Briefings in Bioinformatics, 18(3), 2017, 515–529

doi: 10.1093/bib/bbw082
Advance Access Publication Date: 2 October 2016
Paper

Optimization of cell lines as tumour models by integrating multi-omics data

Ning Zhao*, Yongjing Liu*, Yunzhen Wei*, Zichuang Yan, Qiang Zhang, Cheng Wu, Zhiqiang Chang and Yan Xu

Corresponding author: Yan Xu, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China. Tel.: (+86) 0451-86674768; E-mail: xuyan@ems.hrbmu.edu.cn
*These authors contributed equally to this work.

Abstract

Cell lines are widely used as *in vitro* models of tumorigenesis. However, an increasing number of researchers have found that cell lines differ from their sourced tumour samples after long-term cell culture. The application of unsuitable cell lines in experiments will affect the experimental accuracy and the treatment of patients. Therefore, it is imperative to identify optimal cell lines for each cancer type. Here, we review the methods used to evaluate cell lines since 2005. Furthermore, gene expression, copy number and mutation profiles from The Cancer Genome Atlas and the Cancer Cell Line Encyclopedia are used to calculate similarity between tumours and cell lines. Then, the ideal cell lines to use for experiments for eight types of cancers are found by combining the results with Gene Ontology functional similarity. After verification, the optimal cell lines have the same genomic characteristics as their homologous tumour samples. The contaminated cell lines identified in previous research are also determined to be unsuitable *in vitro* cancer models here. Moreover, our study suggests that some of the commonly used cell lines are not suitable cancer models. In summary, we provide a reference for ideal cell lines to use in *in vitro* experiments and contribute to improving the accuracy of future cancer research. Furthermore, this research provides a foundation for identifying more effective treatment strategies.

Key words: cell line; multi-omics; cancer in vitro model; optimization

Introduction

Cell lines have been used in a large number of scientific experiments as *in vitro* models of malignant tumours. Research has shown that the genetic alterations in tumour-derived cell lines are similar to their corresponding tumours [1]. Therefore, the correct use of these cell lines could greatly accelerate the understanding of cancer biology. However, some studies have demonstrated genetic differences between cancer cell lines and their corresponding tumour samples after long-term artificial culture, which could be because of improper cultivation or the occurrence of frequent mutations [2–9].

Cell lines have already been widely applied to assist in the diagnosis and treatment of cancer and in the development of new treatment strategies for cancer. Therefore, improperly using cell lines for research not only can distort the result of the experiments but also has the potential to cause patients to receive an incorrect or potentially hazardous treatment [10]. Nevertheless, a large amount of research still uses improper cell lines. These biased experimental results will affect the credibility of cancer research and increase the difficulty of finding new treatments. Thus, it is imperative to evaluate the suitability of the cell lines that are used as *in vitro* models of tumours.

Ning Zhao is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Yongjing Liu is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Yunzhen Wei is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Zichuang Yan is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Qiang Zhang is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Cheng Wu is a postgraduate student in College of Bioinformatics Science and Technology at Harbin Medical University.
Zhiqiang Chang is a lecturer for College of Bioinformatics Science and Technology at Harbin Medical University.
Yan Xu, PhD, is a professor for College of Bioinformatics Science and Technology at Harbin Medical University.
Submitted: 3 April 2016; **Received (in revised form):** 23 August 2016

Increasing evidence has indicated that cancer-related genes generally have differential expression, copy number alterations and mutations. Genetic alteration is a major mechanism underlying the transformation of normal cells to cancerous cells [11, 12]. Driver alterations can cooperatively disrupt a cascade of crucial biological pathways and thereby provide cells with significant growth advantages [13, 14].

With the improvement of high-throughput sequencing technology, a large number of gene expression differences, DNA copy number alterations and mutations in individual cancers were revealed [15–17]. Therefore, it is possible to evaluate the suitability of cell lines on the genomic level. Domcke *et al.* [18] used a fitness score to evaluate cell lines with the same genomic characteristics as high-grade serous ovarian cancer (HGSOC) tumour samples. They demonstrated that the most commonly used HGSOC cell lines are not appropriate, but that cell lines more analogous to tumour samples are not often wielded in the laboratory. However, their research only concerned one cancer subtype. Olarerin-George *et al.* [19] analysed the RNA sequencing (RNA-seq) data of 9395 robust primate samples to evaluate the extent of mycoplasma contamination in cell culture. The results showed that 11% of the samples were contaminated. However, they only studied the samples at the expression level and expounded on the phenomenon of cell line contamination, but did not indicate which cell line would be a better *in vitro* model. Experimenters are badly in need of a list of cell lines that are closer to the source tumour samples to perform *in vitro* experiments across cancer types. With such a list, they could research cancer mechanisms more accurately and create more effective treatments.

Different types of cancers have specific genomic characteristics [20, 21], even from the same tissue [22]. Therefore, a cell line is the ideal representation of the original tumour if it has the same expression, copy number alterations and driver mutation or other genome variations as the tumour samples. The Cancer Genome Atlas (TCGA) provides the gene expression, copy number and mutation profiles of a wide variety of cancers [23]. Similarly, the Cancer Cell Line Encyclopedia (CCLE) offers gene expression, copy number and mutation profiles of multiple types of tissues [24]. Using this information, we can compare various cell lines with tumour samples in these data sets.

In this study, we first reviewed the methods used to evaluate cell lines since 2005. Most of these methods did not use high-throughput data and concentrated on only one cancer or a few cell lines. Next, we downloaded the gene expression, copy number and mutation profiles of tumour samples and cell lines from TCGA and CCLE, respectively. Similarity in the expression, copy number and mutation frequency and type between tumour samples and cell lines was calculated to identify the cell lines with similar genetic characteristics as tumour samples. Then, the Gene Ontology (GO) term network affinity between the samples was calculated to screen cell lines on the functional level. Finally, we identified optimal cell lines as *in vitro* models for eight types of cancers: breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), glioblastoma multiforme (GBM) and brain lower grade glioma (LGG). We further verified the cell lines represented by A2780 that were contaminated and were no longer suitable for use as a cancer model. Overall, the similarity between the cell lines and their corresponding tumour samples was studied on the genomic level. The mechanisms for which these cell lines are used as *in vitro* models were illuminated more deeply, which has acute significance. This work provides a list of optimal cell lines and will help researchers choose cancer models and improve the accuracy of their experiments.

Materials and methods

Gene expression, copy number and mutation profiles

Expression, copy number and mutation profiles for eight types of cancers (BRCA, OV, KIRC, KIRP, COAD, READ, GBM and LGG) were downloaded from TCGA. Specifically, KIRC and KIRP are two subtypes of renal cancer. COAD and READ are two subtypes of intestinal cancer. GBM and LGG are two subtypes of glioma. The samples which associated with more than one barcode were deleted. The quantity of the samples for each profile is shown in Table 1.

The profiles for the cell lines were downloaded from CCLE if the cell line had expression, copy number and mutation profiles available. The number of samples and the cancer type for each cell line are shown in Table 2.

Series GSE2109 was downloaded from the Gene Expression Omnibus (GEO) [25]. In total, 353 breast, 400 intestine, 281 kidney and 199 ovary samples were used in this study to confirm our results by using data from different sources.

Cancer-related genes

Cancer-related genes were downloaded from four databases (Cosmic [26], GAD, Phenopedia [27] and OMIM [28]). The number of related genes for each cancer is shown in Table 3.

The similarity between samples

For TCGA copy number data, we chose hg19 and then used Gistic2.0 [29] to obtain a copy number matrix. Only the rows in the expression and copy number matrixes for cancer-related genes were retained. Genes with missing values in >5% of the samples were removed, and we then obtained the other missing values using the k-nearest neighbour method. For mutation data, the silent mutations were deleted. Because CCLE detected

Table 1. The number of samples of each profile

Cancer	CNA	GE	GM
Breast invasive carcinoma (BRCA)	515	521	509
Ovarian serous cystadenocarcinoma (OV)	585	588	92
Kidney renal clear cell carcinoma (KIRC)	485	72	213
Kidney renal papillary cell carcinoma (KIRP)	275	16	168
Colon adenocarcinoma (COAD)	413	155	216
Rectum adenocarcinoma (READ)	162	69	81
Glioblastoma multiforme (GBM)	575	519	316
Brain lower grade glioma (LGG)	523	27	516

Table 2. The number of samples and corresponding cancer for each cell type

Tissue	Number of samples	Corresponding cancer
Breast	59	BRCA
Ovary	51	OV
Kidney	35	KIRC, KIRP
Central nervous system	59	GBM, LGG
Large intestine	59	COAD, READ
Small intestine	1	

Table 3. The number of cancer-related genes

Cancer	Cosmic	GAD	Phenopedia	OMIM	Total
Breast	21	882	1433	35	1488
Ovary	16	858	1156	89	1204
Kidney	23	1334	1490	78	1691
Glioma	26	111	634	16	668
Colorectum	22	435	1083	60	1203

the mutation of only 33 genes [30], we studied the mutation of these 33 genes (Supplementary Table S1) in TCGA data as well.

Regarding the different data platforms of TCGA and CCLE, the DWD method [31] in the CONOR package [32] was applied for multiplatform normalization. After normalization, the expression and copy number values were accurate to nine decimal places. Tiny fluctuations in these values may greatly influence the rank. Two decimal places were reserved so we could put several slightly different values as the same rank to reduce the influence of numerical fluctuation.

Kendall Rank Correlation Coefficient [33] (KRCC, τ) was applied to calculate the correlation of expression and copy number between the tumour (X) and cell line (Y) samples:

$$\tau = \frac{C - D}{\sqrt{(N_3 - N_1)(N_3 - N_2)}}$$

where C is the number of concordant pairs, and D is the number of discordant pairs.

$$N_1 = \sum_{i=1}^{s} \frac{1}{2} U_i (U_i - 1)$$

$$N_2 = \sum_{i=1}^{m} \frac{1}{2} V_i (V_i - 1)$$

$$N_3 = \frac{1}{2} N(N - 1)$$

N is the length. The elements with the same value are grouped. There are s groups in X and m groups in Y. U_i and V_i are the number of elements in the ith group.

The corresponding P value was computed by either the exact permutation distributions (for small sample sizes) or large-sample approximations.

The threshold of correlation coefficient was decided by the median of the distribution of the correlation between cancer samples. The significance threshold for the correlation was set to $P < 0.05$ based on other works. To confirm the threshold was indeed statistically significant, we repeated the calculation of the correlation 100 times with permuted data, obtained by randomly switching the genes' IDs. When $P < 0.05$, the same correlation coefficient threshold as the real data was set to the permuted data. And the numbers of correlated pairs of real and permuted data were compared [34].

The screening of variant genes

Differentially expressed genes were selected. For each gene, the fold-change [35] (*Fold_c*) value ($\log_2$ed) of its expression and the median expression of normal samples were calculated. The threshold of *Fold_c* is >2 or <0.5. Fold-change significance was assessed using a *t*-test. After 1000 simulations, the significance threshold of corrected P value was set to false discovery rate (FDR) [36] $\leq$0.05 [37].

Genes that underwent copy number alterations were selected. The discretization of copy number was defined according to the standard of Ciriello *et al.* [20]. Briefly, the copy number value of protein-coding genes in tumour samples within the range [−0.1,0.1] indicates that the gene is diploid and does not have a copy number alteration, as denoted by '0'; greater than 0.1 indicates a copy number amplification, denoted by '1'; and less than −0.1 indicates a copy number deletion, denoted by '−1'.

The construction of the GO term network

Human blood pressure terms were picked out from gene2go (18 March 2015 updated), which was downloaded from NCBI. The GO term network was built from GO: 0008150 [38]. Only considering 'is_a' relationships, the final network includes 9536 terms and 13 076 edges.

GO enrichment analysis

Cumulative hypergeometric inspection was applied to GO enrichment analysis:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where N is the whole number of genes. M is the number of genes on a term. n is the intersection of interested gene set and N. i is the intersection of M and n. Significance threshold of hypergeometric test was set to $P < 0.05$ and FDR ≤ 0.05.

The similarity of networks

Three types of variant genes of each sample were enriched to GO terms, and each term was coloured. A term was coloured yellow, blue or red if one, two or three types of variation existed, respectively.

The network similarity score (*Score*) was calculated for each pair of tumour and cell line samples, defined as

$$Score = \frac{T \cap C}{T \cup C} \times S_c$$

where T is the number of enriched GO terms of the tumour sample and C is the number of enriched GO terms of the cell line. S_c is the number of GO terms that have the same colour in both the tumour and cell line samples.

Results

Methods for cell line evaluation

Researchers have long been aware of the misidentification and contamination of cell lines. Over the past decade, many studies have assessed the conformity of cell line models. We summarized the methods used to evaluate cell lines since 2005 (Table 4). These methods are generally based on DNA profiling, mutation or expression and are used to evaluate the cancer cell lines.

Table 4. Methods for cell line evaluation

Method	Number of cell lines	Cancer	PMID
Cytogenetic and genetic fingerprint analysis	1	Colon	15771911
Sterility test	21 (including five human cell lines)	–	15939285
PCR-based assay and 'barcode region' amplification and sequencing	67 (including two human cell lines)	–	17934781
p53 mutation status and activity	1211	–	18277095
PCR amplification and DNA sequencing	11 (including three human cell lines)	–	18553210
STR and SNP array analysis	40	Thyroid cancer	18713817
STR examinations	6	Adenoid cystic carcinoma	19557180
68 publications	360 (including 324 human cell lines)	–	20143388
DNA microsatellite STRs, p53 nucleotide polymorphisms and microsatellite instability	51	Ovarian and endometrial cancers	22710073
CLIFF	300	Solid tumours and leukaemias	23356232
STR analysis	1	Bladder cancer	23500642
Genetic similarity	47	High-grade serous ovarian cancer	23839242
RNA-seq sequence data analysis	9395 (primates and rodents)	–	25712092
STR and SNP profiles	3587	–	25877200
STR profiling methods	380	–	26116706
STR DNA profiling	1	Oesophageal squamous cell carcinoma	26432330
Mass spectrometric fingerprints of intact mammalian cells coupled with ANNs	1	Embryonic stem cells	26821236
This work: multi-omics data	264	Eight kinds of cancers	

Analysis of short tandem repeats (STRs) is the standard test for authenticating cell lines, as recommended by the American Type Culture Collection (ATCC) Standards Development Organization Workgroup ASN-0002 [39]. Nearly half of the cell line studies that we evaluated used STR. Ayyoob *et al.* [40] used a STR DNA profiling method to confirm unique identity of a newly established oesophageal squamous cell carcinoma cell line (YM-1). Ye *et al.* [41] presented a comprehensive survey of cross-contamination in 380 samples from 113 independent sources in China using STR profiling methods. They found that 25% of the cell lines were cross-contaminated and that 85.51% of cell lines established in China were misidentified. Jäger *et al.* [42] performed STR analysis on the cell line KU7 found that the KU7 cell line had been cross-contaminated with HeLa cells before 1984 at the source institution. Phuchareon *et al.* [43] performed DNA fingerprint analysis on six ACC cell lines using STR examinations and found that all six cell lines had been contaminated with other cells. Using more computational methods, Somaschini *et al.* [44] developed a software tool called Cell Line Identity Finding by Fingerprinting (CLIFF) to compare the STR profiles of 300 widely used tumour cell lines. However, STR profiling could only separate individual cell lines within a single species.

Single-nucleotide polymorphism (SNP) genotyping is another DNA profiling method that can be used to track biosamples [45]. Furthermore, some studies have integrated STR and SNP profiles to evaluate cell lines. Yu *et al.* [46] present a framework for cell line annotation linked to STR and SNP profiles, and provide a catalogue of synonymous cell lines. Schweppe *et al.* [47] evaluated thyroid cancer-derived cell lines using STR and SNP array analysis. They found that only 23 of the 40 cell lines tested had unique genetic profiles. Interestingly, some cell lines were found to be derivatives of the same cell line, and some cell lines were misidentified.

P53 mutation is the most common genetic abnormality found in human cancers [48]. Berglind *et al.* [49] analysed the p53 status of 1211 cell lines. Their results clearly confirmed that misidentified cell lines are still a silent and neglected danger and that extreme care should be taken because an incorrect p53 status could lead to disastrous experimental interpretations. Korch *et al.* [50] obtained cell lines from multiple institutions and analysed them using DNA microsatellite STR, p53 nucleotide polymorphisms and microsatellite instability. Their results demonstrate significant misidentification, duplication and cross contamination.

For the interspecies identification of cell cultures, several polymerase chain reaction (PCR)-based methods have recently been described. Liu *et al.* [51] described a PCR-based method for the rapid identification and authentication of closely related cell lines. The method enables the routine identification of cell cultures among closely related species of various cell lines. Cooper *et al.* [52] developed a multiplex PCR-based assay to rapidly identify the most common cell culture species and quickly detect cross-contaminations among these species. Then, the 'barcode region' is amplified and sequenced by using a universal primer mix targeted at conserved sequences in the cytochrome c oxidase I gene. These assays can be used to accurately determine the species of cell lines. Unlike STR, PCR-based methods can only be used to identify cell lines from different species.

In addition to the tests described above, other studies have used different methods. Mirjalili *et al.* [53] conducted a 2 year study to detect the microbial contaminations and causative organisms of the cell lines in the Cell and Gene Bank, Biotechnology Department of Razi Vaccine and Serum Research Institute. The sterility test demonstrated that 39% of the cell lines were contaminated. Valletta *et al.* [54] developed and validated a rapid and routinely applicable method for evaluating cell culture cross-contamination levels based on mass spectrometric fingerprints of intact mammalian cells coupled with artificial neural networks (ANNs). Melcher *et al.* [55] performed a cytogenetic and genetic fingerprint analysis of the putative normal colon epithelial cell line NCOL-1 and found that it was probably derived from the colon carcinoma cell line LoVo, which likely occurred through cross-contamination. Capes-Davis *et al.* [56] summarized 360 cross-contaminated cell lines drawn from 68 references. These cell lines were classified according to the contamination type and time.

With the development of sequencing technology, there are more and more genomic profiles in public databases. Olarerin-George *et al.* [19] surveyed NCBI's RNA-seq archive of 884 series to assess the prevalence of mycoplasma contamination in cell culture. They found that 11% of these series were contaminated. Domcke *et al.* [18] analysed a panel of 47 ovarian cancer cell lines and identified those that had the highest genetic similarity to ovarian tumours. From these, they identified several rarely used cell lines that more closely resemble cognate tumour profiles than the commonly used cell lines and proposed that these cell lines are the most suitable models of ovarian cancer.

In the present work, we examined the expression, copy number and mutation profiles as well as functional data. From this analysis, we chose the optimal cell line models for eight types of cancers by calculating KRCC and functional similarity between the samples. This method used more complete types of genomic data and evaluated more cancer types.

Different cancer samples can indeed be distinguished using the KRCC

To verify whether different cancer samples can be distinguished by calculating KRCC, we downloaded the expression data of eight types of cancers from TCGA. Classification of the breast cancer molecular subtypes from the TCGA data was obtained from Koboldt *et al.* [57].

To begin the analysis, we first computed the correlation among the samples of each cancer and between each pair of cancers. The correlation was measured by a uniform threshold ($|correlation\ coefficient| > 0.4$ and $P < 0.05$) to facilitate the comparison of the results. As shown in Table 5, the values are the maximum/average numbers of related samples. The results suggested that the number of similar samples of the same cancer is clearly greater than the number of similar samples among the other cancers, especially when comparing the average value. The number of correlated samples between two cancers that originated from the same organ is higher than for other cancers originating from different organs. For example, almost all of the cancers have no correlated LGG sample, except GBM (26/23.85) and LGG (26/26). The details of the calculated results are shown in Supplementary Table S2.

Next, to evaluate the luminal (347 samples) and basal (88 samples) subtypes of BRCA, the correlation between samples from the same subtype or different subtypes was estimated (the same threshold $|correlation\ coefficient| > 0.4$ and $P < 0.05$ was set as above) (Supplementary Table S3). Notably, the maximum and average numbers of related luminal samples to luminal samples are 345 and 340.06, respectively, and the maximum and average numbers of related luminal samples to basal samples are 321 and 169.34, respectively. Interestingly, luminal samples have a strong correlation to basal samples. However, the relevant sample proportion for basal subtypes was significantly lower than that of the luminal subtype.

Therefore, using KRCC to calculate the correlation between samples can effectively distinguish samples from different cancers or different cancer subtypes. Furthermore, if a cell line has low correlation with tumour samples, it should not be a suitable *in vitro* model. Therefore, KRCC can be used to filter cell lines that are more similar within a specific cancer type.

The workflow of our approach is depicted in Figure 1. The expression, copy number and mutation profiles of eight types of cancers and six types of cell lines were downloaded from TCGA and CCLE, respectively. The profiles of cancer-related genes were extracted (Figure 1A). By applying KRCC, the inter-sample similarity between cell lines and tumours was calculated to obtain candidate cell lines (Figure 1B). Furthermore, to evaluate the functional correlation, the variant genes (differential expressed genes, copy number-altered genes and mutated genes) were detected and mapped to the GO term network to extract specific three-colour subnets for each tumour sample and cell line. The similarity score of each pair of the three-colour subnets was calculated. The optimal cell lines were confirmed if the previous candidate cell lines had functional similarity with the tumours as well (Figure 1C).

The expression, copy number, mutation and functional correlation of the cell lines and tumour samples

This work used KRCC to calculate the similarity in expression and copy number for tumour samples and cell lines. In addition, we used the similarity scores of GO term networks to evaluate their functional closeness (Figure 2A, Supplementary Figures S1A and S2A). The similarity in mutation profile was assessed by evaluating mutations in, e.g. TP53, PIK3CA and KRAS.

The distributions of correlation coefficient of cancer samples are shown in Supplementary Figures S3–S5. Then, the thresholds were decided by the median of the distribution. The thresholds of correlation and the functional similarity score in each cancer are shown in Table 6. 100 permutations were performed using the strategy described in Method section. The results showed that no permutation meets the threshold we set to the real data so that no correlated pairs were obtained. The distributions of correlation coefficient of permutations are shown in Supplementary Figure S6. Count score was the number of

Table 5. The number of related samples of each cancer

	BRCA (521)	OV (575)	COAD (155)	READ (69)	GBM (562)	LGG (27)	KIRC (72)	KIRP (16)
BRCA	506/442.56	369/118.82	69/8.73	19/3.30	185/6.33	0/0	58/10	6/0.16
OV	430/107.63	565/492.10	100/11.68	48/5.22	231/7.73	1/0	62/6.78	9/0.56
COAD	320/29.64	251/44.23	153/151.05	68/67.24	5/0.07	0/0	59/5.54	5/0.56
READ	273/24.88	223/43.54	153/152.26	68/67.71	26/0.67	0/0	54/5.72	4/0.55
GBM	133/5.87	85/7.91	4/0.02	3/0.81	559/541.84	26/23.85	59/7.95	6/0.20
LGG	0/0	1/0.04	0/0	0/0	545/515.59	26/26	0/0	0/0
KIRC	203/72.51	163/54.61	53/11.94	21/5.54	239/62.31	0/0	70/64.17	14/10.21
KIRP	22/5.13	76/20.38	27/5.44	13/2.38	52/7.13	0/0	67/45.19	15/12.88

The total number of samples is in the brackets after each column name. The values indicate 'the maximum/average value of related samples' to the cancer of each row name. The number of similar samples of the same cancer is clearly greater than the number of similar samples among the other cancers, especially when comparing the average value. The number of correlated samples between two cancers that originated from the same organ is higher than for other cancers originating from different organs.

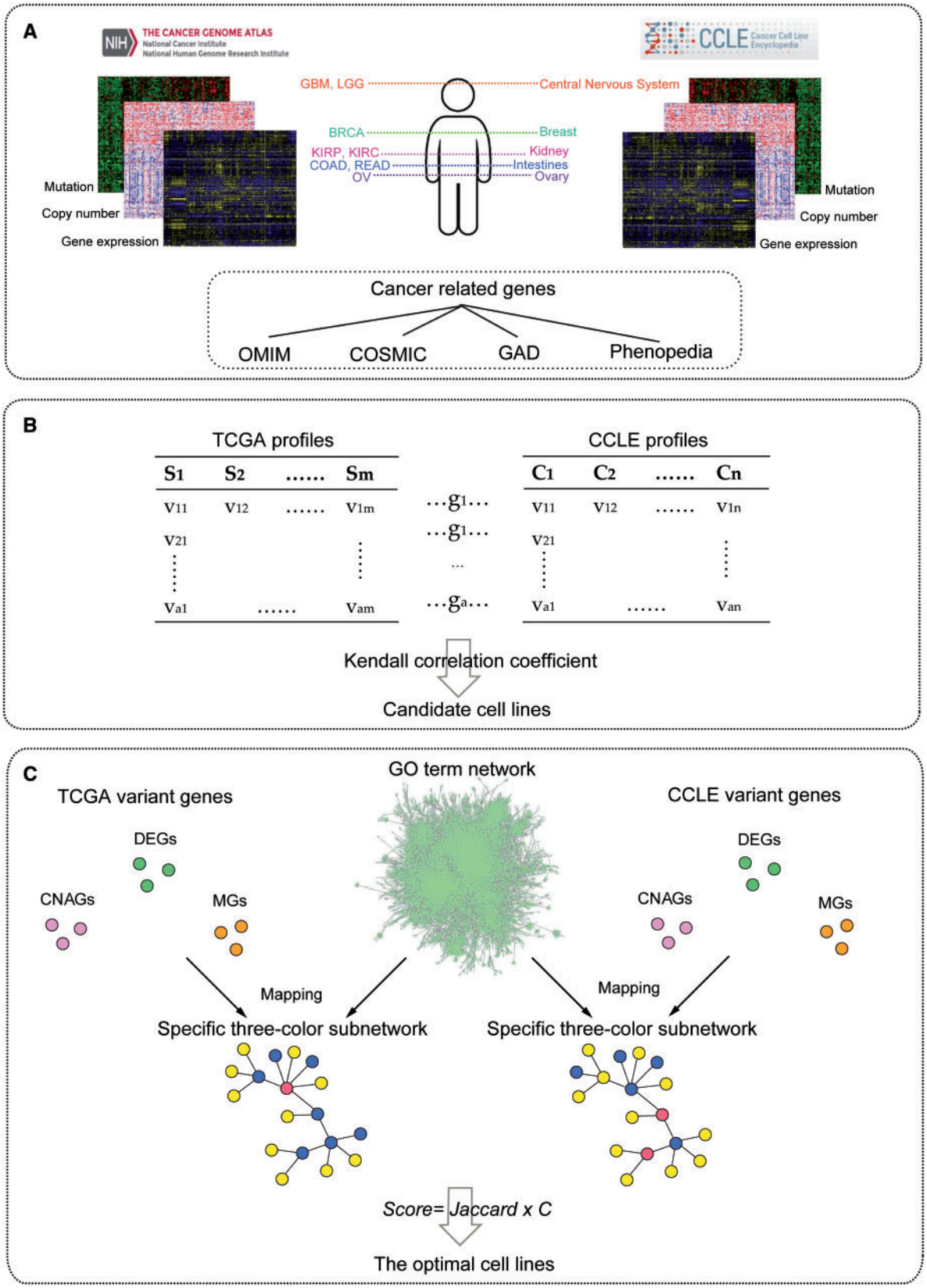

Figure 1. Flowchart of the study. (**A**) Data source of cancer-related genes and gene expression, copy number and mutation profiles of cancers and cell lines. (**B**) Similarity calculation between TCGA and CCLE samples was computed using KRCC to obtain candidate cell lines. (**C**) Variant genes were selected for each sample of TCGA and CCLE, respectively. These genes were then mapped to the GO term network to get specific three-colour subnets. Then, the similarity score of TCGA and CCLE three-colour subnets was calculated. Finally, the optimal cell lines were obtained by the score and the candidate cell lines (DEG = differential expressed gene; CNAG = copy number altered gene; MG = mutated gene.). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

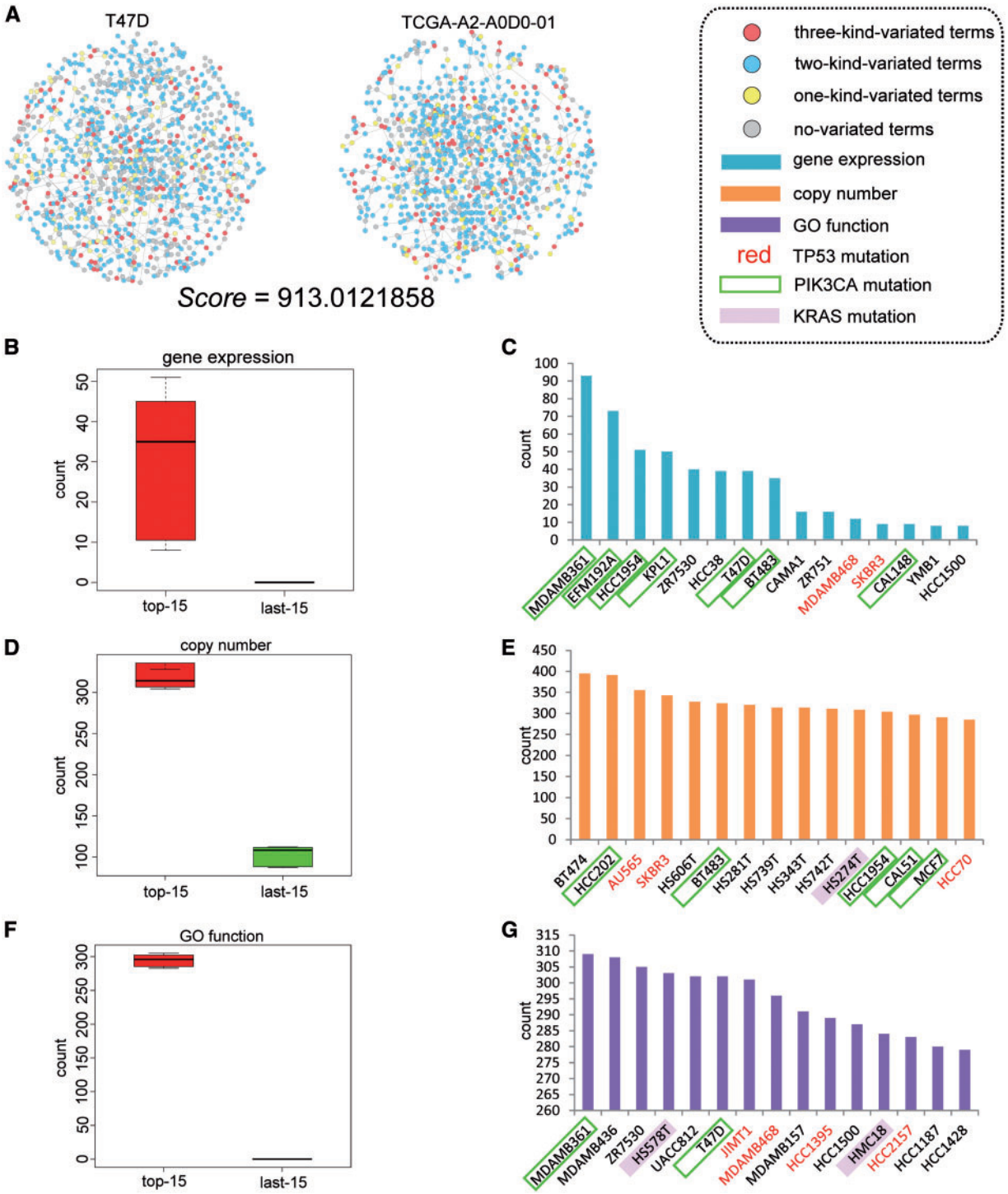

Figure 2. The BRCA top cell lines of expression, copy number and function. (**A**) An example for functional network similarity score. (**B**, **D** and **F**) The comparison of gene expressed, copy number altered and functional count scores of the top 15 and the last 15 cell lines, respectively. (**C**, **E** and **G**) The expressed, copy number altered and functional count scores of the top 15 cell lines, respectively. The mutated genes of each cell line are shown on the labels. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

tumour samples associated with each cell line that met the threshold (Supplementary Table S4). The cell lines were sorted by count scores, and the top cell lines were retained.

To illustrate the similarity between the cell lines and tumour samples without subtype, the top 15 mammary cell lines in terms of expression, copy number and function data were selected (Figure 2). Likewise, the similarity between COAD and READ tumour samples and intestinal cell lines was calculated to evaluate the congruity between cell lines and tumour samples with subtypes. The top 10 cell lines were selected based on their expression, copy number and functional data (Supplementary Figures S1 and S2). The figures show that the top cell lines are significantly similar to more samples than the last cell lines.

As shown in Figure 3A, some cell lines were highly similar to many samples (such as T47D), whereas others were not similar

Table 6. The threshold of similarity between cell lines and each cancer

Type	GE	CN	GO
BRCA	0.49	0.13	264
OV	0.47	0.18	335
KIRC	0.59	0.13	8
KIRP	0.50	0.21	6
GBM	0.56	0.17	22
LGG	0.67	0.09	3
COAD	0.58	0.15	26
READ	0.59	0.20	98

GE, CN and GO columns represent gene expression, copy number and GO function, respectively.

to any of the tumour samples (such as EVSAT). The correlations between cell lines and tumours were not balanced in all aspects. For example, HCC38, HS606T and HS578T in the cell lines were not at all similar to the tumours in terms of their level of expression, copy number and function. Similarly, as shown in Figure 3B and C, the correlations between cell lines and the COAD or READ tumour samples were found to be unbalanced. For instance, SNU61 was correlated with many samples, but the number of correlated samples with HS698T was much less. HUTU80 was only similar to tumours on the level of copy number, and OUMS23 was significantly similar to tumours in terms of function. Interestingly, some cell lines had a higher correlation coefficient with COAD than READ, and vice versa, such as C2BBE1 and LS513.

In summary, different cell lines have different degrees of similarity to tumours, and have a different number of related samples. It is feasible to identify the optimal cell line for cancers based on relevance.

The optimal cell lines for different types cancers

Because of the high heterogeneity in malignant tumour samples, a cell line may not be able to highly correlate with tumour samples on all levels. Therefore, the optimal cell line to use as a model of tumours was defined as the cell line with similarities in at least three out of four of the following: expression, copy number, mutation and function (Table 7).

Regarding breast cancer cell lines, T47D, MDAMB361 and MDAMB468 have high expression and functional scores and possess mutations in PIK3CA or TP53, respectively. However, the copy number score of these cell lines is not very high. HCC1954, SKBR3 and BT483 have the highest expression and copy number scores and have a mutation in PIK3CA or TP53. However, these cell lines do not have a high functional score. Therefore, based on these analyses, we believe that T47D, MDAMB361, MDAMB468, HCC1954, SKBR3 and BT483 are the optimal cell lines of BRCA.

Regarding intestinal cell lines, for both COAD and READ, SNU61 have the top scores in expression, copy number and function and contain mutations in KRAS and TP53. Both SW1463 and SW403 have high expression and copy number ranks as well as mutations in TP53 or KRAS. LS513 has high expression and function ranks and a mutation in KRAS. SNU283 has top scores in expression, copy number and function, but does not has any mutations in the three genes assessed (PIK3CA, TP53 and KRAS). Hence, SNU61 is the optimal cell line for colorectal cancer. SW1463, SW403, LS513 and SNU283 are all tied for 2nd place.

For COAD only, SW948 has the top scores in expression, copy number and function and contain a mutation in KRAS and PIK3CA. LS180 ranks highly in terms of expression and copy number, and has mutations in PIK3CA and KRAS. For READ only, NCIH508 is similar to the tumor samples regarding expression, copy number, and has mutations in TP53 and PIK3CA. Both NCIH747 and SW1116 have high copy number and function ranks as well as mutations in KRAS. Taken together, these results suggested that SW948 is the optimal cell line of COAD, and LS180 comes in second. And NCIH747, NCIH508 and SW1116 are the optimal cell lines of READ.

The optimal cell lines that were identified have similar genomic characteristics to tumour samples

To further validate the optimal cell lines, this work summarized the genomic characteristics of tumour samples. Different types of tumour samples have different genomic characteristics. BRCA, COAD and READ were used as examples.

Most of the BRCA samples have mutations in PIK3CA and TP53 (Figure 4A). Notably, all of the optimal cell lines (T47D, HCC1954, BT483, MDAMB361, SKBR3 and MDAMB468) of BRCA have mutation in at least one of these genes (Figure 4B). Furthermore, as shown in Figure 4C, the top mutated genes in the tumour samples have mutations in cell lines as well. Therefore, the optimal cell lines have the same mutation characteristics as the tumour samples.

The BRCA samples have significant copy number deletions on every chromosome. Additionally, these samples have significant copy number amplifications on chromosomes (chrs) 1, 6, 8, 11, 12, 16, 17 and 20 (Figure 4E). The genes with the highest copy number alterations in BRCA are shown in Figure 4D. As shown in this figure, the percentages of these genes' copy number alterations are similar in these cell lines.

Additionally, we investigated the expression similarity between the tumour samples and the optimal cell lines for BRCA. The top five differentially expressed genes (BRCA1, BRCA2, INHBA, E2F5 and TNFRSF11B) in the tumour samples were detected. As expected, these genes are significantly differentially expressed in the optimal cell lines (Figure 4F).

For COAD and READ, as shown in Figure 5A, most of the COAD and READ samples have mutations in PIK3CA, KRAS and TP53. Consistently, most of the optimal cell lines (SNU61, SW1463, SW403, SNU283, LS180, LS513, SW948, NCIH747, NCIH508 and SW1116) of intestinal cancer have mutations in at least one of those genes (Figure 5B). Furthermore, the top mutated genes in the tumour samples were also mutated in the cell lines (Figure 5C). Thus, the optimal cell lines have the same characteristics as the tumour samples with respect to mutations.

Samples of COAD have significant copy number deletions and amplifications in chrs 6, 10, 17 and 19. Additionally, they have significant copy number amplifications in chrs 8, 12, 13 and 20, and significant deletions in chrs 6, 10, 17 and 19 (Figure 5E). Samples of READ have significant copy number deletions and amplifications in chrs 1, 6 and 20. Additionally, they have significant copy number amplifications in chrs 8, 10, 11, 12, 13 and 17, and significant deletions in chrs 3, 4, 5, 16, 18 and 21 (Figure 5F). The genes with the highest copy number alterations in intestinal cancer are shown in Figure 5D. As shown in this figure, the genes evaluated have significant variation between the different cell lines.

In addition, we also investigated the expression similarity between intestinal tumour samples and the optimal cell lines.

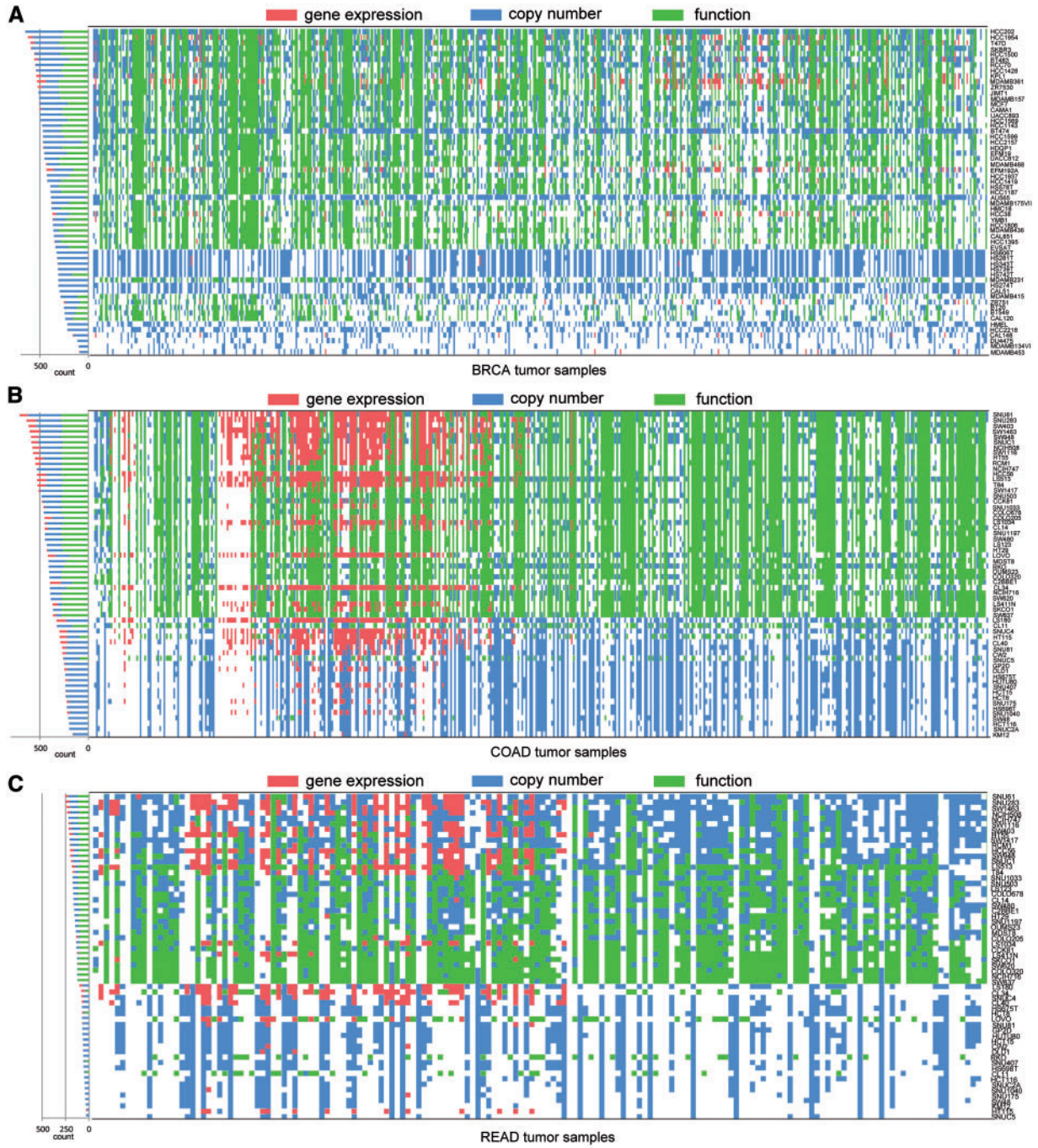

Figure 3. Distribution of related tumour samples of each cell line in BRCA, COAD and READ. In the heatmap, each row represents a cell line, and each column represents a tumour sample. Blocks represent the association of expression, copy number and function, respectively. The histogram on the left shows the total number of each kind of associations. (**A**) BRCA samples with mammary cell lines. (**B**) COAD samples with intestinal cell lines. (**C**) READ samples with intestinal cell lines. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

As shown in Figure 5G and H, the top five differentially expressed genes in COAD (PLA2G4A, EGF, SFRP2, PTPN13 and IGFBP2) and READ (PLA2G4A, ADRA2C, PTPN13, SFRP2 and IGFBP2) were detected. Notably, these genes differ significantly in expression compared with the optimal cell lines.

In conclusion, the optimal cell lines obtained by the method established in this work have similar genomic characteristics as the BRCA, COAD and READ tumour samples.

Some of the commonly used cell lines are not suitable for research, but some of the most suitable cell lines have not been used

It has been confirmed that the optimal cell lines identified by the methods described in this work have the same genomic characteristics as the tumour samples. The next question to address is how frequently these optimal cell lines are used in research.

To this end, we searched the frequency that various cell lines are used as *in vitro* models for various cancers. The number

Table 7. The optimal cell lines for each cancer

Cell line	Tumour type	Source	GE	CN	GO	SM
T47D	BRCA	Ductal carcinoma	7	–	5	PIK3CA
MDAMB361	BRCA	Carcinoma	1	–	1	PIK3CA
MDAMB468	BRCA	Carcinoma	11	–	8	TP53
HCC1954	BRCA	Ductal carcinoma	3	12	–	PIK3CA
BT483	BRCA	Ductal carcinoma	8	6	–	PIK3CA
SKBR3	BRCA	Carcinoma	12	4	–	TP53
OC316	OV	Adenocarcinoma	5	13	–	TP53
TYKNU	OV	Undifferentiated carcinoma	–	3	1	TP53
OC314	OV	Serous carcinoma	1	10	–	TP53
OVSAHO	OV	Adenocarcinoma	10	11	10	–
OVCAR4	OV	Carcinoma	2	6	6	–
KURAMOCHI	OV	Undifferentiated carcinoma	11	12	7	–
KMRC20	KIRC	Renal cell carcinoma	2	4	1	–
VMRCRCZ	KIRP	Renal cell carcinoma	1	5	1	–
TUHR4TKB	KIRP	Carcinoma	1	9	4	–
TUHR14TKB	KIRP	Carcinoma	1	3	5	–
CAS1	GBM	Astrocytoma Grade IV	7	9	–	TP53
LN229	GBM	Astrocytoma Grade IV	1	6	15	ERBB2
T98G	GBM	Astrocytoma Grade IV	2	11	3	–
YKG1	GBM	Astrocytoma Grade IV	2	5	7	–
LN340	LGG	Astrocytoma Grade IV	13	3	2	–
CAS1	LGG	Astrocytoma Grade IV	1	4	–	TP53
HS683	LGG	Glioma	7	–	14	TP53
SNB19	LGG	Astrocytoma Grade IV	13	–	2	TP53
U251MG	LGG	Astrocytoma	4	15	14	TP53
SW948	COAD	Adenocarcinoma	13	9	1	KRAS, PIK3CA
SNU61	COAD	Carcinoma	7	1	1	KRAS, TP53
LS180	COAD	Adenocarcinoma	5	12	–	KRAS, PIK3CA
SW1463	COAD	Adenocarcinoma	2	13	–	KRAS, TP53
SW403	COAD	Adenocarcinoma	12	2	–	KRAS
LS513	COAD	Adenocarcinoma	6	–	12	KRAS
SNU283	COAD	Carcinoma	3	4	8	–
SNU61	READ	Carcinoma	9	1	1	KRAS, TP53
SW1463	READ	Adenocarcinoma	3	7	–	KRAS, TP53
SW403	READ	Adenocarcinoma	10	5	–	KRAS
NCIH747	READ	Adenocarcinoma	–	2	7	KRAS
SW1116	READ	Adenocarcinoma	–	9	6	KRAS
NCIH508	READ	Adenocarcinoma	8	4	–	PIK3CA, TP53
LS513	READ	Adenocarcinoma	4	–	10	KRAS
SNU283	READ	Carcinoma	5	6	13	–

The optimal cell lines for each cancer. The table showed the rank of each cell line in gene expression (GE), copy number (CN) and GO term network similarity and somatically mutated (SM) genes of each cell line. It will be marked a '–' if the cell line is not on the top. A cell line is considered as a candidate if it is on the top in more than three aspects of the four.

of PubMed abstracts mentioning one of the CCLE cell lines was searched (built on 19 January 2016) using several punctuation alternatives for the cell line names. Interestingly, many of the optimal cell lines have rarely been used in research. Conversely, some of the popular cell lines that have been used frequently in research are not on our list of optimal *in vitro* models.

The literature references for the mammary cell lines are shown in Figure 6A. T47D is ranked 3rd, SKBR3 is ranked 4th, MDAMB468 is ranked 6th and MDAMB361 is ranked 12th based on actual usage frequency. However, BT483 and HCC1954 are not conventional cell lines. The literature reference for the intestinal cell lines is shown in Figure 6B. Four of the optimal cell lines are not used frequently, with the exception of LS180 and SW1116, which is ranked 12thand 14th, respectively. Of note, HT29, which is the most popular cell line, has a low similarity with the tumour samples. This cell line is ranked nearly 30th for expression, copy number or function in COAD and READ, except for the function of READ.

We next set out to extra determine whether our method is reliable. Cross-contaminated or misidentified cell lines were obtained from the International Cell Line Authentication Committee [56]. Nine of the cell lines used in this work are known to be contaminated (Table 8). None of these cell lines was found in the list of optimal cell lines, and they have generally low scores for each of the variations. For example, the scores of BT20 for expression, copy number and function are only ranked no more than the 40th. However, BT20 is the 9th most used cell line (Figure 6A).

Our results are similar to those of Domcke *et al.* [18]. Four of our optimal ovary cell lines (KURAMOCHI, OVSAHO, OVCAR4 and TYKNU) are defined as 'likely high-grade serous' in their work. Although the scores of their top-ranked cell lines in our work are lower than those of our optimal cell lines, the cell lines still have relative high scores and are considered good cell line models by our analysis (Supplementary Table S5). Additionally, they claimed that the two most frequently used cell lines, SKOV3

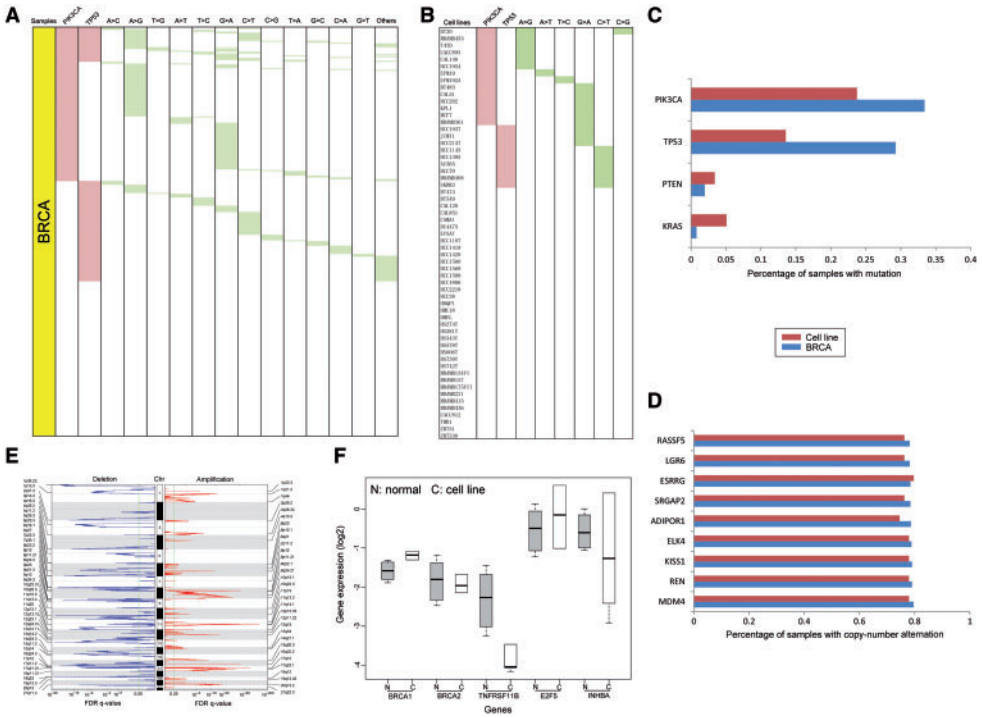

Figure 4. Comparison of mutation, copy number and expression between tumour samples and cell lines of breast cancer. (**A**)The mutation of BRCA samples. (**B**) The mutation of mammary cell lines. (**C**) The proportion of mutated samples of cancer's top mutated genes in tumour samples and cell lines. (**D**) The proportion of copy number-altered samples of cancer's top variant genes in tumour samples and cell lines. (**E**) The tumour samples' copy number of breast cancer. (**F**) The expression of the top 5 differentially expressed genes of tumour samples in normal samples and cell lines. N represents normal samples and C represents the optimal cell lines we detected. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

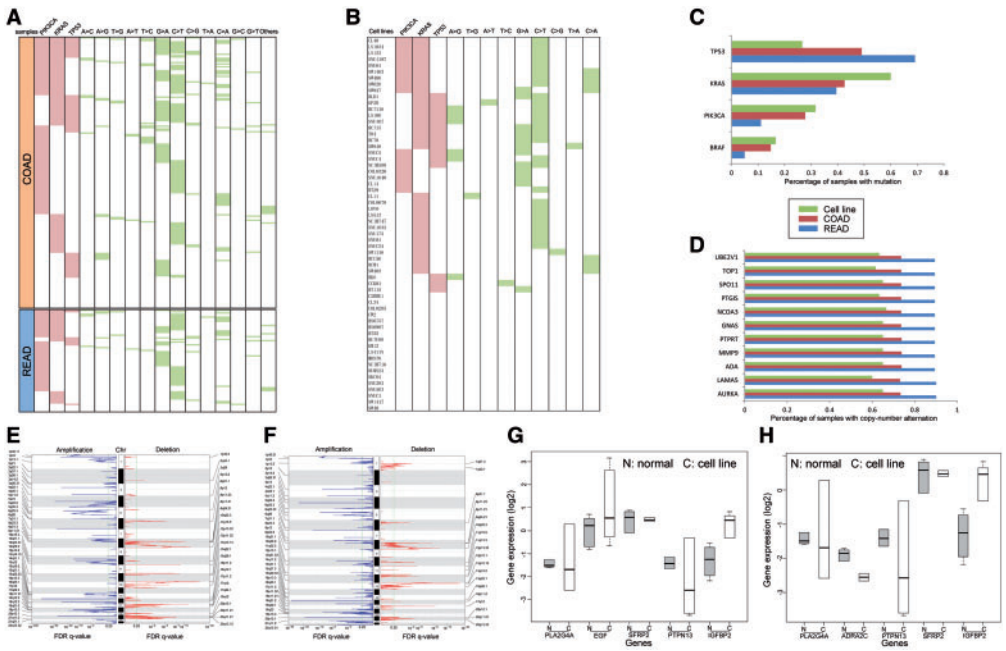

Figure 5. Comparison of mutation, copy number and expression between tumour samples and cell lines of intestinal cancer. (**A**) The mutation of intestinal cancer samples. (**B**) The mutation of intestinal cell lines. (**C**) The proportion of mutated samples of cancer's top mutated genes in tumour samples and cell lines. (**D**) The proportion of copy number-altered samples of cancer's top variant genes in tumour samples and cell lines. (**E**) The tumour samples' copy number of COAD. (**F**) The tumour samples' copy number of READ. (**G** and **H**) The expression of the top 5 differentially expressed genes of COAD or READ in normal samples and cell lines, respectively. N represents normal samples and C represents the optimal cell lines we detected. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

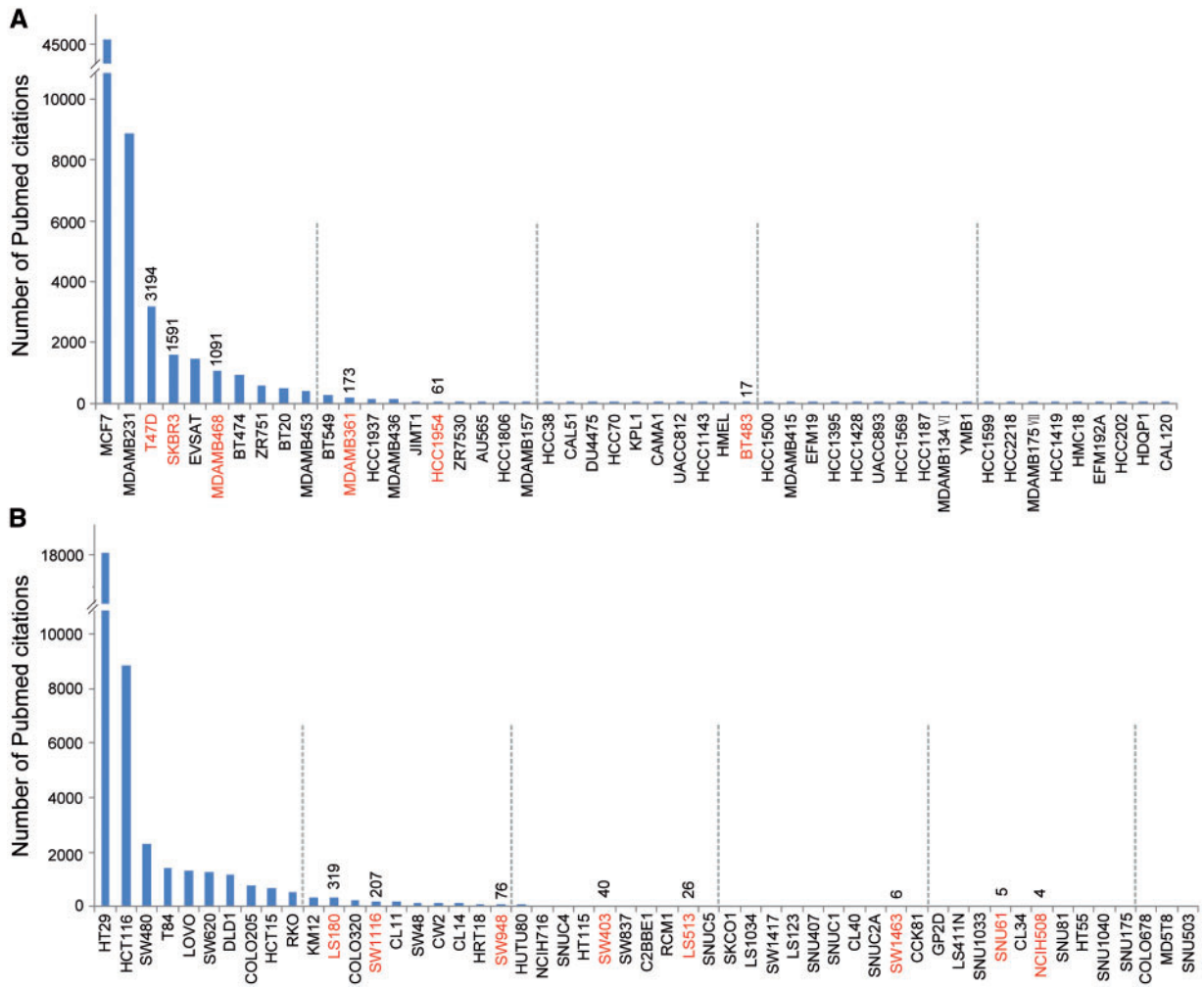

Figure 6. The literature references of cell lines in breast cancer and intestinal cancer. The optimal cell lines detected in this work are highlighted. And the references count is marked above the optimal cell lines. A dotted line is placed every 10 cell lines. (**A**) The literature references of mammary cell lines. (49 cell lines are shown. CAL148, CAL851, HCC2157, HS274T, HS281T, HS343T, HS578T, HS606T, HS739T and HS742T have no application article.) (**B**) The literature references of intestinal cell lines (53 cell lines are shown. HCC56, HS675T, HS698T, NCIH747, OUMS23, SNU1197 and SNU283 have no application article.). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 8. Known contaminated cell lines

Cell line	Claimed type	Contaminating cell line	Actual type	Reference (PMID)	GE	CN	GO	GM
BT20	Breast	HeLa	cervix	6451928	–	–	–	PIK3CA
KPL1	Breast	MCF-7	Breast	DSMZ	4	–	–	PIK4CA
YMB1	Breast	ZR-75-1	Breast	23136038	14	–	–	–
GOS3	Glioma	U-343 MG	Glioblastoma	DSMZ	8	–	13	–
LN319	Glioma	LN-992	Glioblastoma	22570425	–	–	–	–
SF767	Glioma	ME-180	cervix	22570425	–	–	11	
LN443	Glioblastoma	LN-444	Glioblastoma	22570425	–	–	15	–
SNB19	Glioblastoma	U-251 MG	Glioblastoma	17254797	–	–	–	TP53
U118MG	Glioblastoma	U-138 MG	Glioblastoma	ATCC	–	–	–	–

GE, CN and GO columns represent the gene expressed, copy number altered and functional ranks of the cell line in this work, respectively. It will be marked a '–' if the cell line is not on the top. GM column is the mutated gene in the corresponding cell line.

and A2780, which together account for 60% of the publications in their cell line panel, are poorly suited as models for HGSOC. In our work, the similarity between ovarian tumour samples and the two cell lines is noticeably low. Our results confirmed that the cell lines are unsuitable *in vitro* models of ovarian cancer.

In addition, Wilding *et al.* [58] found that in colorectal cancer studies, most of the more commonly referenced cell lines are in fact mismatch repair deficient (DLD1, RKO, HCT116, LOVO, LS174T, SW48). These cell lines were not suitable cell line models in our study, either.

Finally, to verify the consistency of our results, we calculated the similarity between cell lines and tumour samples using data from a different source. Series GSE2109 was downloaded from the GEO database, which summarizes the expression data

Table 9. The number of related GEO samples of optimal cell lines

Cell line	Cancer type (total number of samples)	Number of related samples
BT483	Breast (353)	319
MDAMB361	Breast (353)	314
HCC1954	Breast (353)	301
MDAMB468	Breast (353)	286
T47D	Breast (353)	284
SNU61	Intestine (400)	348
LS180	Intestine (400)	335
SW403	Intestine (400)	332
SW1463	Intestine (400)	331
SNU283	Intestine (400)	330
TUHR4TKB	Kidney (281)	270
TUHR14TKB	Kidney (281)	265
KMRC20	Kidney (281)	263
VMRCRCZ	Kidney (281)	250
SNU8	Ovary (199)	175
OVKATE	Ovary (199)	175
OVCAR4	Ovary (199)	174
OVSAHO	Ovary (199)	169
OC316	Ovary (199)	122
TYKNU	Ovary (199)	22

of multiple cancers. We used the data from four types of cancers (breast, intestine, kidney and ovary) to verify our optimal cell lines ($|correlation\ coefficient| > 0.45$, $P < 0.05$). Based on the results shown in Table 9, the optimal cell line remained similar to most of the tumour samples, except TYKNU. This cell line did not have a high rank in expression when calculated using the TCGA data set previously, but it had a high degree of similarity in copy number, mutation and function of the tumours. In the summary, these results confirmed the optimal cell lines using different sources of data.

Discussion

Cell lines have been widely used as *in vitro* models of cancer in laboratories. However, increasingly, researchers are finding that some cell lines are no longer suitable for *in vitro* cancer studies because they are contaminated or mutated. To improve the accuracy of scientific research, it is necessary to find more suitable cell line models for each cancer type.

In this work, we reviewed the methods for assessing cell line suitability since 2005. Although there are important discoveries revealed by these studies, there are also limitations. First, most of the studies only focus on one cancer or even one cell line; secondly, the majority of the studies did not use omics data. Therefore, in our work, we took advantage of additional types of genomic data to screen for the optimal cell lines for different types of cancers. We used the expression, copy number and mutation profiles to look for the optimal *in vitro* model of cancer by comparing the cell lines and tumour samples. Ultimately, one or more optimal cell lines were identified for eight different types of cancers. The optimal cell lines that we identified have high similarity with tumour samples in more than three of the four aspects (gene expression, copy number alteration, mutation and function), and these cell lines proved to have the same genomic characteristics as the tumour samples.

To verify the feasibility of our method, we used the method to calculate the similarity between samples from the same or different cancers. The results showed that the method can distinguish samples from different cancers and could identify the optimal cell lines for each type of cancer. To obtain more accurate results, we performed multiplatform normalization before calculating the correlation. Extremely strict statistical thresholds were set during all of the calculation processes. The genomic characteristics of the optimal cell lines have been evaluated to verify their accuracy. Finally, the accuracy of our method was verified by known contaminant cell lines from another perspective.

Cell lines could be ideal *in vitro* models of cancer because they are derived from tumour samples and have unlimited proliferation ability. The validation of the same genome characteristics of cell lines and tumour samples illustrates the mechanism from the molecular level.

Domcke *et al.* [18] summarized the genomic characteristics of the HGSOC tumour samples, and sought similar cell lines according to these characteristics. Although they found the optimal cell line accurately, other types of cancers still need to be evaluated. Alternatively, we used the genomic profiles (expression, copy number and mutation) to calculate the similarity between samples. This method identifies cell lines that best approximate the cancer samples and does not require the characteristics of the cancer to be summarized in advance. Therefore, this method can be applied to various cancers. The results will no doubt help researchers find more suitable *in vitro* models for cancer research, which will result in more accurate cancer mechanisms being uncovered. In our work, some cell lines had a high-count score, which illustrated that they were similar to many tumour samples. However, these cell lines did not all have a high correlation coefficient. This may be because of the fact that high frequency, recurrent somatic mutations or DNA copy number alterations are surprisingly rare in most solid tumours [59, 60].

Among the cell lines that were not found to be optimal, a few have outstandingly high correlations with tumour samples based on their expression level. However, these cell lines do not have high correlations with tumour samples based on their copy number or mutation levels. These cell lines may be the most effective in studies that evaluate only one feature (such as HCC38 when considering only expression). However, the cell lines that have low correlations with tumour samples in all aspects may not suitable for cancer research any more (e.g. MDAMB134VI).

Some of the optimal cell lines we identified have extremely low frequency of use research. Although these cell lines have high relevance to cancer, they were not used in many studies because of limited laboratory resources or the unclear definition of the cell line subtypes. Conversely, some of the routine cell lines used in research were not on the optimal list of *in vitro* tumour models, including BT20, EVSAT, BT549 and A2780. The use of these unsatisfactory cell lines may affect the accuracy of the experimental results obtained. Therefore, it is extremely necessary to start using the most highly suitable cell lines.

Moreover, some cell lines may be suitable, but were not considered because there is no expression, copy number or mutation data for them currently. With the flourish of sequencing technology, we will have the opportunity to investigate these cell lines more comprehensively soon.

The present study only compared the expression, copy number, mutation and function of cell lines and tumours. However, there are other factors closely related to cancer (e.g. methylation) that were not considered because of data limitations. It is insufficient that we did not perform biological experiment validation on our results. In the future, we

hope that the results will be further validated, and additional genome characteristics can be considered more comprehensively to perform more detailed and accurate screening of the cell lines.

Key points

- The identification of optimal cancer cell lines facilitates *in vitro* experiments, increases the experimental accuracy and helps identify more effective cancer treatment strategies.
- We summarized the cell line assessment methods used since 2005 and proposed a more comprehensive method that integrated multi-omics data to evaluate cell line similarity to primary tumours.
- We identified the optimal cell lines for eight types of cancers and found that some of the commonly used cell lines are not suitable cancer models.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

National Natural Science Foundation of China (grant number 81372492, in part] and Scientific Research Fund of Heilongjiang Provincial Education Department (grant number 12541278, in part).

References

1. Jones S, Chen WD, Parmigiani G, *et al*. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci USA* 2008;**105**:4283–8.
2. Borrell B. How accurate are cancer cell lines? *Nature* 2010;**463**:858.
3. Ertel A, Verghese A, Byers SW, *et al*. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Mol Cancer* 2006;**5**:55.
4. Stein WD, Litman T, Fojo T, *et al*. 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* 2004;**64**:2805–16.
5. Gillet JP, Calcagno AM, Varma S, *et al*. 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* 2011;**108**:18708–13.
6. Sandberg R, Ernberg I. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proc Natl Acad Sci USA* 2005;**102**:2052–7.
7. Roschke AV, Tonon G, Gehlhaus KS, *et al*. Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res* 2003;**63**:8634–47.
8. Daniel VC, Marchionni L, Hierman JS, *et al*. A primary xenograft model of small-cell lung cancer reveals irreversible changes in gene expression imposed by culture *in vitro*. *Cancer Res* 2009;**69**:3364–73.
9. Tveit KM, Pihl A. Do cell lines in vitro reflect the properties of the tumours of origin? A study of lines derived from human melanoma xenografts. *Br J Cancer* 1981;**44**:775–86.
10. MacLeod RA, Dirks WG, Matsuo Y, *et al*. Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *Int J Cancer* 1999;**83**:555–63.
11. Ricke RM, van Deursen JM. Aneuploidy in health, disease, and aging. *J Cell Biol* 2013;**201**:11–21.
12. Vogelstein B, Papadopoulos N, Velculescu VE, *et al*. Cancer genome landscapes. *Science* 2013;**339**:1546–58.
13. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006;**22**:2291–7.
14. Qiu YQ, Zhang S, Zhang XS, *et al*. Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* 2010;**11**:26.
15. Cao Y, DePinho RA, Ernst M, *et al*. Cancer research: past, present and future. *Nat Rev Cancer* 2011;**11**:749–54.
16. Beroukhim R, Mermel CH, Porter D, *et al*. The landscape of somatic copy-number alteration across human cancers. *Nature* 2010;**463**:899–905.
17. Thomas RK, Baker AC, Debiasi RM, *et al*. High-throughput oncogene mutation profiling in human cancer. *Nat Genet* 2007;**39**:347–51.
18. Domcke S, Sinha R, Levine DA, *et al*. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat Commun* 2013;**4**:2126.
19. Olarerin-George AO, Hogenesch JB. Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Res* 2015;**43**:2535–42.
20. Ciriello G, Miller ML, Aksoy BA, *et al*. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* 2013;**45**:1127–33.
21. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. *Faseb J* 2008;**22**:2605–22.
22. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7.
23. Zhang K, Wang H. [Cancer genome atlas pan-cancer analysis project]. *Zhongguo Fei Ai Za Zhi* 2015;**18**:219–23.
24. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.
25. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013;**41**:D991–5.
26. Forbes SA, Beare D, Gunasekaran P, *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;**43**:D805–11.
27. Yu W, Clyne M, Khoury MJ, *et al*. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* 2010;**26**:145–6.
28. Amberger JS, Bocchini CA, Schiettecatte F, *et al*. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* 2015;**43**:D789–98.
29. Mermel CH, Schumacher SE, Hill B, *et al*. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:R41.
30. MacConaill LE, Campbell CD, Kehoe SM, *et al*. Profiling critical cancer gene mutations in clinical tumor samples. *PLoS One* 2009;**4**:e7887.
31. Benito M, Parker J, Du Q, *et al*. Adjustment of systematic microarray data biases. *Bioinformatics* 2004;**20**:105–14.

32. Rudy J, Valafar F. Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinformatics* 2011;**12**:467.
33. Agresti A. *Analysis of Ordinal Categorical Data*, Vol. **656**. John Wiley & Sons, 2010.
34. Kreimer A, Litvin O, Hao K, *et al*. Inference of modules associated to eQTLs. *Nucleic Acids Res* 2012;**40**:e98.
35. Witten D, Tibshirani R. A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis* 2007;**1776**:58–85.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**1**:289–300.
37. Bertrand D, Chng KR, Sherbaf FG, *et al*. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* 2015;**43**:e44.
38. Shi W, Balazs B, Gyorffy B, *et al*. Combined analysis of gene expression, DNA copy number, and mutation profiling data to display biological process anomalies in individual breast cancers. *Breast Cancer Res Treat* 2014;**144**:561–8.
39. Barallon R, Bauer SR, Butler J, *et al*. Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues. *In Vitro Cell Dev Biol Anim* 2010;**46**: 727–32.
40. Ayyoob K, Masoud K, Vahideh K, *et al*. Authentication of newly established human esophageal squamous cell carcinoma cell line (YM-1) using short tandem repeat (STR) profiling method. *Tumour Biol* 2016;**37**:3197–204.
41. Ye F, Chen C, Qin J, *et al*. Genetic profiling reveals an alarming rate of cross-contamination among human cell lines used in China. *Faseb J* 2015;**29**:4268–72.
42. Jager W, Horiguchi Y, Shah J, *et al*. Hiding in plain view: genetic profiling reveals decades old cross contamination of bladder cancer cell line KU7 with HeLa. *J Urol* 2013;**190**:1404–9.
43. Phuchareon J, Ohta Y, Woo JM, *et al*. Genetic profiling reveals cross-contamination and misidentification of 6 adenoid cystic carcinoma cell lines: ACC2, ACC3, ACCM, ACCNS, ACCS and CAC2. *PLoS One* 2009;**4**:e6040.
44. Somaschini A, Amboldi N, Nuzzo A, *et al*. Cell line identity finding by fingerprinting, an optimized resource for short tandem repeat profile authentication. *Genet Test Mol Biomarkers* 2013;**17**:254–9.
45. Castro F, Dirks WG, Fahnrich S, *et al*. High-throughput SNP-based authentication of human cell lines. *Int J Cancer* 2013;**132**:308–14.
46. Yu M, Selvaraj SK, Liang-Chu MM, *et al*. A resource for cell line authentication, annotation and quality control. *Nature* 2015;**520**:307–11.
47. Schweppe RE, Klopper JP, Korch C, *et al*. Deoxyribonucleic acid profiling analysis of 40 human thyroid cancer cell lines reveals cross-contamination resulting in cell line redundancy and misidentification. *J Clin Endocrinol Metab* 2008;**93**:4331–41.
48. Soussi T, Beroud C. Assessing TP53 status in human tumours to evaluate clinical outcome. *Nat Rev Cancer* 2001;**1**:233–40.
49. Berglind H, Pawitan Y, Kato S, *et al*. Analysis of p53 mutation status in human cancer cell lines: a paradigm for cell line cross-contamination. *Cancer Biol Ther* 2008;**7**:699–708.
50. Korch C, Spillman MA, Jackson TA, *et al*. DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. *Gynecol Oncol* 2012;**127**:241–8.
51. Liu M, Liu H, Tang X, *et al*. Rapid identification and authentication of closely related animal cell culture by polymerase chain reaction. *In Vitro Cell Dev Biol Anim* 2008;**44**:224–7.
52. Cooper JK, Sykes G, King S, *et al*. Species identification in cell culture: a two-pronged molecular approach. *In Vitro Cell Dev Biol Anim* 2007;**43**:344–51.
53. Mirjalili A, Parmoor E, Moradi Bidhendi S, *et al*. Microbial contamination of cell cultures: a 2 years study. *Biologicals* 2005;**33**:81–5.
54. Valletta E, Kucera L, Prokes L, *et al*. Multivariate calibration approach for quantitative determination of cell-line cross contamination by intact cell mass spectrometry and artificial neural networks. *PLoS One* 2016;**11**:e0147414.
55. Melcher R, Maisch S, Koehler S, *et al*. SKY and genetic fingerprinting reveal a cross-contamination of the putative normal colon epithelial cell line NCOL-1. *Cancer Genet Cytogenet* 2005;**158**:84–7.
56. Capes-Davis A, Theodosopoulos G, Atkin I, *et al*. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer* 2010;**127**:1–8.
57. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.
58. Wilding JL, Bodmer WF. Cancer cell lines for drug discovery and development. *Cancer Res* 2014;**74**:2377–84.
59. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**:719–24.
60. Greenman C, Stephens P, Smith R, *et al*. Patterns of somatic mutation in human cancer genomes. *Nature* 2007;**446**:153–8.