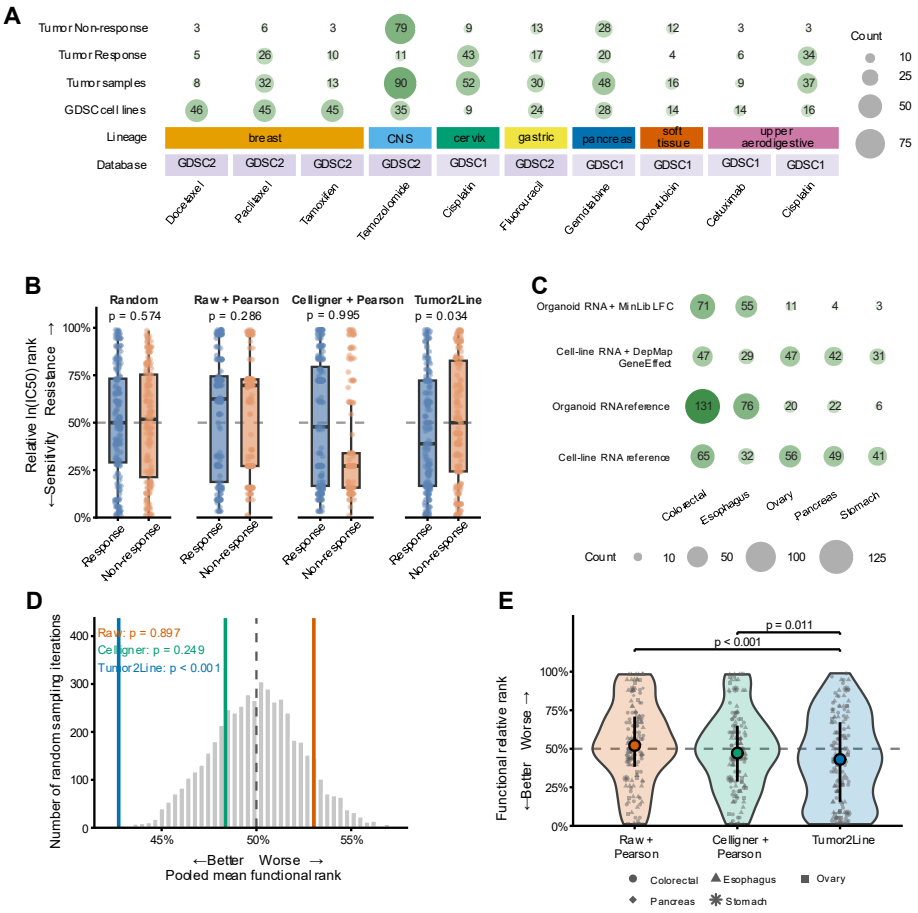

Figure 4 | Tumor2Line improves functional concordance in drug response and genetic dependency. (A) Availability of clinically annotated tumors and GDSC cell lines across cancer lineage–drug combinations. Numbers indicate non-responding tumors, responding tumors, total tumor samples and GDSC cell lines. CNS denotes central nervous system. **(B)** Relative ln(IC50) ranks of cell lines assigned to clinically responding and non-responding tumors by random matching, Raw expression + Pearson, Celligner + Pearson and Tumor2Line. Lower relative ranks indicate greater drug sensitivity. **(C)** Availability of PDO transcriptomic and MiniLib perturbation data, cell-line transcriptomic and DepMap GeneEffect data, and corresponding organoid and cell-line RNA reference cohorts across five cancer types. **(D)** Null distribution of pooled mean functional relative ranks generated from 5,000 random sampling iterations. Colored vertical lines indicate the observed pooled mean ranks for each matching method, with corresponding empirical p values. **(E)** Functional relative rank distributions across Raw expression + Pearson, Celligner + Pearson and Tumor2Line. Individual points represent PDOs, with different shapes indicating cancer lineages. Central points and vertical bars indicate medians and interquartile ranges. Statistical comparisons in **B** were performed using one-sided unpaired Wilcoxon rank-sum tests. In **D**, empirical p values were calculated using a lower-tail randomization test comparing observed pooled mean functional relative ranks with the null distribution generated from 5,000 random sampling iterations. In **E**, pairwise comparisons of Tumor2Line with the

two comparator methods were performed using one-sided paired Wilcoxon signed-rank tests, with PDOs matched by sample ID.

Discussion

Cancer cell lines are indispensable models for investigating cancer biology and therapeutic vulnerabilities[3,28], yet identifying models that faithfully represent individual patient tumors remains a major challenge[10,17]. Here, we developed Tumor2Line, a population-aware framework that improves tumor–cell line matching by incorporating information from the broader tumor population. Beyond recovering experimentally documented tumor–cell line relationships, Tumor2Line enabled a systematic assessment of how faithfully existing cell lines represent human tumors across cancer lineages and biological programs. We found that tumor-cell-intrinsic programs were generally better represented than microenvironment-associated processes, revealing substantial heterogeneity in model fidelity. Furthermore, Tumor2Line-selected cell lines exhibited improved concordance with independent pharmacological and genetic dependency measurements. Together, these findings suggest that cancer model selection should move beyond overall molecular similarity toward a more context-dependent evaluation of biological representativeness and functional relevance.

A central conceptual advance of Tumor2Line is the recognition that cancer model selection can benefit from population-level information rather than relying exclusively on independent pairwise comparisons. By jointly aligning tumor and cell-line populations, Tumor2Line allows the broader distribution of tumor states to inform the selection of models for individual patients[20,21,29]. This approach provides a complementary perspective to conventional similarity-based methods, particularly when tumor and model populations exhibit systematic molecular differences and incomplete correspondence[14,17,30]. Importantly, our findings also highlight that molecular representativeness is not a universal property of an experimental model. Previous pan-cancer studies have similarly demonstrated that tumor–cell line similarity varies across cancer lineages and depends on the molecular features and biological pathways used for comparison[14,16,17]. Even among the most closely matched cell lines, fidelity varied considerably across biological programs and cancer lineages[17]. Programs associated with intrinsic tumor-cell functions were relatively well

preserved, whereas those involving interactions with the tumor microenvironment were less faithfully represented[19,23,31,32]. These observations suggest that the suitability of a cancer model should be evaluated in relation to the specific biological processes and experimental questions being investigated, rather than assuming that a single model can adequately represent all aspects of a patient's tumor.

Several limitations should be considered. First, Tumor2Line can only prioritize models within the available candidate pool and cannot recover biological states that are absent from existing cell-line collections. Consequently, model selection remains dependent on the diversity and coverage of available experimental systems. Second, the current framework primarily relies on transcriptomic profiles, and differences in cellular composition between bulk tumor tissues and cultured cell lines may influence estimates of representational fidelity, particularly for microenvironment-associated programs[17]. Integrating additional molecular and cellular information may enable a more comprehensive assessment of model fidelity. Third, functional validation remains limited by the availability and comparability of clinical and experimental perturbation data. Clinical drug responses reflect factors beyond tumor-cell-intrinsic sensitivity[19,32], while genetic dependency concordance was evaluated using patient-derived organoids rather than primary tumors[33]. Thus, the observed functional associations support the relevance of transcriptome-based model selection but do not establish its ability to predict individual patient treatment outcomes.

More broadly, the population-aware alignment principle underlying Tumor2Line could be extended beyond conventional cancer cell lines to other experimental systems, including patient-derived organoids, xenografts and engineered cellular models[30,34]. Such extensions may enable a more general framework for selecting experimental models according to their fidelity for specific biological states and functional responses. This perspective is also relevant to emerging AI virtual cell models, which increasingly learn cellular responses from large-scale perturbation datasets generated in experimental model systems[35,36] but ultimately seek to predict biological behavior in human tissues[37,38]. A critical challenge is therefore to determine which molecular states and perturbation responses can be reliably transferred across experimental systems and patient tumors[39,40]. By systematically characterizing model representativeness and its biological

limitations, Tumor2Line provides a foundation for more informed experimental model selection and, ultimately, for bridging model-derived discoveries with human tumor biology.

Materials and Methods

Datasets

Transcriptomic data. RNA-seq gene expression profiles of patient-derived tumors and cancer models were obtained from the NCI Patient-Derived Models Repository (PDMR; https://pdmdb.cancer.gov/web/apex/f?p=101:1)[22]. Gene-level RSEM quantification files were downloaded, and TPM values were extracted and merged into a sample-by-gene expression matrix. The PDMR dataset was used to establish the tumor–cell line benchmark for evaluating matching performance. To construction of the reference transcriptomic space, RNA-seq gene expression profiles of tumor tissues were obtained from the Treehouse Tumor Compendium (https://xena.treehouse.gi.ucsc.edu/)[41] which integrates samples from the Treehouse Childhood Cancer Initiative[5], the Therapeutically Applicable Research to Generate Effective Treatments program (TARGET)[6] and The Cancer Genome Atlas (TCGA)[7]. the Cancer Cell Line Encyclopedia (CCLE) cell-line expression profiles were obtained from the Cancer Dependency Map (DepMap; https://figshare.com/articles/dataset/DepMap_19Q4_Public/11384241/2)[1,2]. Corresponding sample annotations and gene annotation files were obtained from the Celligner dataset[17]. RNA-seq expression profiles of patient-derived tumor organoids and their corresponding clinical and cancer-type annotations were obtained from the supplementary datasets of Herranz-Ors et al[42]. The organoid expression profiles were extracted from *data_RNAseq_ALL-organoids.csv* and used for transcriptomic matching and downstream functional validation.

Drug-response data. Clinical drug-response data from TCGA patients, including treatment records and clinical response annotations, were obtained from Ding et al[43]. Cell-line drug-sensitivity measurements were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org), including GDSC1 and GDSC2[8]. Fitted ln(IC50) values were used as the quantitative measure of cell-line drug sensitivity.

CRISPR dependency data. Organoid CRISPR–Cas9 screening data were obtained from Herranz-Ors et al., with gene-level perturbation effects represented by log-fold-change (LFC) values from the MiniLibCas9 screening resource[42]. Cell-line CRISPR dependency profiles were obtained from the DepMap (https://depmap.org/portal/), with Chronos-derived GeneEffect scores used to quantify gene dependency[3]. Clinical and tumor-type annotations for organoids were retrieved from the corresponding supplementary datasets[42]. Gene identifiers were harmonized, and shared genes between the organoid LFC and DepMap GeneEffect matrices were retained for functional concordance analyses.

Cancer Hallmark gene sets. Gene sets representing 13 cancer hallmarks[24,44] were obtained from Sibai M et al. (**Supplementary Table 4**)[23]. The 50 Hallmark (H) gene sets were retrieved from the Molecular Signatures Database (MSigDB)[45] using the R package msigdbr (version 25.1.1; **Supplementary Table 5**).

Data preprocessing

Transcriptomic data preprocessing. Gene expression matrices were organized with samples as rows and genes as columns. The filtered TPM values were then transformed as $\log_2(\text{TPM} + 1)$, and the top 10,000 genes with the highest standard deviations were selected for downstream analyses. Data already provided on a $\log_2(\text{TPM} + 1)$ scale were not transformed again. For reference PCA construction, the 10,000 genes with the highest across-sample variability were selected from the TARGET, TCGA and Treehouse and CCLE datasets. Gene-wise means and standard deviations were calculated from this reference dataset and retained for subsequent projection of external samples.

Cancer lineage harmonization. PDMR samples were classified using the corresponding TCGA cancer lineage annotations. TCGA tumors and CCLE cell lines were assigned lineage labels according to the published Celligner annotations[17]. Organoid cancer types were mapped to the corresponding cancer lineages using their clinical and tumor-type annotations. A harmonized

lineage mapping table was used throughout the analyses. All tumor–cell line matching procedures were restricted to samples and candidate cell lines belonging to the same cancer lineage.

Drug-response data preprocessing. TCGA clinical treatment records were curated to harmonize drug names and retain evaluable clinical responses, including complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD)[43,46]. GDSC cell-line drug-sensitivity data were mapped to CCLE/DepMap identifiers and filtered for valid fitted ln(IC50) values. The two datasets were harmonized by drug identity and cancer lineage, while GDSC1 and GDSC2 assays were processed separately.

Gene dependency data preprocessing. Gene identifiers were harmonized between organoid CRISPR screening data and DepMap cell-line dependency profiles[42]. Only genes shared between the two datasets were retained to construct aligned gene-dependency matrices for downstream functional validation.

Tumor2Line computational framework

Shared transcriptomic representation using scaled PCA. Tumor2Line comprises two main steps: constructing a shared transcriptomic representation using scaled PCA and performing population-aware tumor–cell line matching using uOT. First, to establish a common transcriptomic representation, we constructed a global reference PCA space using the combined Treehouse tumor, TARGET, TCGA and CCLE cancer cell-line expression compendium from Celligner [17]. Gene expression profiles were transformed to $\log_2(\mathrm{TPM} + 1)$ and the 10,000 genes with the highest across-sample variability in the reference dataset were retained. Let $X \in R^{N \times G}$ denote the combined expression matrix containing N samples and G genes. Each gene was centered and standardized across samples:

$$\tilde{X}_{ig} = \frac{X_{ig} - \mu_g}{\sigma_g}$$

where X_{ig} represents the expression of gene g in sample i, and μ_g and σ_g denote the mean and standard deviation of gene g, respectively.

PCA was performed on the standardized expression matrix, and the minimum number of principal components (PCs) explaining at least 90% of the cumulative variance was retained:

$$Z = \tilde{X} V_K$$

where $V_K \in \mathbb{R}^{G \times K}$ contains the loading vectors of the first K PCs, and $Z \in \mathbb{R}^{N \times K}$ represents the resulting PC scores.

PDMR tumor and cell-line expression profiles were independently preprocessed using the same selected genes and reference-derived gene-wise scaling parameters. Their standardized expression profiles were then projected into the fixed reference PCA space using the reference loading matrix. A separate PCA construction strategy was used for PDO-based functional validation, as described in the CRISPR-based functional concordance section.

To reduce the dominance of high-variance PCs in downstream distance calculations, the retained PC scores were further centered and standardized:

$$\hat{Z}_{ik} = \frac{Z_{ik} - \mu_k^{PC}}{\sigma_k^{PC}}$$

where μ_k^{PC} and σ_k^{PC} represent the mean and standard deviation of PC k across samples, respectively. The resulting scaled PC coordinates were used for tumor–cell line matching.

Population-aware matching using unbalanced optimal transport. Within each cancer lineage, pairwise tumor–cell line distances were calculated using the scaled PC coordinates. For $\boldsymbol{n}$ tumors and $\boldsymbol{m}$ candidate cell lines, the Euclidean distance was defined as:

$$\mathrm{D_{ij}} = \|\hat{\mathrm{z}}_\mathrm{i}^{\mathrm{T}} - \hat{\mathrm{z}}_\mathrm{j}^{\mathrm{M}}\|_2$$

where $\hat{z}_i^T$ and $\hat{z}_j^M$ represent the scaled PC coordinates of tumor i and cell line j, respectively. The distances were normalized by the median positive pairwise distance within the corresponding lineage to obtain the transport cost matrix:

$$\mathrm{C_{ij}} = \frac{\mathrm{D_{ij}}}{\mathrm{median}(\mathrm{D_{uv}} : \mathrm{D_{uv}} > 0)}$$

Uniform initial masses were assigned to tumors and candidate cell lines, with $a_i = 1/n$ and $b_j = 1/m$. Tumor2Line then estimated a non-negative transport coupling matrix $\Gamma \in \mathbb{R}_+^{\mathrm{n \times m}}$ by solving the entropically regularized uOT problem[47]:

$$\Gamma^{\star} = \underset{\Gamma \geq 0}{\operatorname{argmin}}$$

$$[\langle \Gamma, C \rangle + \varepsilon \, \mathrm{KL}(\Gamma \parallel a \otimes b) + \rho_T \, \mathrm{KL}(\Gamma \mathbf{1}_m \parallel a) + \rho_M \, \mathrm{KL}(\Gamma^{\mathrm{T}} \mathbf{1}_n \parallel b)]$$

Here, $\langle \Gamma, C \rangle$ denotes the total transport cost, and KL represents the generalized Kullback–Leibler divergence[48]. The parameter ε controls entropic regularization, whereas ρ_T and ρ_M control deviations from the tumor and cell-line marginal distributions, respectively. Smaller marginal relaxation parameters permit greater deviations from the prescribed masses.

The primary analysis used $\varepsilon = 0.1$ and $\rho_T = \rho_M = 1$. The coupling matrix was computed using generalized Sinkhorn iterations[49], with a maximum of 5,000 iterations and a convergence tolerance of 10^{-8}[50].The effects of transport parameters and distance metrics were further examined in sensitivity analyses.

Tumor-to-cell-line ranking. For each tumor $\boldsymbol{i}$, the optimal transport coupling was normalized across candidate cell lines to obtain tumor-specific matching scores:

$$w_{ij} = \frac{\Gamma_{ij}^{\star}}{\sum_{l=1}^{m} \Gamma_{il}^{\star}}$$

where w_{ij} represents the relative transport mass assigned from tumor i to candidate cell line j. Candidate cell lines were ranked in descending order of w_{ij}, with the highest-ranked cell line designated as the Top1 cell line.

Benchmark analysis using PDMR datasets

Benchmarking known tumor–cell line pairs. The PDMR benchmark was initially assembled from documented tumor–cell line pairs[22]. After restricting the analysis to samples with compatible cancer lineage annotations and evaluable transcriptomic profiles, 177 tumor samples were retained. To ensure a one-to-one evaluation framework, duplicate cell-line counterparts associated with the same PairID were resolved before benchmarking. This procedure yielded 177 unique tumor–cell line pairs for the primary benchmark (Supplementary Table 1; Fig. 2A). For each tumor, the documented cell-line counterpart was ranked among candidate models of the same cancer lineage, providing a reference for evaluating matching accuracy.

Baseline matching approaches. Three transcriptomic representations were considered: standardized gene expression (raw expression), Celligner-aligned expression and scaled PCA. Celligner[17] alignment was reproduced following the published contrastive PCA[51]–mutual nearest neighbor[52] (cPCA–MNN) framework, using the reported parameter settings. Pearson correlation ranked candidate cell lines independently for each tumor, whereas uOT incorporated the joint tumor and cell-line populations into the matching process. All configurations used identical lineage-restricted candidate pools, allowing the contributions of molecular representation and matching strategy to be evaluated separately.

Matching performance metrics. Recovery of documented tumor–cell line relationships was evaluated using Top1 accuracy, Top3 accuracy and mean reciprocal rank (MRR; Fig. 2B, C). Let $\boldsymbol{R_i}$ denote the rank of the documented cell line for tumor $\boldsymbol{i}$, and $\boldsymbol{N}$ the number of evaluable tumors. The indicator function $\mathbb{I}(\mathbf{A})$ equals 1 if condition **A** is satisfied and 0 otherwise. Top1 and Top3 accuracy represent the proportions of tumors whose documented counterparts were ranked first or among the top three candidates, respectively:

$$\text{Top1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(R_i = 1)$$

$$\text{Top3} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(R_i \leq 3)$$

MRR summarizes the positions of documented counterparts across the complete candidate rankings, assigning greater weight to higher-ranked matches:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{R_i}$$

Cumulative recovery curves were generated across increasing rank thresholds, and performance variability was estimated from 1,000 tumor-level bootstrap resamples (Fig. 2D).

Ablation studies. The contributions of scaled PCA and uOT were examined by comparing the complete Tumor2Line framework with three reduced configurations. These included raw expression with uOT (Tumor2Line without scaled PCA) to assess the contribution of scaled PCA,

and scaled PCA with Pearson correlation (Tumor2Line without uOT) to assess the contribution of uOT, and scaled PCA with uOT (Tumor2Line) as the full model. Matching performance was evaluated using Top1 accuracy, Top3 accuracy and MRR on the same lineage-restricted PDMR benchmark (Fig. 2E). Statistical comparisons were performed using exact McNemar tests for Top1 and Top3 accuracy and paired Wilcoxon signed-rank tests for MRR.

Sensitivity and robustness analyses. The robustness of Tumor2Line was examined under alternative feature-selection, dimensionality-reduction and transport settings. We varied the number of highly variable genes, retained PCs, entropic regularization parameter $\boldsymbol{\varepsilon}$, marginal relaxation parameters $\boldsymbol{\rho_T}$ and $\boldsymbol{\rho_M}$, and transport distance metrics (Euclidean, Manhattan and cosine). Matching accuracy was evaluated against the documented PDMR pairs under each configuration (Supplementary Fig. 3).

Sensitivity of PCA-space construction to PDMR lineage coverage. To evaluate the sensitivity of matching performance to PDMR lineage coverage, we compared three PCA strategies: PDMR-only PCA, jointly constructed PCA using PDMR and external reference datasets (Treehouse, TARGET, TCGA and CCLE), and External PCA constructed exclusively from the external reference datasets, with PDMR samples subsequently projected into the fixed reference space. Six cohort sizes (3, 6, 9, 12, 15, and 18 lineages) were evaluated. For cohort sizes below 18, three random lineage orderings were generated, whereas all 18 lineages were evaluated once. The same tumor–cell line pairs and candidate pools were used across all three strategies within each cohort. Within each lineage, uOT matching was performed using the same parameters described above. Performance was assessed using Top1 accuracy, Top3 accuracy, and MRR (Supplementary Fig. 1).

Characterization of PCA scaling. To investigate the influence of PC scaling on tumor–cell line matching, raw and scaled PCA coordinates were compared under identical uOT settings, including the 793-PC solution explaining 90% of the cumulative variance (**Supplementary Fig. 2a, b**). For each retained PC, tumor–cell line source effects were quantified using lineage-adjusted partial R^2, derived from linear models with and without sample source. Significance was assessed using nested-model F test, and associations between source effects and PC explained variance were

evaluated using Spearman correlation (**Supplementary Fig. 2c**). The five PCs with the largest source partial R^2 values were further characterized by Gene Ontology Biological Process[53] over-representation analysis (ORA). For each PC, the top 20% of genes ranked by squared PCA loadings were tested using hypergeometric tests against the PCA loading gene universe, with Benjamini–Hochberg correction applied separately for each PC (**Supplementary Fig. 2d**).

True-pair discrimination across transcriptomic representations. To assess the ability of different transcriptomic representations (raw expression, Celligner-aligned expression and scaled PCA) to distinguish documented tumor–cell line pairs from alternative candidates, we calculated the true-pair similarity advantage for each PDMR tumor. Pearson correlations were computed between each tumor and its lineage-compatible candidate cell lines using standardized raw expression, Celligner-aligned expression and scaled PCA coordinates. The similarity advantage was defined as the correlation with the documented cell-line counterpart minus the highest correlation among all alternative candidates. Positive values indicated that the documented counterpart was more similar to the tumor than any alternative cell line. Differences among representations were evaluated using paired Wilcoxon signed-rank test (Fig. 2F).

Effect of population context on uOT matching. The contribution of tumor population context was investigated using colorectal adenocarcinoma (COAD) and pancreatic adenocarcinoma (PAAD) samples from PDMR. Within each lineage, focal tumors and the complete candidate cell-line pool were held fixed, while increasing numbers of randomly sampled auxiliary tumors were incorporated into the matching population. Tumor2Line was rerun across repeated samplings at each auxiliary tumor count. True-pair rank improvement was defined as the baseline rank without auxiliary tumors minus the mean rank across repeated samplings. Positive values indicated improved matching. Rank improvements were summarized at the focal-tumor level, and their relationship with auxiliary tumor count was assessed using linear trend analysis (Fig. 2G). Linear trend significance was assessed using linear regression, with the slope (β) and corresponding p value reported.

Representational fidelity analysis in TCGA and CCLE

Whole-transcriptome representational fidelity. Whole-transcriptome representational fidelity was assessed by calculating Spearman correlations between each TCGA tumor and its Tumor2Line-selected Top1 CCLE cell line across shared genes. Two reference groups were established within each cancer lineage: random cell lines and other TCGA tumors. For the random reference, each tumor was compared with randomly selected same-lineage CCLE cell lines, whereas the tumor reference was constructed from pairwise expression correlations between distinct tumors of the same lineage. The same correlation method was applied to all three groups to evaluate the molecular similarity of Tumor2Line-selected models relative to random cell lines and within-lineage tumors (Fig. 3B). Statistical comparisons were performed using the unpaired Wilcoxon rank-sum test.

Cancer Hallmark expression-pattern fidelity. Cancer Hallmark representational fidelity was assessed using the same approach, restricting Spearman correlation calculations to genes belonging to each of the 13 cancer Hallmarks[23,24]. For each Hallmark, expression-pattern similarity was calculated between TCGA tumors and their Tumor2Line-selected Top1 cell lines, randomly selected same-lineage cell lines, or other tumors within the same lineage. The analysis was further extended to 50 MSigDB Hallmark gene sets[45] (Fig. 3C; Supplementary Fig. 5A).

Tumor-referenced fidelity index. To account for intrinsic molecular similarity among tumors within the same cancer lineage, we defined the tumor-referenced fidelity index (TRF) as the ratio of tumor–cell line similarity to within-lineage tumor similarity. For each lineage $\boldsymbol{l}$ and Hallmark $\boldsymbol{h}$, TRF was calculated as:

$$\mathrm{TRF}_{lh} = \frac{\mathrm{median}_{i \in l}\left(\rho_{ih}^{\mathrm{Tumor2Line}}\right)}{\mathrm{median}_{i \in l}\left(\rho_{ih}^{\mathrm{WithinLineage}}\right)}$$

where $\rho_{ih}^{\mathrm{Tumor2Line}}$ denotes the Spearman correlation between tumor i and its selected Tumor2Line Top1 cell line across Hallmark h genes, and $\rho_{ih}^{\mathrm{WithinLineage}}$ represents the median correlation between tumor i and other tumors of the same lineage (**Fig. 3D, F; Supplementary Fig. 5**).

Robustness of Hallmark representational fidelity to gene exclusion. To exclude potential bias arising from the overlap between genes used for transcriptomic matching and those used for Hallmark-level fidelity assessment, we repeated the analysis after excluding the union of genes from all 13 Hallmark programs from the matching representation (**Supplementary Fig. 4**). Specifically, we tested Hallmark fidelity by excluding all genes belonging to the 13 cancer Hallmarks from the transcriptomic features used for matching. Scaled PCA and uOT matching were repeated using the remaining genes, while the excluded Hallmark genes were retained for subsequent fidelity evaluation. TRF was recalculated from the newly selected Top1 cell lines and compared with the original lineage–Hallmark TRF estimates using Spearman correlation.

Within-lineage Hallmark heterogeneity. Within-lineage heterogeneity was quantified from the distribution of tumor–tumor Hallmark expression-pattern correlations. For each lineage and Hallmark, the interquartile range (IQR) of tumor-level within-lineage similarity was calculated, and the median IQR across eligible lineages was used to represent Hallmark-level heterogeneity. Its association with Hallmark-level TRF was examined using Spearman correlation (**Fig. 3E**).

Tumor-microenvironment dependence. Tumor-microenvironment (TME) dependence was estimated independently of cell-line matching using TCGA tumor transcriptomes. Immune and stromal scores were calculated using the R package tidyestimate (version 1.1.1)[54]. Within each cancer lineage, Spearman correlations were computed between individual Hallmark gene expression and ESTIMATE[55] immune or stromal scores across tumors. Gene-level TME dependence for each lineage l and Hallmark h was defined as:

$$\mathrm{TME}_{lh} = \max\left\{\frac{1}{|G_{lh}|}\sum_{g\in G_{lh}}\left|\rho_{lg}^{\mathrm{Immune}}\right|,\quad \frac{1}{|G_{lh}|}\sum_{g\in G_{lh}}\left|\rho_{lg}^{\mathrm{Stromal}}\right|\right\}$$

Here, G_{lh} denotes the set of evaluable genes in Hallmark h for lineage l, and $\rho_{lg}^{\mathrm{Immune}}$ and $\rho_{lg}^{\mathrm{Stromal}}$ represent their Spearman correlations with immune and stromal scores, respectively. Genes with non-finite expression or zero variance were excluded. The association between TME dependence and TRF across lineage–Hallmark combinations was evaluated using Spearman correlation (**Fig. 3G, H**).

Clinical pharmacological concordance

Integration of clinical response and drug-sensitivity data. Clinical pharmacological concordance was evaluated by integrating TCGA clinical treatment responses with GDSC cell-line drug-sensitivity measurements. Clinical outcomes were classified as Response (complete or partial response, CR/PR) or Non-response (stable or progressive disease, SD/PD)[43]. Only lineage–drug–assay contexts with at least three tumors in each response group and at least six corresponding cell lines with available drug-sensitivity measurements were retained for downstream analyses. Within each cancer lineage and drug combination, a primary GDSC[8] assay was selected based on clinical-group coverage, the number of evaluable tumor–cell line assignments and the number of tested cell lines. GDSC1 and GDSC2 were treated as separate assays (**Fig. 4A**).

Relative drug-sensitivity ranking. Within each lineage–drug–assay context, cell lines were ranked in ascending order of fitted ln(IC50). The relative drug-sensitivity rank of cell line *j* in context c was defined as:

$$r_{jc} = \frac{q_{jc} - 0.5}{m_c}$$

where q_{jc} denotes the ascending rank of the cell line and m_c denotes the number of cell lines with valid drug-sensitivity measurements in that context. Tied ln(IC50) values were assigned average ranks. Lower relative ranks indicate greater drug sensitivity. Each tumor was assigned the relative rank of its method-specific Top1 matched cell line for the corresponding drug and assay.

Random matching reference and clinical response comparisons. Random matching was used as a reference by assigning tumors to randomly selected same-lineage cell lines with available drug-sensitivity measurements. Random assignments were repeated according to the prespecified resampling procedure, and the resulting relative drug-sensitivity ranks were summarized at the corresponding tumor or context level. Response and non-response relative drug-sensitivity ranks were compared using one-sided unpaired Wilcoxon rank-sum test, with lower ranks expected in the Response group (**Fig. 4B**).

CRISPR-based gene dependency concordance

Integration of organoid and cell-line CRISPR perturbation profiles. The primary analysis included cancer lineages with at least three evaluable organoids and at least three corresponding cell lines with available DepMap GeneEffect profiles[3,42]. Transcriptomic matching was performed independently of CRISPR data using Raw expression + Pearson, Celligner + Pearson, and Tumor2Line, with candidate cell lines restricted to those with available DepMap GeneEffect profiles (**Fig. 4C**).

PDO-specific transcriptomic representation and matching. For PDO-based functional validation, a separate lineage-specific PCA space was constructed by jointly analyzing PDO and corresponding lineage-matched cell-line expression profiles, rather than using the reference PCA space established for primary tumor matching. The candidate cell-line pool was restricted to cell lines with available transcriptomic and DepMap GeneEffect data. Tumor2Line matching was then performed using the same uOT framework described above, and the resulting Top1-matched cell lines were used for downstream CRISPR-based functional concordance analyses.

Functional relative ranking. For each organoid, Spearman correlations were calculated between its MiniLib LFC profile and the GeneEffect profiles of candidate cell lines across all shared genes. Candidate cell lines were ranked in descending order of correlation, with higher correlations indicating greater functional concordance. The functional relative rank (FRR) of the Top1 matched cell line selected by method m for organoid i was defined as:

$$FRR_{i,m} = \frac{(r_{i,m} - 0.5)}{N_i}$$

where $r_{i,m}$ denotes the rank of the matched cell line based on functional concordance, and N_i denotes the number of evaluable candidate cell lines in the corresponding lineage. FRR values were expressed as percentages, with lower values indicating greater functional concordance.

Evaluation of functional concordance. The functional concordance of each matching method was evaluated against a random-matching reference. Empirical p values were calculated by comparing the observed pooled mean functional relative rank with the null distribution generated from 5,000 random sampling iterations using a lower-tail test (**Fig. 4D**). Method comparisons were

performed on the same organoids using one-sided paired signed-rank tests to determine whether Tumor2Line achieved lower FRR values than Raw expression + Pearson and Celligner + Pearson (**Fig. 4E**). Within each lineage, method comparisons were performed using one-sided paired Wilcoxon signed-rank tests, with organoids matched by ID. Observed functional relative ranks were also compared with organoid-specific random-reference ranks using one-sided paired Wilcoxon signed-rank tests. Paired mean FRR differences and their 95% bootstrap confidence intervals were calculated, with negative differences favoring Tumor2Line (**Supplementary Fig. 7**).

Statistical analysis

All analyses were performed in R (version 4.4.3) using RStudio. Data visualization was primarily performed using ggplot2 (version 4.0.0), with patchwork (version 1.3.2) used to assemble multi-panel figures. Tumor microenvironment-associated immune and stromal scores were calculated using tidyestimate (version 1.1.1). Random seeds were specified where required for reproducible resampling and randomized analyses. Statistical analyses included paired and unpaired Wilcoxon tests (rank-sum and signed-rank tests, as appropriate), exact McNemar tests, Spearman correlation, linear regression, bootstrap resampling, and permutation-based empirical significance testing. Multiple-testing correction was performed using the Benjamini–Hochberg or Holm procedure where specified.

Data availability

All transcriptomic, pharmacological and CRISPR dependency datasets analyzed in this study were obtained from publicly available resources, including PDMR, Treehouse, TARGET, TCGA, CCLE/DepMap, GDSC and published patient-derived organoid datasets. The corresponding data sources, accession information and specific files used for analysis are described in the Methods.

Code availability

The Tumor2Line R package and associated analysis code are available on GitHub at

https://github.com/Duan-Lab1/Tumor2Line.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions to our manuscript.

Author contributions

B.D. conceived and supervised the study. H.Z. developed the methodology, performed the analyses. B.D. and H.Z. wrote the manuscript with assistance from other authors. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 61572361)

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information. Supplementary Figures 1–7 and Supplementary Tables 1–9 accompany this manuscript.

Correspondence and requests for materials should be addressed to Bin Duan (binduan@sjtu.edu.cn).

References

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

2 Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).

3 Tsherniak, A. *et al.* Defining a cancer dependency map. *Cell* **170**, 564–576. e516 (2017).

4 Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).

5 Vaske, O. M. *et al.* Comparative tumor RNA sequencing analysis for difficult-to-treat pediatric and young adult patients with cancer. *JAMA network open* **2**, e1913968 (2019).

6 Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).

7 Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).

8 Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41**, D955–D961 (2012).

9 Salvadores, M., Fuster-Tormo, F. & Supek, F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science advances* **6**, eaba1862 (2020).

10 Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature communications* **4**, 2126 (2013).

11 Trastulla, L., Noorbakhsh, J., Vazquez, F., McFarland, J. & Iorio, F. Computational estimation of quality and clinical relevance of cancer cell lines. *Molecular Systems Biology* **18**, MSB202211017 (2022).

12 McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).

13 Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).

14 Yu, K. *et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature communications* **10**, 3574 (2019).

15 Najgebauer, H. *et al.* CELLector: genomics-guided selection of cancer in vitro models. *Cell systems* **10**, 424–432. e426 (2020).

16 Sinha, R., Luna, A., Schultz, N. & Sander, C. A pan-cancer survey of cell line tumor similarity by feature-weighted molecular profiles. *Cell reports methods* **1** (2021).

17 Warren, A. *et al.* Global computational alignment of tumor and cell line transcriptional profiles. *Nature communications* **12**, 22 (2021).

18 Dimitrieva, S. *et al.* Biologically relevant integration of transcriptomics profiles from cancer cell lines, patient-derived xenografts, and clinical tumors using deep learning. *Science Advances* **11**, eadn5596 (2025).

19 Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature medicine* **24**, 541–550 (2018).

20 Schiebinger, G. *et al.* Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943. e922 (2019).

21 Bunne, C. *et al.* Learning single-cell perturbation responses using neural optimal transport. *Nature methods* **20**, 1759–1768 (2023).

22 Sun, H. *et al.* Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidates for targeted treatment. *Nature communications* **12**, 5086 (2021).

23 Sibai, M. *et al.* The spatial landscape of cancer hallmarks reveals patterns of tumor ecological dynamics and drug sensitivity. *Cell Reports* **44** (2025).

24 Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer discov* **12**, 31–46 (2022).

25 Dvorak, H. F. Tumor stroma, tumor blood vessels, and antiangiogenesis therapy. *The Cancer Journal* **21**, 237–243 (2015).

26 van Staveren, W. C. *et al.* Human thyroid tumor cell lines derived from different tumor types present a common dedifferentiated phenotype. *Cancer research* **67**, 8113–8120 (2007).

27 Landa, I. *et al.* Comprehensive genetic characterization of human thyroid cancer cell lines: a validated panel for preclinical studies. *Clinical Cancer Research* **25**, 3141–3151 (2019).

28 Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).

29 Demetci, P., Santorella, R., Chakravarthy, M., Sandstede, B. & Singh, R. Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. *Journal of Computational Biology* **29**, 1213–1228 (2022).

30 Peng, D. *et al.* Evaluating the transcriptional fidelity of cancer models. *Genome medicine* **13**, 73 (2021).

31 Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine* **19**, 1423–1437 (2013).

32 Junttila, M. R. & De Sauvage, F. J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501**, 346–354 (2013).

33 Thorel, L. *et al.* Patient-derived tumor organoids: a new avenue for preclinical research and precision medicine in oncology. *Experimental & Molecular Medicine* **56**, 1531–1551 (2024).

34 ElHarouni, D. *et al.* A compendium of next-generation patient-derived models for diverse cancers. *Nature*, 1–14 (2026).

35 Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods* **21**, 1470–1480 (2024).

36 Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature methods* **21**, 1481–1491 (2024).

37 Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

38 Xing, E. P. & Song, L. A world model of the virtual cell. *Cell* **189**, 5831–5844 (2026).

39 Ma, N. *et al.* Virtual cells: progress and challenges from mechanistic modeling to data-driven integration. *Briefings in Bioinformatics* **27**, bbag562 (2026).

40 Howe, L. J. *et al.* Evaluating transportability of in vitro cellular models to in vivo human phenotypes using gene perturbation data. *Nature Communications* **17**, 513 (2025).

41 Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nature biotechnology* **38**, 675–678 (2020).

42 Herranz-Ors, C. *et al.* A tumour-derived organoid biobank maps cancer gene dependencies. *Nature*, 1–10 (2026).

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

44 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646–674 (2011).

45 Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).

46 Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer* **45**, 228–247 (2009).

47 Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation* **87**, 2563–2609 (2018).

48 Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis* **274**, 3090–3123 (2018).

49 Liero, M., Mielke, A. & Savaré, G. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae* **211**, 969–1117 (2018).

50 Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013).

51 Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications* **9**, 2134 (2018).

52 Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421–427 (2018).

53 The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research* **49**, D325–D334 (2021).

54 Aragaki, K., Yoshihara, K., Vegesna, R., Kim, H. & Verhaak, R. tidyestimate: A Tidy Implementation of 'ESTIMATE'. *R package version 1.1.1* (2023).

55 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**, 2612 (2013).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytables.xlsx
- Supplementaryfigures.docx