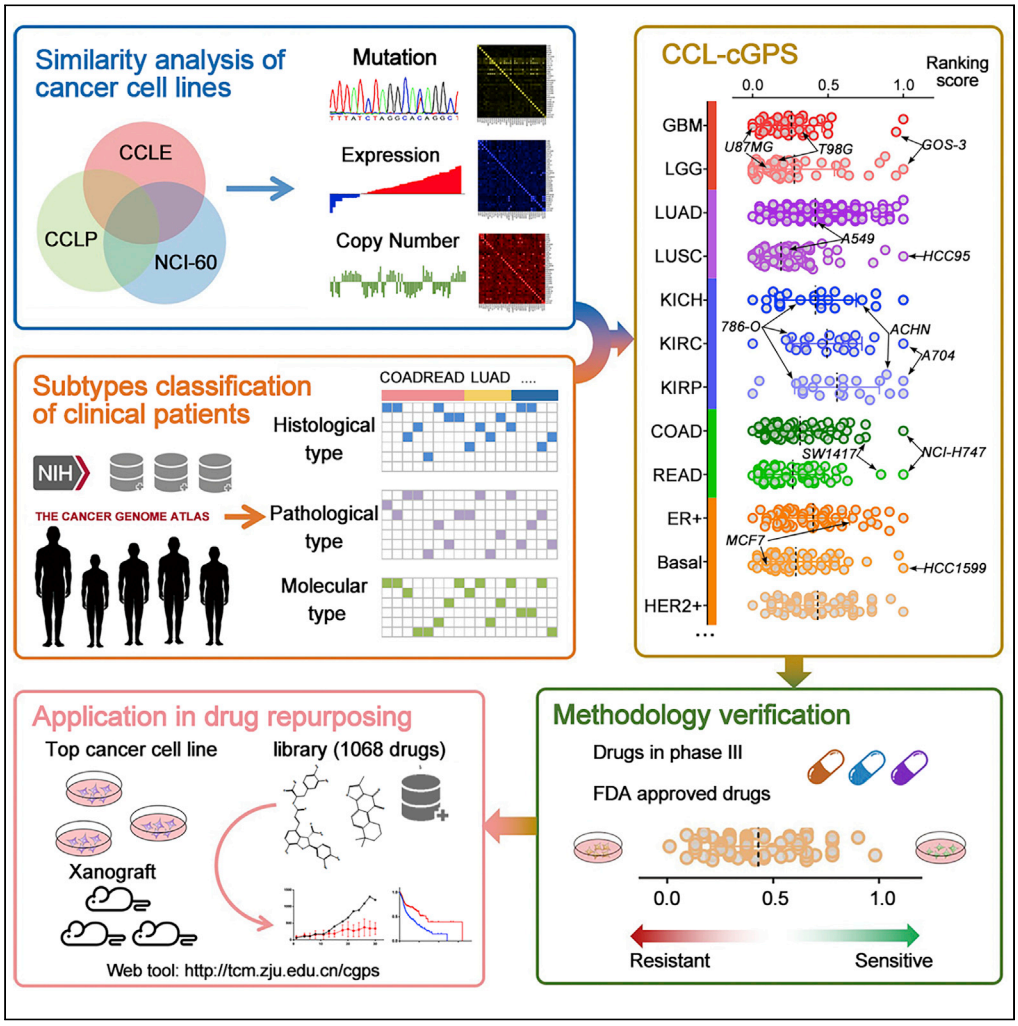
iScience

CellPress
OPEN ACCESS

Article

A Clinical Genomics-Guided Prioritizing Strategy Enables Selecting Proper Cancer Cell Lines for Biomedical Research



Xin Shao, Yi Wang, Xiaoyan Lu, ..., Ni Ai, Meidan Ying, Xiaohui Fan

fanxh@zju.edu.cn

HIGHLIGHTS

Cell lines were ranked by the resemblance of transcriptional signatures to tumors

Among 44 tumor subtypes, CCL-cGPS provides proper cell lines for each subtype

CCL-cGPS was verified by the computational analysis, *in vitro* and *in vivo* assays

A web tool was developed to guide the selection of the most suitable cell lines

Shao et al., iScience *23*, 101748
November 20, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101748

iScience

CellPress
OPEN ACCESS

Article

A Clinical Genomics-Guided Prioritizing Strategy Enables Selecting Proper Cancer Cell Lines for Biomedical Research

Xin Shao,[1,4] Yi Wang,[1,4] Xiaoyan Lu,[1,4] Yang Hu,[1] Jie Liao,[1] Junying Li,[1] Xuechun Chen,[1] Yunru Yu,[1] Ni Ai,[1] Meidan Ying,[2] and Xiaohui Fan[1,3,5,*]

SUMMARY

Selecting appropriate cell lines to represent a disease is crucial for the success of biomedical research, because the usage of less relevant cell lines could deliver misleading results. However, systematic guidance on cell line selection is unavailable. Here we developed a clinical Genomics-guided Prioritizing Strategy for Cancer Cell Lines (CCL-cGPS) and help to guide this process. Statistical analyses revealed CCL-cGPS selected cell lines were among the most appropriate models. Moreover, we observed a linear correlation between the drug response and CCL-cGPS score of cell lines for breast and thyroid cancers. Using RT4 cells selected by CCL-GPS, we identified mebendazole and digitoxin as candidate drugs against bladder cancer and validate their promising anticancer effect through *in vitro* and *in vivo* experiments. Additionally, a web tool was developed. In conclusion, CCL-cGPS bridges the gap between tumors and cell lines, presenting a helpful guide to select the most suitable cell line models.

INTRODUCTION

Immortalized cell lines have been widely employed in the field of biomedical research such as drug discovery and development over the past decades due to their advantages of being readily accessible and easily maintained, compared with primary tissue (Sharma et al., 2010; Yamori, 2003). A wide range of cell lines have been developed as *in vitro* models for the research of various diseases (Allen et al., 2005; Glennon et al., 2019; Li et al., 2019; Qiu et al., 2019). For example, for the study of breast cancer alone, about 50 different cell lines have been reported (Neve et al., 2006), while more than 200 cell lines are available for lung cancer-related studies (Gazdar et al., 2010). However, varying degrees of genomic differences between cell lines and patient specimens have been reported, especially in cancer research (Domcke et al., 2013; Ertel et al., 2006; Gillet et al., 2011; Li et al., 2014; Sandberg and Ernberg, 2005; Sinha et al., 2017; Stein et al., 2004), emphasizing the great importance in appropriate cell line selection (Holliday and Speirs, 2011; Horvath et al., 2016; Ross and Perou, 2001; Sun and Liu, 2015). Hence, a fundamental question arises—how to find a proper cell line with maximal resemblance to its corresponding pathological specimen, to ensure that the subsequent response evaluation gives meaningful results.

Thanks to the fast development of sequencing techniques, several major datasets of cell line or tumor genomics have been made publicly available, such as the NCI-60 project (Ross et al., 2000), the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), the COSMIC Cell Lines Project (CCLP) (Garnett et al., 2012), and the Cancer Genome Atlas (TCGA). Preliminary analysis on these large-scale genomic datasets showed significant discrepancy in the gene activities between the various subtypes of tumors and the commonly used cell lines in the literature of drug discovery and mechanistic studies (Figure S1). Therefore, it is more than necessary to develop a strategy to select cell lines with great biological similarity with tumor specimens. Here we propose a clinical Genomics-guided Prioritizing Strategy for Cancer Cell Lines (CCL-cGPS), to address this issue (workflow see Figure 1).

Interestingly, we found that for over 80% of tumor subtypes under investigation, the most widely used cell lines did not overlap the top selections by CCL-cGPS. We therefore performed statistical assays, including the confirmation analysis and cluster analysis, to validate the CCL-cGPS outcomes. The results suggested that those cell lines selected by CCL-cGPS were among the most appropriate *in vitro* models for most

[1]Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[2]Institute of Pharmacology and Toxicology, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[3]State Key Laboratory of Component-Based Chinese Medicine, Tianjin 301617, China

[4]These authors contributed equally

[5]Lead Contact

*Correspondence: fanxh@zju.edu.cn

https://doi.org/10.1016/j.isci.2020.101748

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

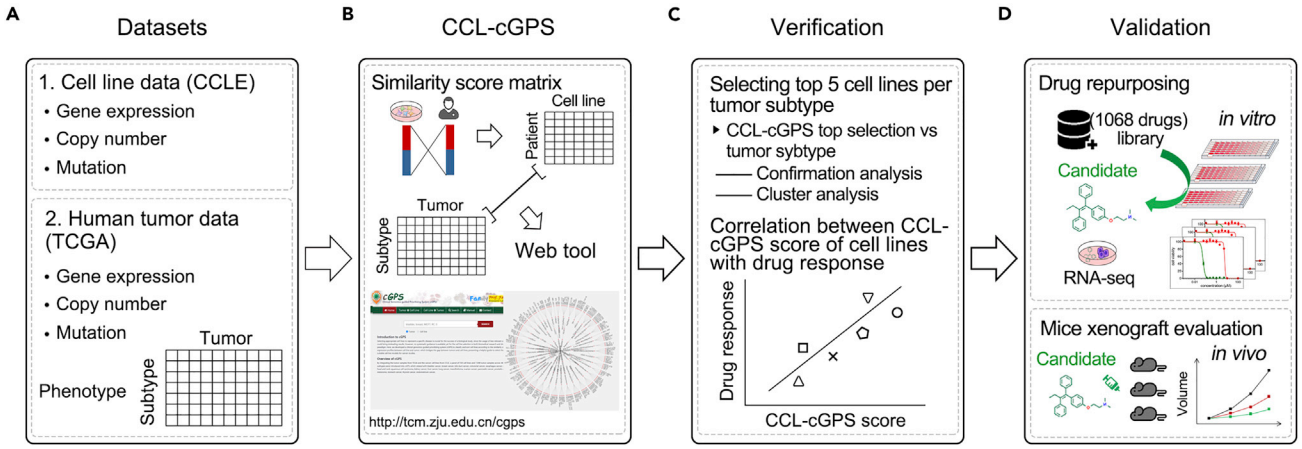


Figure 1. Workflow of CCL-cGPS

(A) Preparation of cell line genomics data from CCLE and tumor genomics data from TCGA including copy number variation, gene mutation, and expression profiles. Tumor samples in each cancer type were classified with respect to three subtype categories, which are histological type, pathological type, and genotype to generate the tumor-subtype matrix.

(B) Clinical Genomics-guided Prioritizing Strategy for Cancer Cell Lines (CCL-cGPS). First, similarity scores were calculated for pairs of patient sample and cell line based on gene expression data to generate the patient-cell line similarity score matrix. Next, the patient-cell line similarity score matrix and the tumor-subtype matrix were merged to generate the cell line-subtype matrix containing the top selections of cell line by CCL-cGPS for each tumor subtype.

(C) Verification of CCL-cGPS top selections of cell line by three means including the confirmation analysis, cluster analysis of cell lines together with tumor patients, and the correlation analysis between the CCL-cGPS scores of cell lines and the drug responses of FDA-approved medications.

(D) Validation of CCL-cGPS top selection from drug repurposing level including the *in vitro* high-throughput screening of 1,068 FDA-approved drugs on recommended cell line by CCL-cGPS and the *in vivo* mice xenograft evaluation of the screened candidate drug compared with the first-line drug.

tumor subtypes. Moreover, the reliability of CCL-cGPS was confirmed by the high correlation between the *in vitro* responses of US Food and Drug Administration (FDA)-approved drugs and the CCL-cGPS ranking scores of cell lines. Last, we further validated the CCL-cGPS dependability through drug repurposing for papillary bladder cancer against which there is a lack of effective medication. As a result, we identified mebendazole and digitoxin as the candidate drugs by screening the FDA-approved drug library with RT4 cell line, which was favored by CCL-cGPS. The subsequent data generated from the *in vitro* cytotoxicity assay and *in vivo* xenograft experiments both corroborated the promising inhibitory effects of both drugs against papillary bladder cancer. Furthermore, a web tool (http://tcm.zju.edu.cn/cgps) was developed to facilitate the usage of CCL-cGPS in the broader scientific community, allowing users to browse the CCL-cGPS selected cell lines across 44 tumor subtypes. Our CCL-cGPS serves as a helpful guide for investigators to determine the most suitable cell line model for *in vitro* biomedical studies. The results of our findings may advance the understanding of the relation between cell lines and clinical specimens by bridging the gap between them, and help increase the success rate in biomedical research.

RESULTS

Similarity Ranking of Cell Lines in CCL-cGPS

In the current study, we analyzed the genomic data of cell lines obtained from multiple resources (Cerami et al., 2012), including NCI-60, CCLE, and CCLP, and found a significant correlation of these data between the datasets (Figures S2A and S2B). Tumor samples were classified into histopathological and molecular subtypes according to their phenotypic and genotypic information, which are closely associated with their distinct treatment response. As a result, 720 cell lines (Table S1) and 7,101 tumor samples across 44 tumor subtypes (Table S2) were introduced into CCL-cGPS. Briefly, the transcriptional signature was aligned to track down the candidate cell lines closely resembling each tumor sample. The best matched cell line for a specific tumor subtype was then determined after considering the tissue of origin and the normalized ranking of these candidate cell lines. The CCL-cGPS ranking score is mainly dependent on two parameters, the number of selected gene signatures and the number of cell lines used for normalization. A systematic investigation showed that the CCL-cGPS ranking scores exhibited high levels of stability and consistency (Figure S3A), and particularly so when 100 differentially expressed genes (DEGs) were examined and the top five best matched cell lines were adopted (Figure S3B).

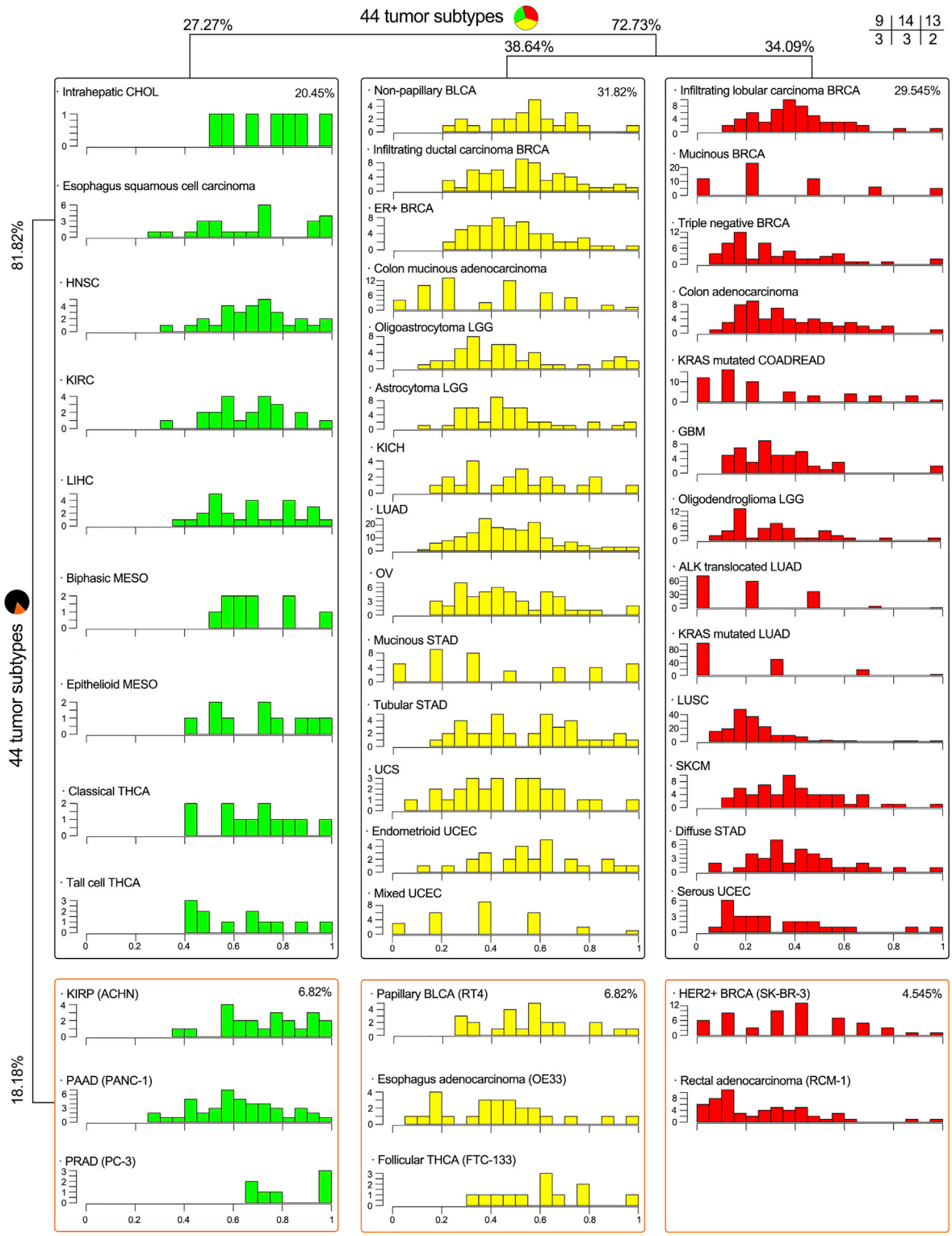

44 tumor subtypes
27.27%
72.73%
38.64%
34.09%
9 | 14 | 13
3 | 3 | 2
81.82%
18.18%
44 tumor subtypes
Intrahepatic CHOL
20.45%
Esophagus squamous cell carcinoma
HNSC
KIRC
LIHC
Biphasic MESO
Epithelioid MESO
Classical THCA
Tall cell THCA
Non-papillary BLCA
31.82%
Infiltrating ductal carcinoma BRCA
ER+ BRCA
Colon mucinous adenocarcinoma
Oligoastrocytoma LGG
Astrocytoma LGG
KICH
LUAD
OV
Mucinous STAD
Tubular STAD
UCS
Endometrioid UCEC
Mixed UCEC
Infiltrating lobular carcinoma BRCA
29.545%
Mucinous BRCA
Triple negative BRCA
Colon adenocarcinoma
KRAS mutated COADREAD
GBM
Oligodendroglioma LGG
ALK translocated LUAD
KRAS mutated LUAD
LUSC
SKCM
Diffuse STAD
Serous UCEC
KIRP (ACHN)
6.82%
PAAD (PANC-1)
PRAD (PC-3)
Papillary BLCA (RT4)
6.82%
Esophagus adenocarcinoma (OE33)
Follicular THCA (FTC-133)
HER2+ BRCA (SK-BR-3)
4.545%
Rectal adenocarcinoma (RCM-1)

Figure 2. Distribution of CCL-cGPS Ranking Scores for Cell Lines Across 44 Tumor Subtypes.

y Axis: frequency; x axis: CCL-cGPS ranking score. Green: most cell lines with high CCL-cGPS ranking scores are similar to tumors. Yellow: CCL-cGPS ranking scores for cell lines are distributed evenly. Red: most cell lines with low CCL-cGPS ranking scores are different from tumors. The black or orange border represents the discordant or concordant cases in which the commonly used cell line falls outside or within the selected cell lines by CCL-cGPS, respectively. The commonly used cell line was labeled beside the subtype in the concordant cases.

According to the distribution pattern of cell line ranking scores, all tumor subtypes were divided into three groups (Figure 2). In two of the three distribution pattern groups, most cell lines exhibited low resemblance to patient tumor samples (yellow and red ones in Figure 2), including those of bladder, breast, colorectal, glioma, lung, ovarian, melanoma, stomach, and endometrium cancers, which might be a result of high heterogeneity or imperfect establishment of those cell lines.in these cancers. Therefore, one should be particularly cautious in picking the right cell line for biomedical research of these tumor types. In the third situation, most cell lines are transcriptionally similar to tumor specimens (green ones in Figure 2), including those of bile duct, head and neck squamous cell, liver cancer, mesothelioma (MESO), and pancreatic and prostatic cancers, wherein one of the reasons might be that these cancers are not highly heterogeneous compared with the cancers with so much well-appreciated diversity, e.g., breast cancer. To study tumors in this category, there could be many options of cell lines. Intriguingly, the CCL-cGPS found that esophagus cancer cell lines are more similar to esophagus adenocarcinoma than to esophagus squamous cell carcinoma, whereas thyroid cancer cell lines presented high resemblance to classical or tall cell thyroid carcinoma (THCA). Moreover, a bulk of kidney cancer cell lines showed higher similarity with renal clear cell carcinoma (KIRC) and papillary renal cell carcinoma (KIRP), than chromophobe renal cell carcinoma (KICH), which was in accordance with previous copy number variation (CNV)-based data (Sinha et al., 2017).

The goal of CCL-cGPS is to identify the cell lines best representing each tumor subtype. To overcome the bias introduced by the sample number variation between subtypes, a permutation test was conducted wherein significantly enriched cell lines possessing top five ranking scores were regarded as the CCL-cGPS top selections (Table S3). Concordantly, CCL-cGPS selected cell lines contained a high percentage of DEGs (median, 78.37% across 169 CCL-cGPS top selected cell lines) which were conserved with the corresponding tumor patients (Table S4). For example, OVSAHO selected for ovarian cancer includes 1,110 DEGs, 1,026 of which are conserved in patients with ovarian cancer, while the most commonly used PC-3, also selected by CCL-cGPS, contains 830 out of 913 conserved DEGs (90.81%) for prostatic cancer. Same concordance is observed in conserved pathways between the highly ranked cell lines and associated tumor samples, e.g., ER^+ breast cancer (23 common pathways, e.g., cell cycle, DNA repair, autophagy, transcriptional regulation by *TP53*, regulation of cell cycle process).

Next, we compared the CCL-cGPS selected cell lines with the most cited ones in previous literature, as retrieved from PubMed, for 18 cancers (Tables 1 and S1). Our results indicated that only eight commonly used cell lines, namely, bladder cancer cell line RT4, breast cancer cell line MCF7, colorectal cancer cell line RCM-1, esophagus cancer cell line OE33, kidney cancer cell line ACHN, pancreatic cancer cell line PANC-1, prostatic cancer cell line PC-3, and thyroid cancer cell line FTC-133, overlapped the CCL-cGPS selections, pertaining to eight tumor subtypes, namely, papillary bladder cancer, HER2-positive breast cancer, rectal adenocarcinoma (READ), esophagus adenocarcinoma, KIRP, pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), and follicular THCA. These results highlighted a discordance rate of 81.82% between the widely cited cell lines and those preferred by CCL-cGPS (Figure 2). Furthermore, among the cell lines with the even distribution of ranking scores across 17 subtypes, we found only three cases wherein the most commonly used cell line for a tumor subtype was within those selected by CCL-cGPS. In comparison, two cases were observed for cell lines with mostly low ranking scores (Figure 2). Taken together, these results warranted more consideration on cell lines selection for *in vitro* studies of these tumor subtypes. For example, MCF7 is the most commonly used cell line for ER^+ breast cancer in biological research. However, CCL-cGPS showed a low-ranking score of MCF7 in representing ER^+ breast invasive carcinoma (BRCA), contrary to the popular assumption (Formisano et al., 2019; Hinohara et al., 2018; Li et al., 2018). The relatively low resemblance of MCF7 to ER^+ BRCA could be attributed to the ongoing genotypic and phenotypic evolution during the continual culture (Holliday and Speirs, 2011; Wang et al., 2006), which is not uncommon to many cell lines.

Verification of CCL-cGPS Top Selection of Cell Lines

Next, the accuracy of CCL-cGPS selection was verified by the confirmation analysis and the cluster analysis, followed by the correlation assay performed on the CCL-cGPS ranking scores of the cell lines and their drug

Tumor Type	TCGA Label	Tumor Subtype	Commonly Used Cell Lines	Selected by CCL-cGPS	Biological Traits	Hit Rate
Bladder	BLCA	P	RT4	√	–	100%
Bladder	BLCA	NP	T24	–	–	60%
Breast	BRCA	ID	MCF7	–	80% Ductal	80%
Breast	BRCA	IL	MDA-MB-134-VI	–	–	100%
Breast	BRCA	Mucinous	–	–	–	–
Breast	BRCA	ER^+	MCF7	–	60% ER^+	60%
Breast	BRCA	$HER2^+$	SK-BR-3	√	100% $HER2^+$	0%
Breast	BRCA	TN	MDA-MB-231	–	60% TN	0%
Bile duct	CHOL	Intrahepatic	HuCCT1	–	–	–
Colorectal	COADREAD	COAD	HT-29	–	100% AD	80%
Colorectal	COADREAD	COMAD	–	–	67% AD	67%
Colorectal	COADREAD	READ	RCM-1	√	100% AD	60%
Colorectal	COADREAD	$KRAS^+$	HCT116	–	80% $KRAS^+$	–
Esophagus	ESCA	AD	OE33	√	25% AD	50%
Esophagus	ESCA	SCC	TE-1	–	100% SCC	100%
Glioma	GBM	GBM	U-87_MG	–	–	–
Glioma	LGG	A	U-87_MG	–	80% A	20%
Glioma	LGG	OA	–	–	–	40%
Glioma	LGG	OD	Hs_683	–	0% OD	40%
HNSC	HNSC	HNSC	FaDu	–	100% SCC	–
Kidney	KICH	KICH	–	–	–	25%
Kidney	KIRC	KIRC	786-O	–	40% KIRC	40%
Kidney	KIRP	KIRP	ACHN	√	–	20%
Liver	LIHC	LIHC	Hep_G2	–	80% LIHC	–
Lung	LUAD	LUAD	A549	–	0% AD	80%
Lung	LUAD	$KRAS^+$	A549	–	0% KRAS+	60%
Lung	LUAD	ALK^+	A549	–	–	–
Lung	LUSC	LUSC	SK-MES-1	–	60% SCC	–
Mesothelioma	MESO	Epithelioid	–	–	–	100%
		Biphasic	MSTO-211H	–	–	100%
Ovarian	OV	OV	SK-OV-3	–	–	–
Pancreas	PAAD	PAAD	PANC-1	√	–	–
Prostate	PRAD	PRAD	PC-3	√	50% AD	–
Melanoma	SKCM	SKCM	A-375	–	–	–

Table 1. Verification of CCL-cGPS Top Selection of Cell Lines

(Continued on next page)

Tumor Type	TCGA Label	Tumor Subtype	Commonly Used Cell Lines	Selected by CCL-cGPS	Biological Traits	Hit Rate
Stomach	STAD	Diffuse	MKN-45	–	0% Diffuse	0%
Stomach	STAD	Mucinous	–	–	–	100%
Stomach	STAD	Tubular	MKN74	–	0% Tubular	75%
Thyroid	THCA	Classical	–	–	–	75%
Thyroid	THCA	Follicular	FTC-133	√	33% Follicular	33%
Thyroid	THCA	Tall cell	–	–	–	100%
Endometrium	UCEC	Endometrioid	RL95-2	–	–	0%
Endometrium	UCEC	Serous	–	–	–	0%
Endometrium	UCEC	Mixed	–	–	–	100%
Endometrium	UCS	UCS	ESS-1	–	–	–

Table 1. Continued

Biological characteristics are the histopathological or molecular traits of selected cell lines by CCL-cGPS for each tumor subtype. Hit rate is the percentage of selected cell lines clustered with the corresponding tumor patients by integrated analysis of cell lines and tumor patients for each tumor subtype. - means not available. √ means the commonly used cell line is also selected by CCL-cGPS. P: papillary. NP: non-papillary. TN: triple negative. ID: infiltrating ductal. IL: infiltrating lobular. COMAD: colon mucinous adenocarcinoma. AD: adenocarcinoma. SCC: squamous cell carcinoma. A: astrocytoma. OA: oligoastrocytoma. OD: oligodendroglioma.

responses. Theoretically, the histopathological or molecular traits of CCL-cGPS selected cell lines should be consistent with those of tumor samples for each tumor subtype. We focused on 22 subtypes with known histopathological or molecular profiles of both cell lines and tumor samples (Table S4). As a result, for 17 of these subtypes, each had at least one CCL-cGPS favored cell line exhibiting the same traits as its corresponding tumor (Table 1). For example, for ER^+ breast tumor, three out five cell lines selected by CCL-cGPS (HCC1428, ZR-75-30, and MDA-MB-134-VI) highly expressed *estrogen receptor 1* (*ESR1*) gene. In the case of $HER2^+$ breast tumor, the CCL-cGPS selected SK-BR-3 and HCC202 cell lines both exhibited upregulated expression of *erb-b2 receptor tyrosine kinase 2* (*ERBB2*), whereas the gene expressions of *ESR1*, *ERBB2*, and *progesterone receptor* (*PGR*) were uniformly downregulated in 60% of the CCL-cGPS preferred cell lines (HCC1599, CAL-85-1, and HCC1143) for triple-negative breast tumor. Concordantly, four of five CCL-cGPS selections for infiltrating ductal breast tumor (namely, HCC2218, HCC2157, HCC1599, and ZR-75-30) indeed originated from ductal carcinoma.

Adenocarcinoma and squamous cell carcinoma are two common histological tumor subtypes present in esophagus cancer (Du et al., 2019), as well as in colorectal, lung, pancreas, and prostate tumors. Notably, for adenocarcinoma, such as colon adenocarcinoma (COAD) and READ, 100% CCL-cGPS selected cell lines originated from adenocarcinoma. Besides, the top two selected cell lines for PRAD, i.e., VCaP and PC-3, were both derived from adenocarcinoma. For squamous cell carcinoma such as esophagus squamous cell carcinoma, head and neck squamous cell carcinoma (HNSC), and lung squamous cell carcinoma (LUSC), 100%, 100%, and 60% of the CCL-cGPS recommended cell lines stemmed from squamous cell carcinoma, respectively. Meanwhile, *KRAS* mutation was observed for 80% of the selected cell lines for *KRAS*-mutated colorectal tumor. However, in some tumor subtypes, the CCL-cGPS selected cell lines exhibited entirely different traits compared with the tumors, including oligodendroglioma of brain lower-grade glioma (LGG), lung adenocarcinoma (LUAD), *KRAS*-mutated LUAD, and diffuse and tubular stomach adenocarcinoma (STAD). This observation might not be surprising, as some cell lines were likely to develop mutated biological traits as previously described. For instance, the broadly cited kidney cell line ACHN, originated from KIRC, appeared to be a poorly differentiated carcinoma with a predominantly sarcomatoid pattern, as evidenced by the H&E staining in ACHN xenograft mice (Sinha et al., 2017), indicating ACHN preferably resembles KIRP rather than KIRC.

Subsequently, the similarity of transcriptional profiles between cell lines and tumor samples was analyzed by principal-component analysis dimensionality reduction and k-means clustering, two widely used

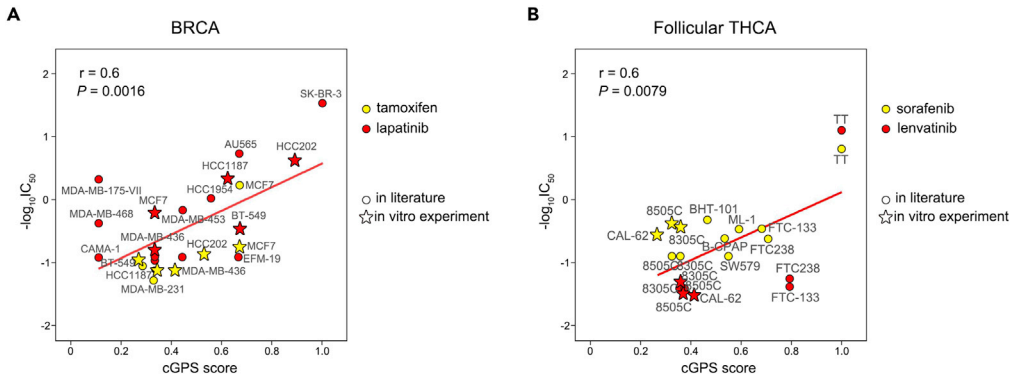


Figure 3. Correlation between CCL-cGPS Ranking Score and Drug Response

(A) IC_{50} of FDA-approved drugs tamoxifen against ER^+ BRCA and lapatinib against $HER2^+$ BRCA in breast cancer cell lines; x axis: CCL-cGPS ranking scores of cell lines for ER^+ or $HER2^+$ BRCA.

(B) IC_{50} of FDA-approved drugs sorafenib and lenvatinib against differentiated thyroid cancer in thyroid cancer cell lines; x axis: CCL-cGPS ranking scores of cell lines for follicular THCA.

approaches to explore the relationship among individuals with high-dimensional data. Through the statistical analysis, we sought to verify the reliability of CCL-cGPS selection. Tumors differentiated into at least two subtypes were included for cluster analysis except for the mucinous breast cancer, which had no cell line assigned by CCL-cGPS. Ideally, selected cell lines by CCL-cGPS would be predominantly clustered with tumor samples of the same subtype. Among the 31 included tumor subtypes (Table S4), at least one selected cell line in each of the 26 subtypes was hit with the cluster of the corresponding tumor samples (Table 1 and Figure S4). Concordantly, the hit rate of matched cell lines selected by CCL-cGPS reach 100% for eight tumor subtypes, including papillary bladder urothelial carcinoma (BLCA), infiltrating lobular BRCA, esophagus squamous cell carcinoma, epithelioid and biphasic MESO, mucinous STAD, tall cell THCA, and mixed uterine corpus endometrial carcinoma (UCEC). Nevertheless, there existed some deviant cases wherein the selected cell lines were hardly clustered with the tumors, such as $HER2^+$ and triple-negative BRCA, diffuse STAD, and endometrioid and serous UCEC.

Furthermore, we tested the *in vitro* inhibitory activities of FDA-approved drugs against the cell lines selected by CCL-cGPS. It is expected that the cell lines with high ranking scores would be more sensitive to the drugs than those with lower ones. With a focus on cell lines showing certain dissimilarity with tumors, we selected three tumor subtypes, i.e., ER^+ BRCA in discordant case, follicular THCA with an even distribution, and $HER2^+$ BRCA with mostly low ranking scores. The IC_{50} values of tamoxifen for ER^+ BRCA and lapatinib for $HER2^+$ BRCA in breast cancer cell lines, and those of lenvatinib and sorafenib for differentiated thyroid cancer (DTC) in thyroid cancer cell lines, were retrieved from PubMed, PubChem, and CCLE. Meanwhile, cell viability assays on breast and thyroid cancer cell lines were conducted to determine the IC_{50} values of the drugs (Table S5). As expected, the cell lines with higher CCL-cGPS ranking scores were associated with lower IC_{50} values of the drugs (sensitive), and vice versa (resistant). Moreover, a linear correlation was observed between the drug responses and the CCL-cGPS ranking scores of cell lines (Figure 3). In addition, our experimental data of drug response were in line with those from previous reports involving cell lines MDA-MB-436, 8305C, 8505C, etc. (Barretina et al., 2012; Tohyama et al., 2014)

Notably, CCL-cGPS selected cell lines SK-BR-3 and HCC202 for $HER2^+$ BRCA, which were perfectly concordant $HER2^+$ biological traits, also showed high sensitivity to FDA-approved drug lapatinib against $HER2^+$ BRCA (Figure 3A). Interestingly, despite being derived from medullary thyroid carcinoma (MTC), the CCL-cGPS favored cell line TT for follicular THCA was concordantly clustered with follicular thyroid tumor samples. Moreover, across all thyroid cell lines, only TT was sensitive to lenvatinib and sorafenib, the approved drugs against DTC (Figure 3B). It is noteworthy that TT also exhibited high sensitivity to cabozantinib and vandetanib (IC_{50} = 0.04 and 0.47 μM, respectively, Verbeek et al., 2011; IC_{50} = 0.048 μM and 0.7 nM, respectively, Mologni et al., 2013), two more anticancer drugs approved for treating MTC. These results corroborated that the commonly used cell line TT is an appropriate *in vitro* model for DTC- or MTC-related study.

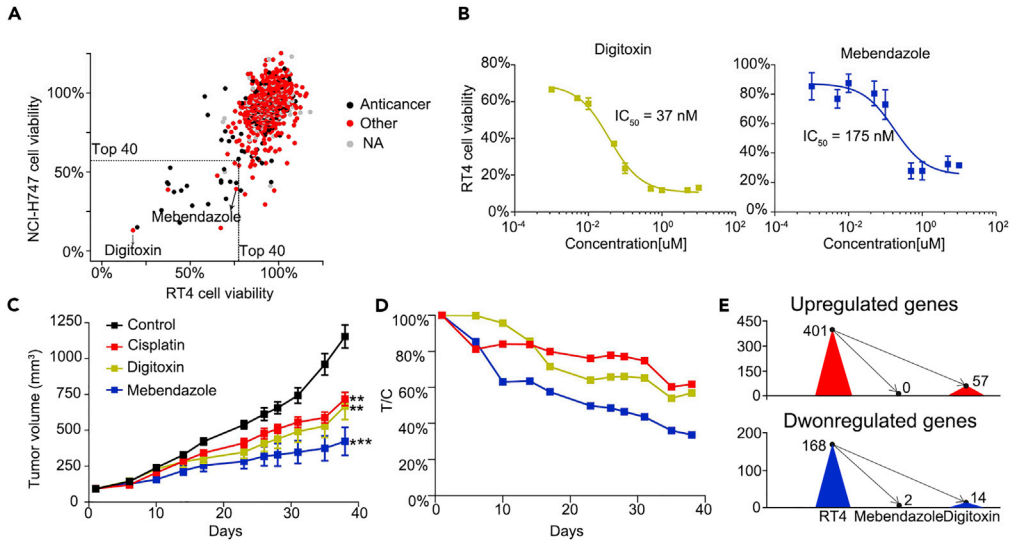


Figure 4. Application of CCL-cGPS on Drug Repurposing *In Vitro* and *In Vivo*

(A) High-throughput screening of 1,068 FDA-approved drugs in bladder cancer cell line RT4 and colorectal cancer cell line NCI-H747. Other: non-anticancer agents. NA: not available.

(B) Dose-response curve of digitoxin and mebendazole on RT4. IC_{50} values were calculated by non-linear simulation (data represent mean ± SEM of three experiments performed in triplicates).

(C) Tumor volume of RT4 xenograft mice treated with saline, cisplatin (2 mg/kg every three day, i.p.), digitoxin (0.5 mg/kg/day, i.p.), and mebendazole (50 mg/kg every other day, i.g.) for consecutive 38 days.

(D) The relative tumor growth rate of mice treated with cisplatin, digitoxin, and mebendazole.

(E) The number of upregulated and downregulated genes from RT4 after mebendazole and digitoxin. **p < 0.0021; ***p < 0.0002.

Validation on CCL-cGPS Top Selection

There is a lack of effective drug for treating bladder and colorectal cancers. To overcome this, we used the RT4 cell line selected by CCL-cGPS for papillary BLCA, and NCI-H747 cells for COAD and READ, to screen potential anticancer agents from an in-house drug library consisting of 1,068 FDA-approved drugs. High-throughput screening was performed with RT4 and NCI-H747 at the concentration of 1 μM. Unsurprisingly, the majority of non-anticancer drugs showed weak inhibitory effects against RT4 and NCI-H747, whereas the cell viability was significantly reduced by the popular anticancer drugs, such as bortezomib, vincristine, and paclitaxel (Figure 4A). From the resultant top 40 inhibitors, we selected the first and the last non-anticancer drugs for further validation, namely, digitoxin commonly used as cardiovascular medication and mebendazole treating parasitic infestations. As a result, strong inhibitory effects of digitoxin (IC_{50} = 37 nM) and mebendazole (IC_{50} = 175 nM) in RT4 cells were unprecedentedly observed (Figure 4B).

To further confirm the *in vivo* efficacy of digitoxin and mebendazole in treating papillary bladder tumor, RT4 cell-derived xenograft tumors were established in BALB/c nude mice, which were then administered with vehicle control (0.9% sodium chloride, 0.1 mL/10 g/day, intraperitoneally [i.p.]), cisplatin (2 mg/kg every three day, i.p.), digitoxin (0.5 mg/kg/day, i.p.), or mebendazole (50 mg/kg every other day, intragastrically [i.g.]) until the tumor volume reached about 100 mm^3 7 days after tumor inoculation. Compared with the control group, the drug-treated mice achieved a significant reduction of tumor volume (Figure 4C). Remarkably, on the 38th day of the treatment, both digitoxin and mebendazole exhibited stronger inhibitory effects on tumor growth than cisplatin, the first-line drug for bladder cancer. At the same time, the relative tumor growth rate of mice treated with mebendazole reached 33.7% (Figure 4D), which was significantly lower than that with cisplatin. In aggregate, these results indicated that mebendazole is a potentially promising therapeutic medication combating papillary bladder cancer.

Finally, we further examined the regulation of mebendazole and digitoxin on the transcriptional level of DEGs in RT4 cells. A total of 401 highly expressed and 168 lowly expressed genes were initially observed in RT4 cells obtained from the CCLE resource. These two numbers were sharply decreased to 0 and 2 upon the treatment of mebendazole, and to 57 and 14 after the digitoxin administration, respectively (Figure 4E).

In other words, mebendazole and digitoxin regulated 99.6% and 87.5% of the 569 DEGs in RT4 cells, respectively. The remarkable perturbation capability of both drugs on the gene expression accorded with their strong inhibition effects on the growth of RT4 cell line detected both *in vitro* and *in vivo*.

Web Tool of CCL-cGPS

To facilitate the application of CCL-cGPS, a user-friendly web tool was developed (http://tcm.zju.edu.cn/cgps). The website presents the overview of CCL-cGPS in "Home" page for users to browse all kinds of tumors, related TCGA labels, tumor subtypes, and CCL-cGPS top selected cell lines (Figure S6A). Primarily, the CCL-cGPS website allows users to query the database through two aspects. Users are allowed to choose the tumor subtype of interest to view the CCL-cGPS selected cell lines for the searched tumor subtype in "Tumor to Cell" page (Figure S6B). On the other hand, users are able to choose the cell line of interest by tissue type from the hierarchical tree to retrieve the appropriate tumor subtypes in "Cell to Tumor" page (Figure S6C), in which the searched cell line is selected by CCL-cGPS.

DISCUSSION

PAM50, known as prediction analysis of microarray 50, was proposed by analyzing the 189 patients with breast cancer (Parker et al., 2009). However, those signatures are chiefly based on breast cancers rather than pan cancer. Therefore, we developed CCL-cGPS as a supplementary approach to better help guide the selection of cell line models that mimic clinical patients. Compared with PAM50, patient-derived signatures used in CCL-cGPS are from a more comprehensive and high-quality genome atlas of tumor patients with next-generation sequencing. We not only verified the CCL-cGPS selected cell lines by confirmation analysis, cluster assay, and correlation assessment between the CCL-cGPS scores of cell lines and the drug responses of FDA-approved medications but also demonstrated the application of CCL-cGPS in at least one area, drug repurposing. In addition, we provided a web tool to browse the cell lines favored by CCL-cGPS for each tumor subtype, providing new insights for drug discovery and mechanistic research. Besides, we performed a systematic investigation on varying the number of patient-derived signatures (100, 200, and 300) to explore the impact of different number of patient-derived signatures on the selection of cell lines.

Cell line-based screening is a fundamental step in biomedical research such as drug discovery, and the right choice of cell line is indispensable for the success of a biological study. However, for many tumor subtypes, too many options of cell lines exist today, making it unrealistic to use all the cells for disease. Currently, many researchers simply pick a commonly used cell line or a readily available one to conduct the experiment with. However, of the 18 primary cancers surveyed, we found only five commonly used cell lines (namely, esophagus cancer cell line TE-1, kidney cancer cell line ACHN, pancreatic cancer cell line PANC-1, prostatic cancer cell line PC-3, and thyroid cancer cell line TT) resembling the corresponding tumors, pertaining to merely 15.9% of the 44 tumor subtypes. The bulk of commonly used cell lines for each tumor subtype hardly mimic the pathological specimens, which indicates that in most cases the frequently used cell lines may not be the best choice.

Surprisingly, the odds seem better to randomly select an available cell line than choosing the commonly used one, because, for 27.27% of the tumor subtypes, most cell lines exhibit low similarity with the tumors, including CHOL, HNSC, liver hepatocellular carcinoma (LIHC), MESO, PAAD, and PRAD. Nevertheless, in the rest 72.73% of the tumor subtypes, most cell lines do not show good resemblance to the tumors; hence in the majority of cases, random selection may lead to a wrong cell line usage. Other than three tumor subtypes for which the commonly used cell lines overlap the CCL-cGPS selections and another 12 subtypes having concordant cell lines, 65.9% of tumor subtypes have mostly dissimilar cell lines, namely, BLCA, BRCA, COADREAD, GBM, LGG, LUAD, LUSC, OV, skin cutaneous melanoma (SKCM), STAD, UCS, and UCEC. For these subtypes, researchers must be extremely careful about choosing the right cell line.

Drug response varies among cell lines because of cell heterogeneity. Increasing evidence indicated the existence of genomic differences between cell line and tumor, although immortalized cell lines are originally sourced from a tumor. As expected, we found that genomic variation events in tumors are not well preserved across the cell lines originated from the same primary site (Figure S2C). Only a few cell lines transcriptionally correlated with the tumors across 22 cancers, such as BRCA, LUAD, and STAD. These findings may not be surprising given that cancer cell lines hardly expressed any tissue- or tumor-specific genes at transcriptional level upon group-wise comparison (Sandberg and Ernberg, 2005). Therefore, we explored

the relationship between cell line and tumor by molecular profile comparisons, and proposed CCL-cGPS to ensure the selected cell lines truthfully reflects the tumor activities at molecular level.

Although significantly high similarities of the genomic data, including CNV, gene expression, and mutation, were observed among the cell lines shared by NCI-60, CCLE, and CCLP, it is noteworthy that gene expression profiles exhibited substantially higher correlation than CNV and mutation between CCLE and NCI-60, with the Pearson's correlation coefficients of matched cell lines all well within the 5% quantile (Figure S2A). Meanwhile, the DEGs' association with common cell lines between CCLE and CCLP surpassed those of copy number variant genes or mutated genes (Figure S2B), indicating that gene expression profiles possessed more concordant characteristics of cell lines. Besides, high concordance between DEGs and CNVs has been extensively documented across cancer cell lines and patients with tumor (Shao et al., 2019). Moreover, it is messenger RNA, rather than upstream genetic CNV or DNA mutation, that acts as a vital mediator affecting the phenotype and drug response (Lee et al., 2018). Hence, gene expression, but not CNV or mutation, was introduced into CCL-cGPS to explore the potential relationship between cell lines and tumors at the molecular level.

Interestingly, no proper cell line currently exists for the study of mucinous BRCA. Similarly, we observed a huge genomic discrepancy between cell lines and liquid cancers, including acute myeloid leukemia (LAML) and diffuse large B-cell lymphoma (DLBC) (Figure S5), and therefore excluded them from CCL-cGPS. However, cell lines derived from solid tumors were distributed indistinguishably as corresponding tumor samples, indirectly corroborating that solid cancer could be mimicked by cell lines. Our CCL-cGPS helps identify the proper cell line best resembling the tumors. To date, a total of 44 tumor subtypes of solid cancers (Table S2) involving 7,101 tumor samples and a collection of 720 related cell lines have been introduced into CCL-cGPS.

This work is the first systematic pan-cancer study on how to select the right cell lines to best represent tumors for biomedical research to date. Selected cell lines by CCL-cGPS showed high sensitivity to FDA-approved drugs, which will remarkably increase, at the early stage, the success rate for drug discovery and development. Despite some discrepancy in biological traits between cell lines and tumors, most of the selected cell lines by CCL-cGPS retained consistent histopathological or molecular profiles of the corresponding tumors, which could be utilized for the mechanistic study. In addition, we demonstrated the application of top selected cell lines by CCL-cGPS in drug repurposing with *in vitro* and *in vivo* experiments and identified potential agents for treating papillary bladder cancer. Fundamentally, CCL-cGPS is a gene expression-prioritized strategy of cell lines selection for biomedical research, which could therefore be utilized to study a wide scope of diseases, such as cardiovascular, liver, and brain dysfunctions, upon the expansion of genomic data of cell lines and pathological specimens.

Limitations of the Study

As revealed in our study, CCL-cGPS can significantly reduce the gap between tumors and cancer cell lines by selecting a proper cell line based on similarity of transcriptomic profiling. However, varying degrees of genomic differences between cell lines and patient specimens have been reported. From a personalized medicine point of view, efforts still need to be made to effectively bridge the gap between patients and cell lines from other aspects.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Xiaohui Fan (fanxh@zju.edu.cn).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The CNV, mRNA expression, and mutation profiles from CCLE, NCI-60 and TCGA were collected from the cBioportal for Cancer Genomics (https://www.cbioportal.org/). Genomic data of CCLP was retrieved from COSMIC Cell Lines Project (v81, https://cancer.sanger.ac.uk/cell_lines). The web of CCL-cGPS is available

at http://tcm.zju.edu.cn/cgps and RNA-seq data of RT4 and source codes of data processing with R are available at github (https://github.com/ZJUFanLab/CCL-cGPS).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101748.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors thank Dr. Weida Tong (National Center for Toxicological Research, U.S. Food and Drug Administration) for the helpful discussions in the development of this project. This work was supported by the National S&T Major Project (2018ZX09201011), the National Natural Science Foundation of China (81822047), and the National Youth Top-notch Talent Support Program (W02070098).

AUTHOR CONTRIBUTIONS

X.F. conceived and designed the study. X.S. and Y.W. collected and analyzed the genomics data. X.S., X.F., and J.Liao implemented the algorithm of cGPS. X.S. and J.Li performed the cluster analysis and confirmation analysis. X.L., Y.H., and X.C. executed the *in vitro* experiments. Y.W., X.C., and Y.Y. screened the FDA-approved drug library. X.S., X.L., and Y.H. carried out the *in vivo* experiments and RNA sequencing-related experiments. X.S., J.Liao, and N.A. designed and constructed the website. All authors wrote the manuscript and read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 25, 2019
Revised: August 1, 2020
Accepted: October 26, 2020
Published: November 20, 2020

REFERENCES

Allen, D.D., Caviedes, R., Cardenas, A.M., Shimahara, T., Segura-Aguilar, J., and Caviedes, P.A. (2005). Cell lines as in vitro models for drug screening and toxicity studies. Drug Dev. Ind. Pharm. *31*, 757–768.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

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

Du, L., Wang, L., Gan, J., Yao, Z., Lin, W., Li, J., Guo, Y., Chen, Y., Zhou, F., Jim Yeung, S.C., et al. (2019). MTA3 represses cancer stemness by targeting the SOX2OT/SOX2 Axis. iScience *22*, 353–368.

Ertel, A., Verghese, A., Byers, S.W., Ochs, M., and Tozeren, A. (2006). Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. Mol. Cancer *5*, 55.

Formisano, L., Lu, Y., Servetto, A., Hanker, A.B., Jansen, V.M., Bauer, J.A., Sudhan, D.R., Guerrero-Zotano, A.L., Croessmann, S., Guo, Y., et al. (2019). Aberrant FGFR signaling mediates resistance to CDK4/6 inhibitors in ER+ breast cancer. Nat. Commun. *10*, 1373.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Gazdar, A.F., Girard, L., Lockwood, W.W., Lam, W.L., and Minna, J.D. (2010). Lung cancer cell lines as tools for biomedical discovery and research. J. Natl. Cancer Inst. *102*, 1310–1321.

Gillet, J.P., Calcagno, A.M., Varma, S., Marino, M., Green, L.J., Vora, M.I., Patel, C., Orina, J.N., Eliseeva, T.A., Singal, V., et al. (2011). Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. Proc. Natl. Acad. Sci. U S A *108*, 18708–18713.

Glennon, E.K.K., Austin, L.S., Arang, N., Kain, H.S., Mast, F.D., Vijayan, K., Aitchison, J.D., Kappe, S.H.I., and Kaushansky, A. (2019). Alterations in phosphorylation of hepatocyte ribosomal protein S6 control plasmodium liver stage infection. Cell Rep. *26*, 3391–3399 e3394.

Hinohara, K., Wu, H.J., Vigneau, S., McDonald, T.O., Igarashi, K.J., Yamamoto, K.N., Madsen, T., Fassl, A., Egri, S.B., Papanastasiou, M., et al. (2018). KDM5 histone demethylase activity links cellular transcriptomic heterogeneity to therapeutic resistance. Cancer Cell *34*, 939–953 e939.

Holliday, D.L., and Speirs, V. (2011). Choosing the right cell line for breast cancer research. Breast Cancer Res. *13*, 215.

Horvath, P., Aulner, N., Bickle, M., Davies, A.M., Nery, E.D., Ebner, D., Montoya, M.C., Ostling, P., Pietiainen, V., Price, L.S., et al. (2016). Screening out irrelevant cell-based models of disease. Nat. Rev. Drug Discov. *15*, 751–769.

Lee, S.I., Celik, S., Logsdon, B.A., Lundberg, S.M., Martins, T.J., Oehler, V.G., Estey, E.H., Miller, C.P., Chien, S., Dai, J., et al. (2018). A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat. Commun. *9*, 42.

Li, H., Wawrose, J.S., Gooding, W.E., Garraway, L.A., Lui, V.W., Peyser, N.D., and Grandis, J.R. (2014). Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: a rational approach to preclinical model selection. Mol. Cancer Res. *12*, 571–582.

Li, T., Han, J., Jia, L., Hu, X., Chen, L., and Wang, Y. (2019). PKM2 coordinates glycolysis with mitochondrial fusion and oxidative phosphorylation. Protein Cell *10*, 583–594.

Li, Z., Razavi, P., Li, Q., Toy, W., Liu, B., Ping, C., Hsieh, W., Sanchez-Vega, F., Brown, D.N., Da Cruz Paula, A.F., et al. (2018). Loss of the FAT1 tumor suppressor promotes resistance to CDK4/6 inhibitors via the hippo pathway. Cancer Cell *34*, 893–905 e898.

Mologni, L., Redaelli, S., Morandi, A., Plaza-Menacho, I., and Gambacorti-Passerini, C. (2013). Ponatinib is a potent inhibitor of wild-type and drug-resistant gatekeeper mutant RET kinase. Mol. Cell Endocrinol. *377*, 1–6.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515–527.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Qiu, Z., Guo, J., Kala, S., Zhu, J., Xian, Q., Qiu, W., Li, G., Zhu, T., Meng, L., Zhang, R., et al. (2019). The mechanosensitive ion channel Piezo1 significantly mediates in vitro ultrasonic stimulation of neurons. iScience *21*, 448–457.

Ross, D.T., and Perou, C.M. (2001). A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. Dis. Markers *17*, 99–109.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Sandberg, R., and Ernberg, I. (2005). Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). Proc. Natl. Acad. Sci. U S A *102*, 2052–2057.

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., and Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med. Genet. *20*, 175.

Sharma, S.V., Haber, D.A., and Settleman, J. (2010). Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. Nat. Rev. Cancer *10*, 241–253.

Sinha, R., Winer, A.G., Chevinsky, M., Jakubowski, C., Chen, Y.B., Dong, Y., Tickoo, S.K., Reuter, V.E., Russo, P., Coleman, J.A., et al. (2017). Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. Nat. Commun. *8*, 15165.

Stein, W.D., Litman, T., Fojo, T., and Bates, S.E. (2004). A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins. Cancer Res. *64*, 2805–2816.

Sun, Y., and Liu, Q. (2015). Deciphering the correlation between breast tumor samples and cell lines by integrating copy number changes and gene expression profiles. Biomed. Res. Int. *2015*, 901303.

Tohyama, O., Matsui, J., Kodama, K., Hata-Sugi, N., Kimura, T., Okamoto, K., Minoshima, Y., Iwata, M., and Funahashi, Y. (2014). Antitumor activity of lenvatinib (e7080): an angiogenesis inhibitor that targets multiple receptor tyrosine kinases in preclinical human thyroid cancer models. J. Thyroid Res. *2014*, 638747.

Verbeek, H.H., Alves, M.M., de Groot, J.W., Osinga, J., Plukker, J.T., Links, T.P., and Hofstra, R.M. (2011). The effects of four different tyrosine kinase inhibitors on medullary and papillary thyroid cancer cells. J. Clin. Endocrinol. Metab. *96*, E991–E995.

Wang, H., Huang, S., Shou, J., Su, E.W., Onyia, J.E., Liao, B., and Li, S. (2006). Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. BMC Genomics *7*, 166.

Yamori, T. (2003). Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. Cancer Chemother. Pharmacol. *52* (*Suppl 1*), S74–S79.

iScience, Volume *23*

Supplemental Information

A Clinical Genomics-Guided Prioritizing Strategy Enables Selecting Proper Cancer Cell Lines for Biomedical Research

Xin Shao, Yi Wang, Xiaoyan Lu, Yang Hu, Jie Liao, Junying Li, Xuechun Chen, Yunru Yu, Ni Ai, Meidan Ying, and Xiaohui Fan

Supplementary Figures

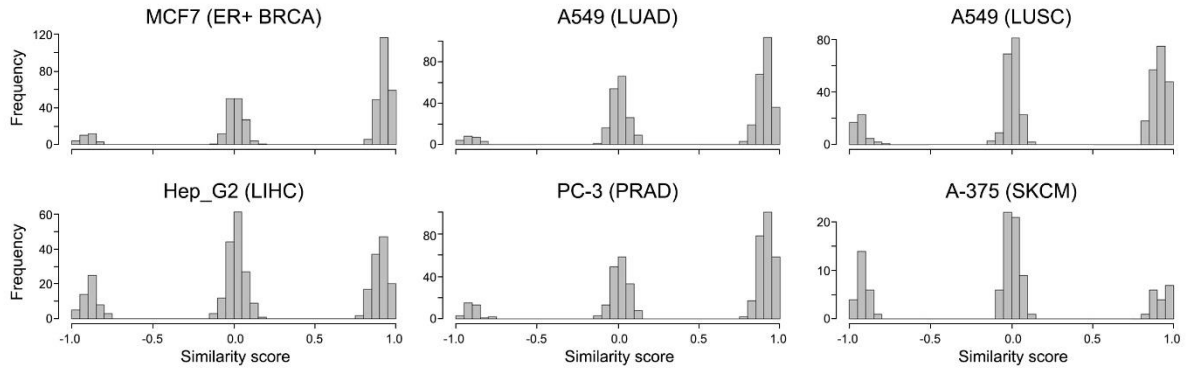


Figure S1. The distribution of similarity score of commonly used cell lines with the corresponding tumor samples, Related to Figure 2. Similarity scores were calculated based on Kolmogorov-Smirnov statistic described as Supplementary method. The commonly used cell line in each tumor was labelled. BRCA: breast invasive carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; LIHC: liver hepatocellular carcinoma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma.

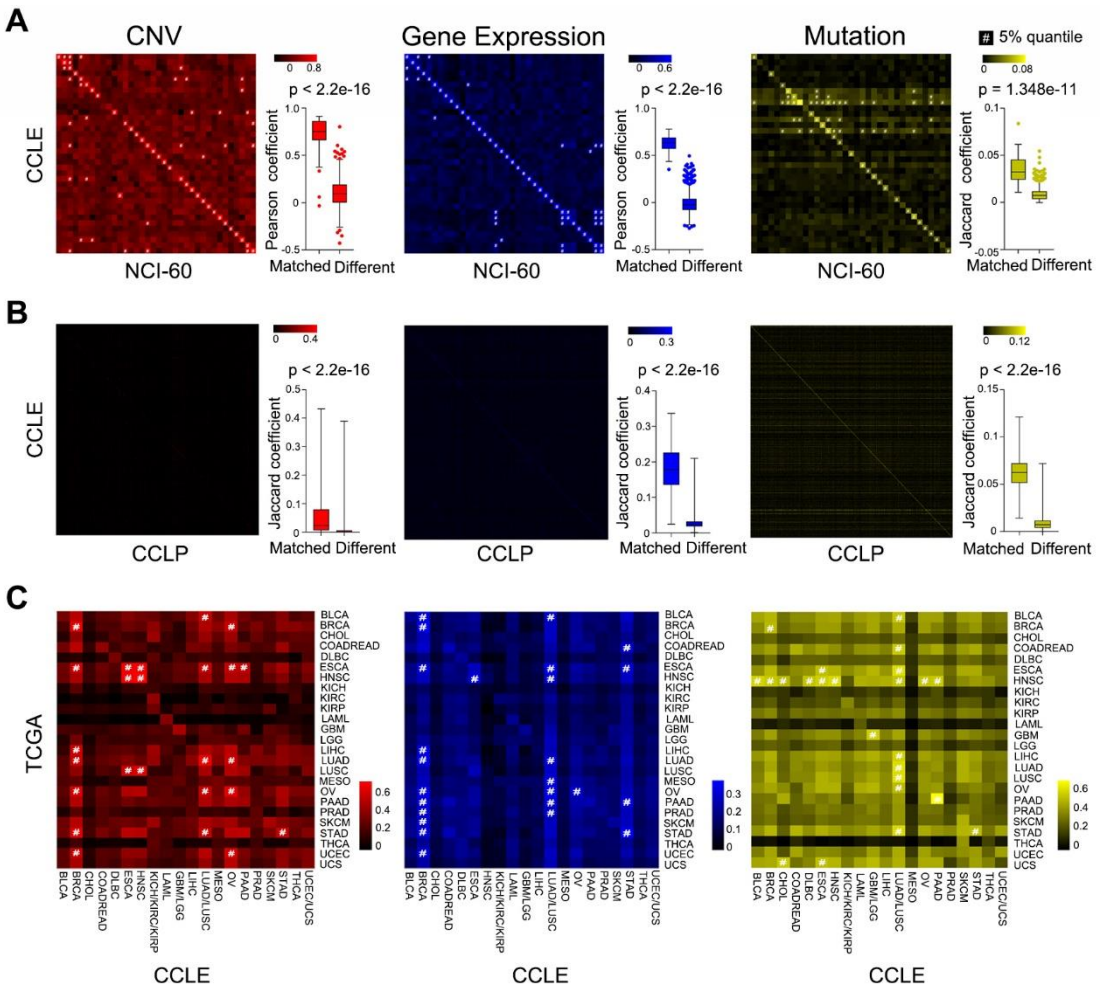


Figure S2. Comparison of genomics data of cell lines among NCI-60, CCLE and CCLP and between cell lines and tumors, Related to Figure 1. Red, blue and yellow represent the comparison of CNV, gene expression and mutation, respectively. **A.** Genomic data correlation comparison (Pearson or Jaccard coefficient) of cell lines shared by NCI-60 and CCLE. # means the coefficient is in top 5% within the matrix. 'Matched' and 'Different' represent on-diagonal and off-diagonal coefficients, respectively. *P* values were obtained from Welch's t-test. **B.** Genomic data correlation comparison (Jaccard coefficient) of cell lines shared by CCLP and CCLE. **C.** Genomics data correlation comparison (Pearson coefficient) between group-wise cell lines from CCLE and tumors from TCGA.

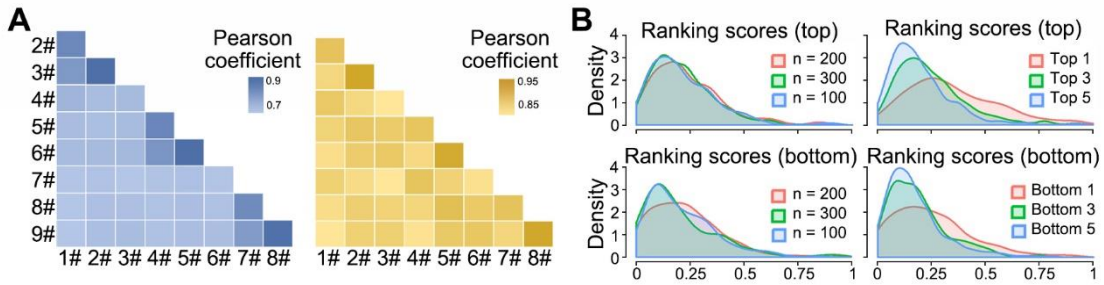


Figure S3. Impact of different parameters on the ranking score from CCL-cGPS, Related to Figure 2. A. Pearson coefficients of 2,341-dimensional vector containing CCL-cGPS scores of the corresponding cell lines across 44 tumor subtypes among nine ranking results with distinct parameters. For blue heatmap, 1# (n = 100, top 1); 2# (n = 100, top 3); 3# (n = 100, top 5); 4# (n = 200, top 1); 5# (n = 200, top 3); 6# (n = 200, top 5); 7# (n = 300, top 1); 8# (n = 300, top 3); 9# (n = 300, top 5). For yellow heatmap, 1# (n = 100, bottom 1); 2# (n = 100, bottom 3); 3# (n = 100, bottom 5); 4# (n = 200, bottom 1); 5# (n = 200, bottom 3); 6# (n = 200, bottom 5); 7# (n = 300, bottom 1); 8# (n = 300, bottom 3); 9# (n = 300, bottom 5). n means the number of signature (DEGs). **B.** Distribution of RSD of the cell lines' ranking scores. Average ranking score being more than 0.5 under each condition for each cell line was selected. X axis: RSD. Y axis: density. The variant signature number conditions (n = 100/200/300) represent RSD of the ranking scores on each cell line across the different number of matched top/bottom candidate cell lines, while the variant matched candidate cell line number conditions (top/bottom 1/3/5) mean RSD of the ranking scores on each cell line across the different number of signature numbers.

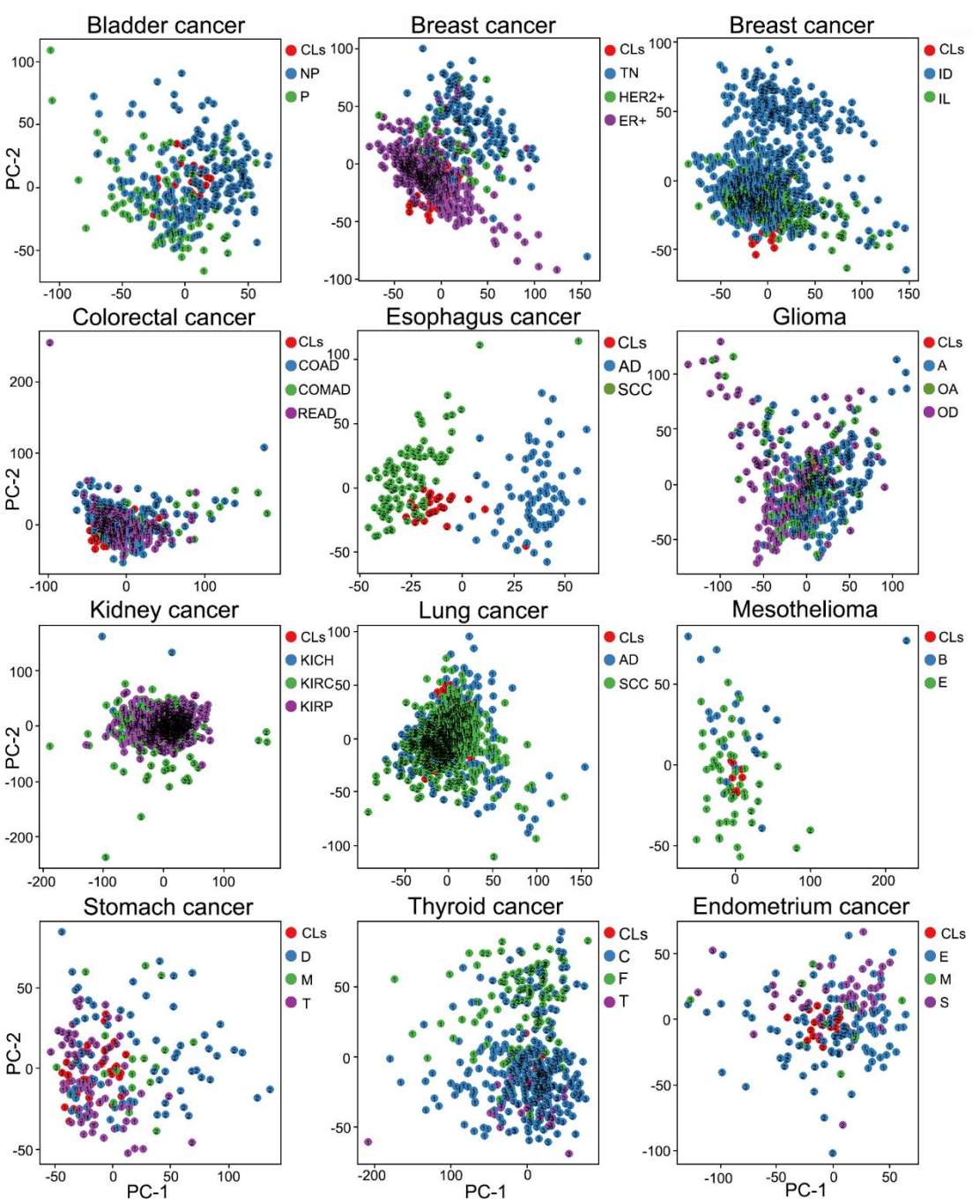


Figure S4. Verification of CCL-cGPS top selection of cell lines by cluster analysis, Related to Table 1. Cell lines and tumor samples were integrated and performed with PCA dimensionality reduction followed by k-means clustering. The cluster information was labelled. CLs: cell lines. NA: not available. P: papillary. NP: Non-papillary. TN: triple negative. ID: Infiltrating ductal. IL: Infiltrating lobular. MU: mucinous for breast and stomach cancer, mixed for endometrium cancer. COAD: colon adenocarcinoma. COMAD: colon mucinous adenocarcinoma. READ: rectal adenocarcinoma. AD: adenocarcinoma. SCC: squamous cell carcinoma. A: astrocytoma. OA: oligoastrocytoma. OD: oligodendroglioma. B: biphasic. E: epithelioid for mesothelioma

and endometrioid for endometrium cancer. D: diffuse. T: tubular for stomach cancer and tall cell for thyroid cancer. C: classical. F: follicular. S: serous.

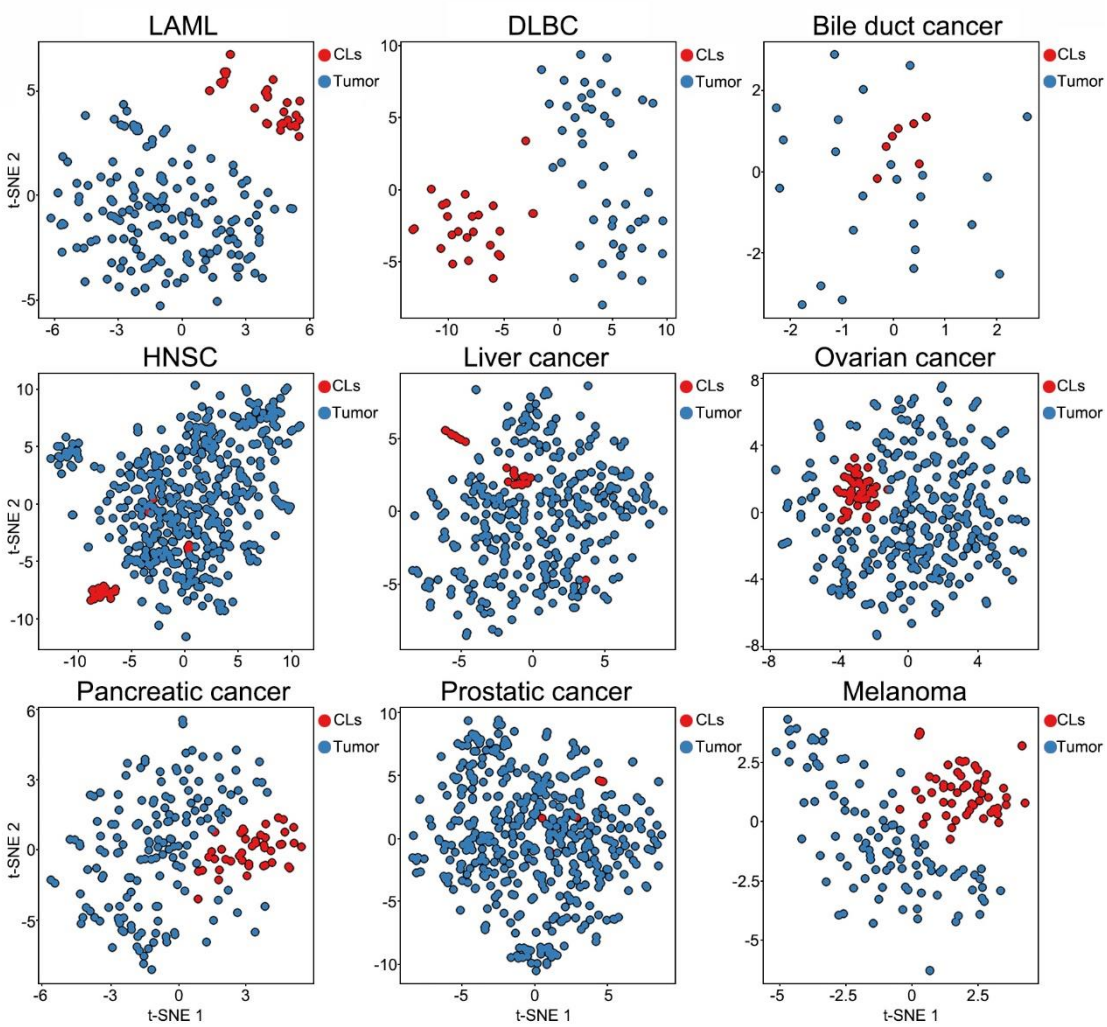


Figure S5. Integrated analysis of cell lines and tumor samples, Related to Table 1. Cell lines and tumor samples were integrated and performed with t-SNE dimensionality reduction. The information of cell lines and tumor samples was labelled. Liquid cancers, including LAML and DLBC, were therefore exclude from CCL-cGPS because of the huge genomic discrepancy between cell lines and cancer samples. CLs: cell lines. DLBC: diffuse large b-cell lymphoma. HNSC: head and neck squamous cell carcinoma. LAML: acute myeloid leukemia.

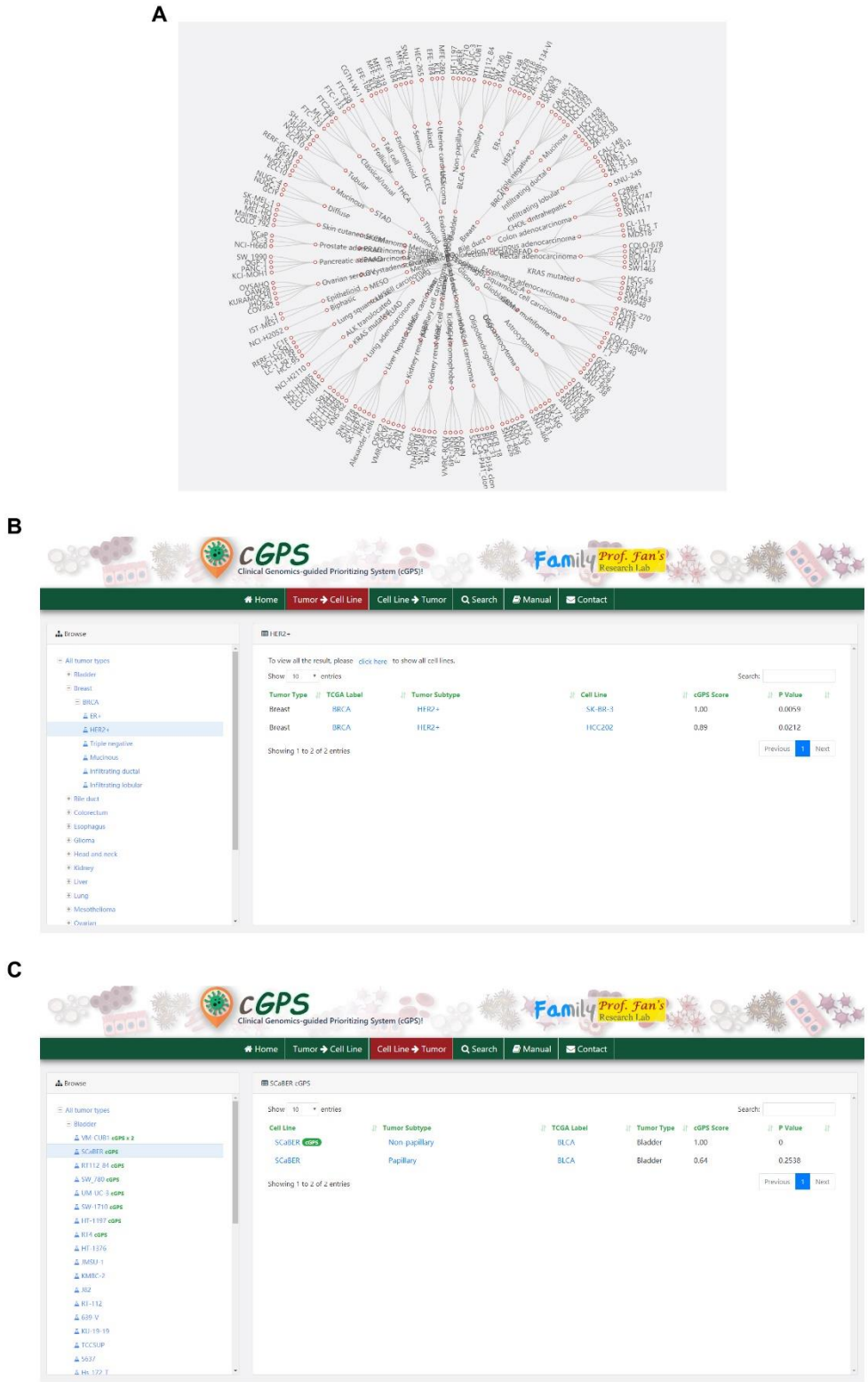


Figure S6. Web tool of CCL-cGPS, Related to Figure 1. A. Overview of CCL-cGPS includes 18 tumors, 23 related TCGA labels, 44 tumor subtypes and 168 CCL-cGPS selected cell lines. **B.** Website shows the results of CCL-cGPS selected cell lines when choosing HER2+ breast tumor

subtype in “Tumor to Cell line” page. **C.** Website shows the results of appropriate tumor subtype when choosing bladder cancer cell line ScaBER in “Cell line to Tumor” page.

Transparent Methods

Genomics data resources

The CNV, mRNA expression and mutation profiles from CCLE, NCI-60 and TCGA were collected from the cBioportal for Cancer Genomics (https://www.cbioportal.org/). Genomic data of CCLP was retrieved from COSMIC Cell Lines Project (v81, https://cancer.sanger.ac.uk/cell_lines).

Identification of gene expression variants and CNV

In terms of segmented data obtained from Affymetrix SNP6.0 array, GISTIC 2.0 was applied to determine the putative copy number calls for CCLE, NCI-60 and TCGA datasets (Values: -2 = homozygous deletion; -1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high level amplification), wherein gene-wise homozygous deletion or high level amplification were regarded as copy number amplified or deleted gene. However, in CCLP, copy number was acquired by PICNIC in which copy number amplification was obtained by the following criteria: (the average genome ploidy < =2.7 AND total DNA segment copy number > =5) OR (average genome ploidy >2.7 AND total DNA segment copy number > =9). While the criteria for copy number deletion was: (the average genome ploidy < =2.7 AND total DNA segment copy number =0) OR (average genome ploidy >2.7 AND total DNA segment copy number < (average genome ploidy – 2.7)).

Gene expression levels were quantified by RSEM from RNA-seq data and mRNA Z scores were computed using the tumors samples that are diploid for the corresponding gene. Differentially expressed genes (DEGs) were further filtered out as Z scores more than 2 (highly expressed or upregulated genes) or less than -2 (lowly expressed or downregulated genes).

Similarity analysis of cancer cell lines from different resources

Pearson's correlation coefficient based on copy number and mRNA Z scores of shared genes was performed to explore the similarity of cell lines for the common ones between CCLE and NCI-60, while Jaccard index was applied to mutated genes to quantify the similarity coefficient

since mutation data was nonnumeric. Regarding the relevance between CCLE and CCLP, Jaccard index based on CNVs including copy number amplified and deleted genes, DEGs or mutated genes was calculated respectively to detect their concordance on genomic data.

Comparison of cancer cell lines to primary tumors

By comparing the primary site or histology of cancers in which tumor samples and cell lines both involved, a panel of 779 relevant cancer cell lines (Supplementary Table S1) and a total of 25 tumor types (Supplementary Table S2) were selected for analysis of cell line versus tumor comparisons. For tumors/cell lines integrated analysis, frequency profiles PD(C) comparisons were applied in which D represents one of the three type of data-omics, C means tumor type or cell lines group, containing n entries: one for genomic variation event from CNV (copy number amplified or deleted genes), gene expression (highly or lowly expressed genes) and mutated genes. The value of the i-th entry of the profile correspond to the percentage of total samples of this tumor type or cell lines group in which the i-th genomic variation event was present. Thus, a total of 25 profiles for tumors and 20 for cell lines were assembled for further computing Pearson's correlation coefficient on shared genes between each pair of 25 tumor types and 20 cell lines groups. All these coefficients were then arranged into pairwise comparison matrices MD containing 25 rows (cancer type; CNV and gene expression; 24 rows for mutation matrix) and 20 columns (cell lines groups) yielding results showed in the heatmaps of Supplementary Figure S1C.

Subtype classification of tumor samples.

For each cancer in TCGA, tumor samples without other malignancy history were selected for subtype classification (at least ten samples). It has been widely revealed that substantial tumor heterogeneity consists of various subtypes from histological, pathological and increasing gene-expression based molecular aspects. Among 11 cancer types in TCGA, namely BLCA, BRCA, CHOL, COADREAD, ESCA, LGG, LUAD, MESO, STAD, THCA and UCEC, we further divided them into 32 subtypes (Supplementary Table S2) according to the specific histological or diagnosis information for the histopathological subtype or the status of specific molecular

markers for the genotype. As the HNSC is related to several primary sites rather than histopathology, HNSC tumor samples were regarded as one subtype. GBM, HNSC, KICH, KIRC, KIRP, LIHC, LUSC, OV, PAAD, PRAD, SKCM and UCS were treated as single subtype for the lack of available histopathological or molecular subtypes. Oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status (positive or negative) by immunohistochemistry were used to determine the genotype of breast invasive carcinoma (BRCA). Besides, *KRAS* mutation and *ALK* translocation were applied as two additional molecular characteristics to classify colon and rectum adenocarcinoma (COADREAD), lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC).

Similarity ranking of cancer cell lines

Gene expression profiles of each cell line and tumor sample were reordered in descending order *(G)* according to the Z scores. For tumor samples, the top and bottom genes (50/100/150) were filtered out as the signature (*n* = 100/200/300) including the upregulated genes (*UG*) and downregulated genes (*DG*), respectively. Cell lines were scored and ranked with respect to one sample based on Kolmogorov-Smirnov statistic as follows. For each cell line *i*, two enrichment score for *UG* and *DG* were respectively computed, namely S^i_{UG} and S^i_{down}. Let *N* be the total number of unique genes in cell line expression profile. Construct a position set *V* of each gene in *G* and sort these elements in ascending order (*V* {1, 2, …, *N*}) such that *V(j)* is the position of gene *j*, where *j* = 1, 2, …, *n* from *UG* or *DG*. Compute the following two values:

$$a = \max_{j=1}^{n} \left\{ \frac{j}{n} - \frac{V(j)}{N} \right\}$$

$$b = \max_{j=1}^{n} \left\{ \frac{V(j)}{n} - \frac{j-1}{N} \right\}$$

For both S^i_{UG} *and* S^i_{down}, set $S^i = a$ if $a > b$ or $S^i = -b$ if $b >$ a. Set $s^i = S^i_{UG} - S^i_{down}$, p = max (s^i) and q = min (s^i) across all the cell lines. The similarity score SS^i is defined as s^i / p where $s^i >= 0$, or $-(s^i / q)$ where $s^i < 0$. Let vector *SS* be the final score of all cell lines in descending order, in

which high or low SS^i was regarded as positive or negative similarity between the cell line and the tumor sample for yielding the histogram plot in Supplementary Figure S1.

Let k be the number of unique cell line among the candidate set C from tumor samples of one cancer subtype. For each cell line i, the frequency m^i was counted based on C. Construct a vector M [m^1, m^2, …, m^k] in descending order. Set x = max (M). The final CCL-cGPS ranking score (RS) is defined as M / x. In terms of the two types of candidate set C^t and C^b generated from the top and bottom cell lines, RS^t and RS^b were computed respectively, in which top cell lines in RS^t or RS^b resemble positively or negatively the specific tumor subtype.

In this study, top 1/3/5 cell lines were selected as the matched candidate cell lines for each tumor sample. For a panel of tumor samples, matched candidate cell lines were pooled together followed by counting and normalization. On the basis of assigned frequency, candidate cell line for each subtype was prioritized in descending order by ranking score.

The distribution pattern of ranking scores

For each cancer subtype, the distribution pattern of ranking scores of cell lines was determined by the ratio of the number of cell lines with ranking scores being more than 0.5 to that with ranking scores of less than 0.5. The subtypes with ratio values of more than 2 were regarded as having most cell lines similar with the tumor, while those with values of less than 0.5 were having most cell lines dissimilar with the tumor. The rests represented the tumor subtypes with evenly distributed cell line ranking scores.

Significantly similar cell lines identification

The permutation test was performed to construct the background model. For each tumor subtype, simulated candidate cell lines (1/3/5) with the same tissue origin were randomly sampled and counted. Repeat this step 10,000 times to obtain a distribution for each simulated candidate cell line. The frequency of the candidate cell line in CCL-cGPS was then compared with this background. The null hypothesis was rejected when the frequency of the candidate cell line was equal to or less than 5% of the random distribution.

Impact of different parameters on CCL-cGPS scores

By varying the number of selected gene signatures (100, 200 and 300) and cell lines (top 1, 3 and 5), we obtained nine combinations. Under each combination, we obtained a 2,341-dimensional vector containing CCL-cGPS scores of the corresponding cell lines towards the corresponding tumor subtypes and transformed CCL-cGPS scores into CCL-cGPS ranks for the corresponding cell lines towards the corresponding tumor subtypes. Then, the Pearson coefficients of CCL-cGPS scores' ranks were calculated between paired combinations, which generates the heatmap in Figure S3A.

Commonly used cell line retrieval in tumor subtypes

Total number of cited articles up to July, 2020 for each cell line was acquired from PubMed with the key words (Detailed in Supplementary Table S1). For each cancer subtype and each associated cell line, the PubMed search term was constructed by using the corresponding Mesh term and the name of the cell line to retrieve the related articles. Taking ER^+ breast cancer subtype and MCF-7 as an example, the search term is ("Breast Neoplasms"[Mesh]) AND (("Receptors, Estrogen"[Mesh]) OR ER OR ESR OR luminal) AND (MCF-7 OR MCF7) to retrieve related studies on ER^+ breast cancer with MCF7. Then, we carefully read full texts, especially the Methods section, to verify whether MCF7 is regarded as an ER^+ breast cancer cell line model by researchers in their studies. Concordant studies were collected and the sum of concordant studies were counted for each cell line for each cancer subtype. Last, cell lines with the same origin were ranked in descending order by the number of concordant studies. The cell line with the most citation was regarded as the commonly used one for the associated cancer subtype.

Verification from confirmation analysis and cluster analysis

Histopathological or molecular information of selected cell lines by CCL-cGPS were collected from CCLE. The gene expressions of ESR1, ERBB2, and PGR were used to classify breast cancer cell lines into molecular subtypes, known as ER^+, $HER2^+$, and triple-negative. The mutation profile of *KRAS* was used to determine the molecular subtype, namely *KRAS* mutated, for

colorectal and lung cancer cell lines. For each cancer, overlapped subtypes between cell lines and tumor samples were retained. Then for each cancer subtype, the percentage of matched selected cell lines with the same histopathological or molecular profiles was calculated to evaluate the biological traits of the selected cell lines by CCL-cGPS.

For each tumor type, the gene expression data (Z score of each gene) of related cell lines and tumor samples were integrated for performing dimensionality reduction using principal components analysis (PCA) with default parameters. Resulted two-dimensioning data was clustered by k-means, in which the cluster number was mainly the number of histopathological or molecular subtypes. For each subtype, the cluster with the most tumor samples was regarded as the matched cluster. The hit rate for each tumor subtype was calculated by the number of CCL-cGPS selected cell lines falling within this matched cluster divided by the total number of CCL-cGPS selected cell lines.

IC_{50} values of FDA-approved drugs

FDA-approved drugs for ER+, HER2+ breast and differentiated thyroid cancer were curated from FDA (https://www.fda.gov/). The specific molecular targeted drugs were searched in PubMed, PubChem and CCLE and checked manually to obtain the reported IC50 values on breast and thyroid cancer cell lines. The widely used antitumor drugs for preferred medications functioning as cytotoxicity (such as cisplatin, fluorouracil, paclitaxel, and their analog, etc.) were excluded, resulting in molecular targeted drugs including tamoxifen for ER+ breast cancer, lapatinib for HER2+ breast cancer, lenvatinib and sorafenib for follicular thyroid cancer (one type of differentiated thyroid cancer). IC50 values of tamoxifen, lapatinib, lenvatinib and sorafenib on breast or thyroid cancer cell lines were obtained from PubChem and CCLE project (Fit type with Sigmoid and average activity SD less than 10 was selected), while IC50 related research articles were retrieved from PubMed by using the keyword “drug name (e.g. tamoxifen) AND IC50 AND cancer type (e.g. breast cancer)” or “drug name (e.g. tamoxifen) AND IC50 AND representative cell line name (e.g. MCF7)”in order to obtain the IC50 values. With respect to the multiple IC50 values of a certain drug towards the same cell line, the median IC50 were used for yielding the point plot in Figure 3.

Conserved DEGs and pathways between CCL-cGPS selected cell lines and tumor patients

For each cell line, highly expressed ($Z > 2$) and lowly expressed ($Z < -2$) genes were treated as DEGs. For each tumor subtype, highly expressed and lowly expressed genes for each patient were combined and compared with highly expressed and lowly expressed genes of the corresponding CCL-cGPS selected cell line, respectively. Genes that were both highly expressed or both lowly expressed in the cell line and the tumor patients were regarded as the conserved DEGs. To compare the conserved pathways between highly ranked cell lines and associated tumor samples, e.g., ER^+ breast cancer, DEGs of a tumor subtype were ordered with frequency across tumor patients and the same number of DEGs as the one in the CCL-cGPS selected cell line were selected for enriching pathways by using Metascape (http://metascape.org/).

Reagents and cell lines

Tamoxifen, lapatinib, cisplatin and digitoxin were purchased from Sigma-Aldrich Co. (St Louis, MO). Lenvatinib and sorafenib were obtained from Selleckchem (USA). Mebendazole was provided by Aladdin Co. (Shanghai, China). FDA-approved drug library was obtained from Target Molecule Co. (TargetMol). Breast cancer cell lines HCC202, MDA-MB-436, HCC1187, BT-549, thyroid cancer cell line CAL-62, colorectal cancer cell line NCI-H747, and bladder cancer cell line RT4 were purchased from Shanghai Xinyu Biological Technology Co., Ltd (Shanghai, China). Breast cancer cell line MCF7 was a gift from Prof. Wang's lab (Hangzhou, China). Thyroid cancer cell lines 8305C and 8505C were provided by Zhejiang Cancer Hospital (Hangzhou, China). The HCC202, MDA-MB-436, BT-549, NCI-H747, 8305C, and 8505C cell lines were maintained in RPMI 1640 (HyClone, USA) with L-glutamine supplemented with 10% fetal bovine serum (FBS, Gibco, USA) and 1% penicillin-streptomycin mixture (PSM, Gibco, USA). The MCF7, HCC1187, and CAL-62 cells were cultured in DMEM/HIGH GLUCOSE (HyClone, USA) supplemented with 10% FBS and 1%PSM. The RT4 was cultured in McCoy's 5A (Modified) Medium (BI, Israel) supplemented with 10% FBS and 1% PSM. All the cell lines were kept in a humidified incubator (Thermo Fisher, USA) at 37 °C with 5% carbon dioxide.

High throughput screening

RT4 and NCI-H747 were automatically seeded into 96-well flat bottom plates at a density of 7 × 10^3 cells per well with 100 µL appropriate growth medium by using Multidrop Combi (Thermo, USA). After 24-hour incubation, the medium was removed and replaced by 90 µL fresh culture, and then each drug from FDA-approved drug library (a total of 1068 drugs were tested in this study) was automatically added to one well with 10 µL culture medium using a Liquid Handler (Tecan Fluent 780, Switzerland). The concentration of all the drugs was set at 10 µM. After 48 hours, the culture medium was discarded and cells were treated with 100 µL 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) of 0.5 mg/mL (Sigma, USA) at 37°C for 4 hours, followed with 100 µL DMSO. The optical density of each well was measured at 580 nm using a microplate reader (Infinite M1000 Pro, TECAN, Germany). The cell viability of drug-treated groups was quantified as a percentage compared to the vehicle control.

Half maximal inhibitory concentration detection

Cell lines were seeded into 96-well flat bottom plates containing 100 µL appropriate growth medium per well at an appropriate density (HCC202, HCC1187, 104 cells/well; MDA-MB-436, BT-549, RT4, NCI-H747, 7x10^3 cells/well; MCF7, 8505C, 2x103 cells/well; 8305C, 1x10^3 cells/well) and the medium was replaced after 24 hours with the corresponding culture medium without FBS and PSM and then treated with the corresponding drugs (tamoxifen and lapatinib for HCC202, MCF7, BT-549, HCC1187 and MDA-MB-436; lenvatinib and sorafenib for 8505C, 8305C, CAL-62; digitoxin and mebendazole for RT4; digoxin and adapalene for NCI-H747). After 48 hours, the culture medium was discarded and cells were replenished with 100 µL MTT solution at 37°C for 4 hours followed with 100 µL DMSO. The cell viability was examined as previously described. The IC_{50} values were calculated by GraphPad Prism 8.0.

RNA-seq experiment

RT4 cells were seeded into cell culture dish (60 mm, Corning, USA) and the medium was replaced after 24 hours with the McCoy's 5A culture medium without FBS and PSM. The cells

were then treated with or without digitoxin at 0.01 μM or mebendazole at 0.1 μM to generate control, digitoxin- or mebendazole-treated groups (n = 3). After 24 hours, the culture medium was discarded and total RNA was extracted using the mirVana miRNA Isolation Kit (Ambion) following the manufacturer's protocol. RNA integrity was evaluated using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The samples with RNA Integrity Number (RIN) ≥ 7 were subjected to the subsequent analysis. The libraries were constructed using TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer' s instructions. Then these libraries were sequenced on the Illumina sequencing platform (Illumina HiSeq X Ten) and 150 bp paired-end reads were generated.

RNA-seq data analysis

The transcriptome sequencing and analysis were conducted by OE biotech Co., Ltd. (Shanghai, China). Raw data (raw reads) were processed using Trimmomatic. The reads containing poly-N and the low-quality reads were removed to obtain the clean reads. Then the clean reads were mapped to reference genome using hisat2. The read counts of each gene were obtained by htseq-count and were normalized with median 1000 and log2. Upregulated genes were identified as those with fold change of more than 2 by comparing the mebendazole- or digitoxin-treated group to the control, while downregulated genes were identified as having fold change being less than -2.

Animal experiments

Male BALB/c nude mice (4-6 week, Silaike Co. Shanghai, China) were housed under specific pathogen-free conditions with a 12 h light/dark cycle. All the cages, food, and water were sterilized. The mice were injected subcutaneously in the right flank with 0.1mL of cell suspension containing 1×10^7 RT4 cells or 3.5×10^6 NCI-H747 cells. Tumors were allowed to grow for approximately 7 days to a volume of 100 mm^3 measured using a caliper before treatment. Tumor-bearing mice (RT4) were randomly allocated into 4 groups (n = 7): cisplatin group (2 mg/kg every three day, i.p.), digitoxin group (0.5 mg/kg/day, i.p.), mebendazole group (50 mg/kg every other day, i.g.), and the vehicle control group (0.9% sodium chloride,

0.1 mL /10 g/day, i.p.). Tumor-bearing mice (NCI-H747) were randomly allocated into 4 groups (*n* = 4): capecitabine group (150 mg/kg every three day, i.g.), digoxin group (1.2 mg/kg/day, i.p.), adapalene group (50 mg/kg every other day, i.g.), and the vehicle control group (0.9% sodium chloride, 0.1 mL /10 g/day, i.p.). The tumor size was measured twice a week using a caliper. After five-week treatment for RT4 and four-week treatment for NCI-H747, mice were sacrificed and the tumors were immediately harvested and weighed. All the animal experimental protocols were approved by the Animal Care and Use Committee of the Zhejiang University School of Medicine.

Statistics

R (version 3.6.1) and GraphPad Prism 8.0 were used for the statistical analysis. R packages of FactoMineR and ggplot2 was used for PCA analysis and data visualization, respectively. IC_{50} values were estimated from the nonlinear regression of dose-response inhibition. Differences between two groups were determined using the Welch's t-test (significant with $P < 0.05$). Differences of tumor size among four groups were assessed by using one-way ANOVA (significant with $P < 0.032$).